THE AUSTRALIAN NATIONAL UNIVERSITY

Research Evaluation and Policy Project

Research School of Social Sciences

Quantitative indicators for research assessment – a literature review

Literature Review
for
ARC Linkage Project:
The Strategic Assessment of Research Performance Indicators

REPP Discussion Paper
05/1
30 March 2005

This is the first draft of a literature review undertaken as part of the ARC Linkage Project, The Strategic Assessment of Research Performance Indicators. While the review is extensive, there are still some sections to complete and a final edit to be undertaken.

REPP Discussion Papers are scholarly papers that report research in progress. They can be downloaded free of charge (PDF)

Research Evaluation and Policy Project
Research School of Social Sciences
The Australian National University
Canberra ACT 0200
Tel: (02) 6125 4849
Fax: (02) 6125 9767
http://repp.anu.edu.au

CONTENTS

List of Indicators

List of Figures

1 Introduction

This literature review was undertaken as part of an Australian Research Council (ARC) project, "Strategic Assessment of Research Performance Indicators", which was funded to examine quantitative performance indicators used in the evaluation of research. The overall aim of the project is to establish a knowledge base on performance measures, containing a comprehensive coverage of indicators and an assessment of their validity, fairness, transparency and impact on research, and the cost of implementation. This review seeks to summarise the 'state of the art' by giving an overview of quantitative indicators that are currently in use, or have been proposed, and locating any assessments of the measures that have already been undertaken. In this way, any gaps in our knowledge of quantitative indicators that need to be addressed in further research can be identified.

Throughout this review we will refer to two groups of quantitative measures – bibliometric indicators and non-bibliometric indicators. While a detailed description of the indicators is left to Section 4, brief definitions of a number of terms are necessary at this point. *Bibliometric indicators* are based on published literature in all its forms – journal articles, monographs, book chapters, conference papers, patents, and the like – and the references these publications contain (i.e. *citations*). *Non-bibliometric measures* encompass all other readily quantifiable indicators, such as ability to attract external funding and measures of esteem (honours and awards, editorship of journals, keynote addresses, etc). They also cover rapidly expanding ventures into web-link analysis.

Given the burgeoning increase in the use of quantitative measures to evaluate research performance, it is surprising that there is scant literature summarising assessments of their use in this context. Reviews have usually been confined to bibliometric indicators (e.g. Narin 1976; selected chapters of Van Raan 1988; Verbeek *et al.* 2002; Moed *et al.* 2004: 2). They do not cover the broader range of quantitative indicators, and do not touch on many important issues that must be understood before they can be applied in the assessment of research.

The sociology of science, which should have provided important background information about relationships between indicators and the research process, has not been interested in quantitative indicators for more than 20 years. With its constructivist turn, researchers in the discipline questioned the validity of bibliometric indicators on very basic principles (Gilbert and Woolgar 1974; Woolgar 1991) and has since ceased to make contributions to the topic. A recent trend is the emergence of a body of literature produced by scientists from outside science studies who have become interested in performance indicators. This literature is scattered across a variety of fields and tends to be fragmented and non-cumulative, often with little reference to the literature that does exist. Our review will bring together the different strands of literature on research performance indicators.

For reasons that will be discussed more fully in Section 3.4, the development of bibliometric indicators has been oriented by the research and publication practices of the natural sciences. With the increasing popularity of research evaluations, efforts are now being made to develop indicators which are more appropriate for the social sciences and humanities. We will give special attention to these new developments.

We start with fundamental questions concerning all quantitative indicators, including a theoretical discussion of the key concepts underlying the use of quantitative measures for the assessment of research performance (section 2.1), followed by discussions of the basic concerns that must be resolved before any attempt is made to utilise these indicators in practice (section 2.2). In section 3, we focus on issues relating to specific types of indicators. Sections 3.1 to 3.3 cover validity and technical questions associated with bibliometric indicators; Section 3.4 summarises the literature covering the use of bibliometrics for the social sciences, arts and humanities; Section 3.5 reviews patent indicators; and Section 3.6 discusses the limited literature on non-bibliometric indicators. Section 4 provides a detailed description of the quantitative indicators, including their calculation, data requirements, and any assessments of their use as measures to evaluate research performance. Given that the literature is far from comprehensive, these descriptions must remain patchy. In conclusion, Section 5 provides a brief overview of the literature, highlighting the major gaps that exist and that need to be addressed in future research.

2 Fundamental steps in designing the assessment process

2.1 Defining key concepts

The intention of any application of performance indicators is to either identify 'high quality' or to find out which research is 'better'. It is therefore not surprising that the application of quantitative indicators has triggered a debate about our understanding of what is meant by 'research quality', and how the various indicators are related to it. In this section, we discuss the key concepts that are used in this debate, namely 'research quality', 'impact', and 'scientific excellence'.

The literature focuses on the relationship of bibliometric indicators to these concepts, as it is this connection that has attracted most investigation. The development of these indicators has been advanced by the bibliometrics community, while other quantitative indicators have not been systematically scrutinised and are applied in an ad hoc and in a much less sophisticated way.

2.1.1 Research quality

In order to clarify what is being measured by performance indictors, a number of attempts have been made to define what is meant by 'research quality'. Cole and Cole suggested that quality might be defined in two ways. Firstly, "A traditional historian of science might apply a set of absolute criteria in assessing the quality of a paper. Those papers which embody scientific truth and enable us to better understand empirical phenomena are high-quality papers" (Cole and Cole 1973: 23). Measuring the quality of work in historical retrospect has the advantage that it does not exclude work that is momentarily fashionable or temporarily ignored. Secondly, Cole and Cole suggest using a social as opposed to an absolute definition of quality. They say that since absolute truth does not exist, high-quality work can be defined "as that which is currently thought useful by one's colleagues. If scientists in their everyday behaviour find a particular idea useful in their work, that idea is a valuable one, and we shall call it a high-quality idea." (ibid: 24).

The relative and social character of research quality is also discussed by other authors (e.g. Martin and Irvine 1983: 70-71; Herbertz and Müller-Hill 1995: 961). Martin and Irvine define quality as "a property of the publication and the research described in it. It describes how well the research is done, whether it is free from obvious 'error', how aesthetically pleasing the mathematical formulations are, and so on". Martin and Irvine claim that "quality is relative, it is socially and cognitively determined; it is not just intrinsic to research but judged by others with differing research interests and social goals". Similarly, Moed *et al.* (1985a: 134) refer to the multidimensional character of research quality and differentiate between a cognitive (importance of the specific content of scientific ideas), methodological (accuracy of methods and techniques) and aesthetic dimension of quality. They contend that these three dimensions of quality cannot be measured by quantitative indicators, but can only be judged by peers.

Van Raan suggests the following definition:

> Quality is a measure of the extent to which a group or an individual scientist contributes to the progress of our knowledge. In other words, the capacity to solve problems, to provide new insights into 'reality', or to make new technology possible. Ultimately, it is always the scientific community ('the peers', but now as a much broader group of colleague-scientists than only the peers in a review committee) who will have to decide in an inter-subjective way about quality (Van Raan 1996: 398).

While the introductory sentence of this quote appears to indicate that research quality is a quantitatively measurable property, Van Raan makes clear immediately afterwards that he, too, regards quality as something that in its complexity can only be judged by peers. This implies that quantitative indicators may be related to quality and measure certain aspects of it, but cannot exhaustively represent quality.

2.1.2 Impact

A way to circumvent the problem of defining 'quality' was to treat it as one characteristic of research among others. In an earlier work, Lawani (1977) claimed that there was a difference between 'impact', 'influence' and 'quality' though he failed to clarify that difference. Similarly, Martin and Irvine introduce a differentiation between 'quality', 'importance' and 'impact' in relation to publications. Making a distinction from their definition of 'quality' (see above), they contend the 'importance' of a publication refers to its potential influence on surrounding research activities if communication channels in science were flawless (Martin and Irvine 1983:70). Finally, they define the 'impact' of a publication as "the actual influence on surrounding research activities at a given time. While this will depend partly on its importance, it may also be affected by such factors as the location of the author, and the prestige, language and availability of the publishing journal" (ibid).

Because of the difficulty in defining a concept of quality that is appropriate when analysts seek to apply quantitative measures, discussions have concentrated on clarifying what 'impact' was. Contrary to attempts to define 'research quality', there is much more agreement on how to define the 'impact of research'. In addition to Martin and Irvine's definition, other authors have defined it along similar lines:

- the actual importance of a paper judged by the scientific community (Dieks and Chang 1976: 249, 265);
- the reception of research work by other scientists (Weingart *et al.* 1988: 396); or
- the effect which a published research finding has on its audience (Phillimore 1989: 263).

Moed distinguishes between 'short-term impact' and 'long-term impact'. The former is the impact of research at the research front a few years after publication which is usually a precondition for a long-term impact. The 'long-term impact' indicates whether and to what degree the research has made a more permanent contribution to scientific advance Moed *et al.* 1985a: 133-134). This has implications for the time-span for which research is evaluated (see Section 3.3.2).

In contrast to the views of Lawani (1977), impact is now regarded as one aspect of research quality that can be measured by quantitative indicators (e.g. Nederhof and Van Raan 1987: 334, Van Raan 1996: 404). Research quality as a whole needs to be assessed by peers. However there is an emerging trend in science policy to regard *impact*, the measurable part of quality, as a *proxy measure for quality* in total.

2.1.3 Scientific excellence

In recent years, 'scientific excellence' has become a buzzword, triggered by the desire of science policy to identify and promote 'excellent research,' and this term is increasingly used in bibliometric studies (e.g. Hicks *et al.* 2000; Van Raan 2000; Tijssen *et al.* 2002; Tijssen 2003; Van Leeuwen *et al.* 2003). Attempts to define 'scientific excellence' have led to the same difficulties encountered with the definition of 'research quality'. For example, Tijssen (2003: 93) suggests using the term 'excellence' as:

> …a comparative expression denoting superiority to others in scientific quality based on the best scores within a given set of comparable entities. Once internationally recognized 'dimensions' are identified and appropriate measurement models have been developed, distinctive elements of research excellence can be subjected to comparative measurements and rendered useful for further analysis.

Once again the problem of defining quality is circumvented by asking which research is 'the best' according to some measurable dimensions of quality. The attempt to measure excellence by Tijssen and others employs indicators based on publication and citation indicators, which are generally accepted as measuring only the impact aspect of scientific quality. Van Raan makes this explicit: "'Scientific excellence' can be operationalised as research groups with an impact above a specific threshold value" (Van Raan 2000: 305).

Both authors operationalise the concept of excellence as a relationship between pieces of research which is characterised by one piece being far ahead of the other ('excelling') in one or more quantitative dimension. This is not a problem in itself as long as the users of studies of 'research excellence' are aware of the reduction that is necessary for measurement. Unfortunately, the everyday meaning of the concept, which is used in science policy discourses, is far more complex than the meaning used in measurements.

Furthermore, there is an increasing tendency to borrow methods from industry (e.g. 'benchmarking' or 'scoreboards'). Grupp and Mogee warn that the use of these methods "may be dangerous because the numbers provided are taken at face value with little discussion of their validity. Substantial space exists for manipulation by selection, weighting and aggregating indicators" (Grupp and Mogee 2004: 86-87).

2.2 How to use quantitative indicators

Applying performance indicators in practice is not a straightforward task. It is important to clarify what role the indicators will play in the assessment of research; which indicators should be selected; and what the unintended consequences their application could be. For all quantitative indicators there is the problem that research practices vary across different fields, and it is necessary to determine level of aggregation to use and the form in which the results will be presented. The vast majority of the literature discusses these problems only for bibliometric indicators. However, they affect all quantitative indicators.

2.2.1 The role of quantitative indicators

What **role** should quantitative indicators play in the evaluation process? The literature reveals two different answers to this question — to replace peer review completely or use it as an aid to peer assessment. The answer is important, as the significance of many of the issues discussed in this review is directly related to the use to which the measures are being put.

(i) Replacement of peer review

Can peer judgements be replaced by quantitative analyses? Oppenheim (1997) suggested that for reasons of costs, the peer judgements in the British Research Assessment Exercise (RAE) should be replaced by citation analyses. However, studies of the correlation between RAE scores and bibliometric measures have shown that, while the correlation is high, there are deviant cases (Warner 2000). Warner cautions that their existence raises concerns about the straight replacement of peer review by bibliometrics where funding decisions are coupled research assessment.

The generally good pattern of correspondence between bibliometric indicators and peer judgements has sometimes led to the tendency to point to bibliometric indicators as objective measures in contrast to the subjective character of the peer review. But it should be remembered that bibliometric indicators themselves are based in part on peer decisions, e.g. journal articles embody the peer evaluations that have led to acceptance for publication (Weingart 2003).

There has been a recent proliferation of institutional rankings, often utilised as public relations tools, which raises the spectre of an increasing displacement of peer review by quantitative methods. We discuss the technical concerns about rankings in Section 2.2.6, but also raise concerns about their use at this juncture. They have little, if any, involvement of peers, and often apply quantitative indicators in a less-than-expert manner (Van Raan, 2005). Van Raan observed a general proliferation of "rapid and, particularly, cheap 'evaluations' with help of standard journal impact factors" (Van Raan 2005) at the instigation of heads of institutions or agencies.

(ii) Enhancing peer review

Most researchers see bibliometric indicators being used, not to replace the peer evaluation, but rather to make the results of research assessment debatable and to offer experts additional information (Van Raan and Van Leeuwen 2002: 614). Peer review can become more 'transparent' by using bibliometric indicators and can counterbalance the shortcomings of the peer review (e.g. Van Raan 1996: 401; Tijssen 2003: 101; Aksnes and Taxt 2004: 40). For example, they can control the rapid decline of research fields and protect them against the operation of 'old boy's networks' in peer reviews (Weingart 2003).

They are also seen by scientists as a useful resource in cases of doubt within panel discussions of peers (Moed and Van Raan 1988: 184). In addition, they can be used to highlight gaps in the knowledge of peers – as "triggers to the recognition of anomalies" (Bourke *et al.* 1999). Where the indicators do not align with peer evaluation, then the reasons must be sought. It may be due to problems with the indicators, or it may be that the experts have an incomplete knowledge of the research they are assessing (Bourke and Butler 1995). Inconsistencies between quantitative data and peer review are likely to trigger additional, deeper analyses of the performance of units being evaluated by those conducting the assessment.

From this discussion it becomes clear that quantitative analyses of research performance should enhance rather than replace peer review. There is widespread consensus on this recommendation in the literature. Many indicators are highly aggregated statistics in which a lot of information is lost (Moed *et al.* 1995: 400). Therefore data from such analyses should always be interpreted with background knowledge about specific publication habits, characteristics of the research field, etc. (Moed *et al.* 1985c; Van Raan 2002: 615; Gläser *et al.* 2004 : 32).

A recent trend which has raised some concern is the increased application of *'amateur' bibliometrics.* Van Raan points to a proliferation of 'rapid and, particularly, cheap 'evaluations' with the help of standard journal impact factors' (2005) at the instigation of heads of institutions or agencies. He adds this pressure from administrators to another cause identified by Weingart – the loss of competence and skill to interpret the indicators due to the increased commercialisation of the monopolist data producer ISI (Weingart 2004) – as the origins of their increased use.

The haphazard 'feeding in' of bibliometric indicators to a peer review, rather than using them as additional independent information, can distort the evaluation when the existence of anomalies is not recognised and investigated. (Gläser 2004a: 5). Gläser identifies the main danger of this practice is that amateur bibliometricians, not understanding all the factors behind the use of such measures, are likely to trust them blindly (Gläser and Laudel 2005). They fear that such practices may lead to a situation where 'bibliliometrics involuntarily takes over because peers do not judge content anymore'. The time pressures placed on reviewers increases the likelihood of this occurring.

2.2.2 Selection of indicators

Since each indicator has different strengths and weaknesses, it has been suggested that evaluations should always incorporate more than one indicator (Martin and Irvine 1983), and that indicators should never be used in isolation, especially if applied to individual groups (Van Raan 1996: 398-400). This is also the proposed standard practice for OECD surveys of R&D activities (Godin 2002a: 7). The Centre for Science and Technology Studies (CWTS) puts this into practice by always using a set of indicators in their evaluative studies (described in Van Leeuwen *et al.* 2003). Their standard indicators are highlighted in the description of individual indicators in Section 4.

The selection of a suitable suite of indicators for a given evaluation task is by no means clear-cut. In a number of studies, Australian researchers have been sent questionnaires asking which indicators best reflect the work in their field, department etc. For example, in a study conducted by Hattie and colleagues (Hattie *et al.* 1991, Tognolini *et al.* 1994, Print and Hattie 1997) scientists rated a large list of indicators divided into six groups. Similar questionnaires were used in a study by Grigg and Sheehan (1989) and by a research group chaired by Linke (NBEET 1993). While the lists were comprehensive, none of the studies came up with a preferred set.

Martin suggests identifying the combination of indicators that provides the strongest correlations and thereby the best combination (Martin and Irvine 1983). However, important information may be lost if indicators are chosen on the basis of their convergence – contradictory results could enhance, rather than detract from, the analysis.

While many indicators have a common starting point – a particular data source – their final form may bear few similarities. There is considerable room for “manipulation by selection, weighting and aggregating indicators.” (Grupp and Mogee 2004: 86-87). These concerns have been specifically raised in relation to bibliometric indicators, where a special session of the major international conference in the discipline was devoted to the issue.[1]

Glanzel noted several reasons why it would be desirable to develop standardised indicators: to increase the reliability of bibliometric results; to improve the validity of bibliometric methods; and to make bibliometric data compatible (Glänzel 1996b: 167). Unfortunately little progress has been made in this direction. Surprisingly, Glänzel who pressed for standardisation, states towards the end of his article that: “Finally, it should also be stressed that standardisation does not necessarily mean that one standard has to be followed by all bibliometricians.” (ibid: 176) However he did follow this with the stipulation that “… all these standards should be properly documented, so that it is guaranteed that any user of bibliometric data will be sufficiently informed about origin and background of the data and possible compatibility problems.” (ibid: 176)

The importance of Glanzel’s last rider was starkly illustrated in the debate on the state of British science in the 1980s and early 1990s[2]. Much of the discussion in this literature concerned the efficacy of different ways of counting publications: argument has turned on the choice of ISI's databases; the use of ‘fractionated’ rather than ‘whole’ counts for publications where authors from more than one country were involved; the use of a constant or fixed journal set versus the use of a dynamic journal set which expands with the introduction of new journals; the publication types to include; whether to use ‘citing-year’ or ‘cited-year’ citation counts.

2.2.3 Anticipating unintended consequences

> … I urge the funding councils to remember that all evaluation mechanisms distort the processes they purport to evaluate. (Roberts 2003: 3)

Any system used to assess research that affects money and/or prestige is likely to affect the behaviour of researchers and administrators. This applies equally to qualitative and quantitative

[1] Fifth International Conference of the International Society for Scientometrics and Informetrics, River Forest, Illinois in 1995. Selected papers from the session were published in *Scientometrics* 35, 1996.

[2] The British bibliometric debate about the state of the nation's science may be followed (by those with considerable stamina) in B.R. Martin, J. and R. Turner, "The writing on the wall for British Science", New Scientist 104 (1984): 25-29; J. Irvine, B.R. Martin, T. Peacock and R. Turner, "Charting the Decline in British Science", Nature 316 (1985): 587-590; B.R. Martin, J. Irvine, F. Narin and C. Sterritt, "The continuing decline of British science", Nature 330 (1987): 123-126; L. Leydesdorff, "Problems with the measurement of national scientific performance", Science and Public Policy 15 (1988): 149-152; J. Anderson, P.M.D. Collins, J. Irvine, P. Isard, B.R. Martin, F. Narin and K. Stevens, "On-line approaches to measuring national scientific output-a cautionary tale", Science and Public Policy 15 (1988): 153-161; T. Braun, W. Glanzel and A. Schubert, "Assessing assessments of British science: some facts and figures to accept or decline", Scientometrics 155 (1989): 165-170; B.R. Martin, J. Irvine, F. Narin, C. Sterritt and K. Stevens, "Recent trends in the output and impact of British science", Science and Public Policy 17 (1990): 14.-26; B.R. Martin, "The bibliometric assessment of UK scientific performance, a reply to Braun *et al.*", Scientometrics 20 (1991): 333-357; T. Braun, W. Glanzel and A. Schubert, "The bibliometric assessment of UK scientific performance-some comments on Martin's reply", Scientometrics 20 (1991): 359-362; L. Leydesdorff, "On the scientometric decline of British science", Scientometrics 20 (1991): 363-367; T. Kealey, "Government-funded academic science..." Scientometrics 20 (1991): 369-394; and "The growth of British science, " Nature 350 (1991):, 370; B.R. Martin and, J. Irvine, "The position of British science", Nature 355 (1992): 760; B.R. Martin, "British science in the 1980s-has the relative decline continued?" Scientometrics 29 (1994): 27-56.

assessments. Two major types of unintended consequences, which can occur as the result of any evaluation, have been identified.

One almost inevitable consequence of repeatedly applying the same type of measure is goal displacement (Perrin 1998) — high scores in the measures become the goal rather than a means of measuring whether an objective (or performance level) has been attained. The re-assessment of the RAE by the House of Commons Science and Technology Committee also devoted considerable time and space to the subject of 'game-playing' by universities in response to the assessment criteria (2004). A number of submissions to their enquiry go to the nub of this issue. One senior academic commented:

> … the improvement in results represented a "morass of fiddling, finagling and horse trading"(ibid 21).

A second effect may be that the research process itself is modified — researchers adapt their behaviour in response to the method of evaluation. This occurs in ways that are more complex and more difficult to observe than goal displacement. The adaptive response that can be observed most easily is a change in publication behaviour. Scientists reported changing their publication strategy by publishing more in international journals making it difficult for nationally oriented journals to attract sufficient manuscripts (De Bruin *et al.* 1993: 40). Although a stronger international orientation in publishing seems desirable, it could lead to the neglect of nationally important topics. Mojon-Azzi *et al.* (2003) investigated publications in opthalmological journals and found some work published twice, perhaps caused by a desire to increase the author's performance in assessment measures based on simple publication counts.

Other behavioural changes have been hypothesised, such as risk avoidance and clinging to the mainstream, which can lead to a reduced diversity of approaches in research. Marginson and Considine (2000: 141-174) claim that Australia's formula-based funding favours research quantity rather than research quality, short-term rather than long-term research, and mitigates against new researchers and emergent approaches. The authors' conclusions were largely reached on the basis of anecdotal evidence rather than resulting from systematic analysis. A recent review of the literature on performance-based funding did not identify any conclusive evidence about when (under what circumstances), how (by what processes) and with what effects researchers adapt to these influences (Gläser *et al.* 2002). A bibliometric study of Australia's scientific output showed a significant increase in the country's journal output in the mid 1990s. The timing of this productivity increase in relation to the introduction of funding formulas suggests a causal relationship (Butler 2003).

2.2.4 Recognising and dealing with the differences between fields

Research fields differ in their publication and citation practices, i.e. the number of publications produced and citations received in a certain time-span (e.g Moed *et al.* 1985c). The most obvious example is the differences that exist between the natural and physical sciences on the one hand, and the humanities, arts and social sciences on the other. There is general agreement that the research methods and orientations in the social sciences and humanities are distinct from those of the experimental sciences, that their communication practices or literatures are thereby differently structured and that this has bibliometric consequences (Hicks 2004, 473; Glänzel and Shoepflin 1999, 31; Luwel *et al.* 1999, 13; Nederhof *et al.* 1989, 427). While quantitative indicators (primarily bibliometric) are relatively straightforward for the evaluation of scientific research, we find that '[w]hen challenged to evaluate scholarly work in the social sciences and humanities, we are rudely forced to work outside this comfort zone' (Hicks 2004, 474). This issue is discussed more fully in a separate section on the humanities, arts and social sciences (3.4).

Even within seemingly coherent fields, sub-discipline variations in publication and citation practices can occur. For example, in high-energy physics theorists tend to publish more frequently than experimentalists (Irvine and Martin 1985: 304). Small fields encompassing a limited number of researchers have lower citation rates than those with many researchers (Van Raan 1996: 403). Fields differ in the kind of output they produce, e.g. whether they produce patents or not. Different types of publication are relevant in different types of research fields. While in some disciplines journal articles are the most important channel of communication, in others books, book chapters, and conference papers play a crucial role (e.g. Small and Crane 1979: 451, Butler 2005??).

Given these differences, the ideal approach would be to confine evaluations within the boundaries of a research field because this would mean comparing the 'same with the same'; a suggestion proposed by Collins (1985: 555-557). However, research assessments, particularly those linked to funding decisions, are just as likely to focus on organisations as they are to focus on fields (Martin and Irvine 1985: 559-560). Research organisations and also subunits like university departments typically encompass more than one research field (also caused by a growing number of interdisciplinary research processes). The dilemma can be solved by using sophisticated indicators which employ reference values to normalise indicators and thus make cross-field comparisons more valid (see Section 2.2.4 for a more detailed discussion of normalisation techniques).

2.2.5 Selecting the level of aggregation

Three factors are identified as particularly significant in determining the appropriate size for the unit of analysis: the focus of assessment, the division of labour within research and statistical relevance.

Focus of Assessment

A controversy between Collins and Martin and Irvine on the assessment of UK astronomy. Collins declared that research must be evaluated within its cognitive boundaries rather than within its institutional boundaries (Collins 1985: 555-557). Martin and Irvine countered this claim by stating that institutions (organisations, groups) rather than fields should be the unit of study in a science policy context because the aim is often to decide which institutions should be supported (Martin and Irvine 1985: 559-560). This dispute can be resolved when the unit of analysis is small enough to ensure that cognitive and institutional boundaries are closely aligned.

Division of labour within research

The performance of small units or individuals cannot always be measured independently where the level of collaboration is high or where specific work roles determine authorship on publications (Gläser *et al.* 2004). Publications become jointly 'owned' and most measures relate to the collaborative effort as a whole, rather than to the component parts. The authors suggest that in every research organisation there exists a 'least evaluable unit', that is "the smallest subunits to which the quantitative performance indicators can be applied" (ibid: 20). Section 3.3.5 discusses in detail the ways multi-authored papers can be handled.

Statistical significance

An additional limitation to the disaggregation of the unit of analysis is that the set of publications on which the analysis is based must be sufficiently large to support the application of statistical methods (Nederhof 1988: 207; Van Raan 1993: 163). The discussion about the validity of citation indicators has shown that it is important to allow for the highly skewed nature of the distribution of citations. Most publications receive relatively few citations, with only a tiny minority being heavily cited (Garfield 1979b). It is possible that the average citation rate of a research unit is high because one article of the group is highly cited, with other publications receiving very few. Concerns relating to this skewed distribution are most critical if the number of publications is small, i.e. less than 50 publications (Moed *et al.* 1995: 411).[3] Van Raan proposes ten or 20 publications per year — the usual output of a research group in the sciences – as a sufficient basis for bibliometric calculations while rejecting those based on a few publications per year (Van Raan 2000: 307-309). As analyses rarely focus on a single year, but instead are calculated across a three to five year window, the two minimal levels proposed are broadly similar.

Moed and Van Raan conclude that the group level is the lowest level sustainable for bibliometric analyses and it is inappropriate at the level of the individual (Moed and Van Raan 1988: 188-190). Van Raan also urges researchers to address the problem by developing better statistics than those based on mean values (Van Raan 1996: 404). One suggestion is to use median values of citations instead of mean values for ten publications or less (Salzarulo and von Ins 2001). Another strategy, recently adopted by CWTS, is to focus the analysis on the high impact end of the spectrum (Van Leeuwen *et al.*, 2003).

2.2.6 Presenting the results

The way in which the results of quantitative analyses are published can have a considerable impact on their usefulness, validity, and the degree to which they are universally accepted. Decisions made throughout the process, such as why a specific set of indicators has been selected, should be documented (Glänzel 1996b: 176).

Since there is no absolute yardstick for research quality, a single number describing one aspect of research quality does not reveal anything useful. There needs to be a benchmark against which performance can be judged (Van Raan 1996: 403). Statistical tests have been developed to check whether a citation rate differs significantly from the world average (Schubert and Glänzel 1983). However, it is usually more important to interpret differences between values (e.g citation rates) for the evaluated units, i.e. to decide whether they can they be considered as 'significant' with a certain probability or should be ascribed to mere chance (Moed *et al.* 1985c: 142). If the numbers of publications are small, only large differences are significant (Nederhof and Van Raan 1993: 334-335).

A popular way of presenting the data is in ranking tables. In the media, in recent months, there has been an explosion of tables ranking institutions on some perceived notion of quality. One ranking in particular, the Jiao Tong University in Shanghai (SJTU 2003) has stirred the imagination and energies of university administrators. Rank orders have considerable appeal because they are simple to construct (Van Raan 2005). The problem with rank orders is that they

---

[3] In research evaluations, undertaken by REPP, for citation analyses the evaluated unit needs to have at least 100 publications. Furthermore, it demands extra caution for citation analyses based on units having between 100 and 200 publications. (e.g. Butler 2001: 149).

'create' quality differences where the data on which the ranking is based might reveal only marginal differences in absolute numbers.

Rankings are usually constructed by aggregating a series of quantitative indicators. Standard bibliometric measures such as publications and citations are often central to the calculations. But those well-versed in the idiosyncrasies of indicator construction alert us to two worrying drivers to the proliferation of these tables. Weingart (2003, 2004) points to the huge influence ISI (who currently have a monopoly on citation data) has as it aggressively markets its off-the-shelf products. This has led to a rapid increased the non-expert use of bibliometric indicators such as rankings. Van Raan (2005) alerts to an influence he judges just as important: "organizations greedy to buy ready-to-go indicators without any competence to understand what is measured".

Steele highlighted the concerns about ranking tables: "In an ideal world, university league tables would be relegated to the intellectual basement", though he went on to accept that this was unlikely to happen (Steele 2004). He did, however, make a plea for co-ordinated analysis of the methodologies applied in the various rankings, and the need to 'establish improved data … and, most important, a wider examination of the implications for scholarly research' (ibid).

Many attempts were made to weight different indicators and to construct a composite index (e.g. Johnes and Johnes 1995; Nagpaul and Roy 2003). This next step is tempting as it produces a final rank order of the evaluated units. But any weighting is inevitably arbitrary (Johnes 1988: 57). There is no empirical basis for the choice of weights and experiments with alternative weighting systems have yet to be conducted (Lawani 1977: 26, Johnes 1988: 57).

3 Evaluating bibliometric indicators

While the previous section dealt with issues that relate to all quantitative indicators, this section discusses the validity and technical concerns surrounding bibliometric indicators. It would be preferable to extend the following considerations to all quantitative indicators. However, a literature review is limited to the information that is offered in prior publications, and both validity and technical issues in relation to non-bibliometric indicators are almost entirely absent from the literature.

3.1 The validity of using quantitative indicators to assess performance

Many of the concerns raised about the use of quantitative measures in the evaluation of science centre on the question of whether their use for such purposes is valid. This issue has already been mentioned in section 2.1 in the discussion of key concepts. There is a very large body of literature on the topic, though it is limited in that most references focus on bibliometrics, rather than quantitative indicators more generally.

3.1.1 Publication numbers – quantity not quality

A simple count of the number of publications is usually viewed as a measure of the productivity of research, or quantity rather than quality (Moravcsik 1973: 268). Indicators that are based on publication counts do not receive the same attention as citation indicators in discussions of validity because their connection to research quality is quite remote. The fact that an article has been published in a scholarly journal usually implies it has undergone a thorough peer review process by several reviewers, which might be interpreted as a sign of quality. However, it constitutes nothing more than an offer to the scientific community. It is the subsequent use of that offer that certifies the actual quality of an article (Gläser 2004b: chap. 2.4, 2.5).

While a Cole and Cole (1967) study of 120 researchers found a correlation between citations and publications, they also found that, at the individual level, citation measures could not be replaced by publication measures. Problems were encountered with ‘mass producers’, who publish a large number of papers but get only few citations, and ‘perfectionists’, who publish little but get high citations for their work (ibid: 382).

Rinia’s more extensive analysis of research groups in condensed matter physics found no significant correlation between the number of journal publications and peer review judgements (Rinia *et al.* 1998: 101). Moed also found no correlation between the quality and the pure number of publications (Moed *et al.* 1985b: 153). The contradictory results may be a statistical consequence of the relatively small output analysed by Cole and Cole in their studies.

Despite the almost universally acknowledged shortcomings of publication indicators as performance measures, a review of articles published in *Scientometrics* showed there was a tendency to employ only a single indicator in assessments, and that a count of publications was the most frequently used indicator (Martin 1996: 352-354; see also Toutkoushian *et al.* 2003).

3.1.2 What does a citation mean?

Much of the discussion that can be found in the literature on the validity of using citations to judge performance rests on the meaning that can be attached to a reference given by an author. Central to the concerns about their validity is a discussion of the precise motivation for one publication citing another. We summarise the main theoretical concerns that have been voiced about this use of citation analysis, and in the following section (3.1.3) detail some of the empirical studies that have been undertaken to test their validity.

For a review of these issues see Smith 1981 and MacRoberts and MacRoberts 1989; 1996. In her paper, Smith lists the fifteen reasons for citing another paper that Garfield (1965: 189) had previously enumerated[4].

While the meaning of citations presents one set of theoretical concerns about their application as performance indicators, there are additional concerns that have featured in the literature:

- citations do not reflect all influences on a scientific work, e.g. techniques and theories that become assimilated into the existing body of scientific knowledge (the 'obliteration phenomenon', King 1987: 265); influences that appear in acknowledgements (Cronin *et al.* 1993)[5]; and informal communication (Edge 1979);
- Work that is ahead of its time gets few citations (Cole and Cole 1967: 380; also called the "Mendel syndrome", Garfield 1979a: 364);
- The chance of being cited depends on factors other than the quality of the work: the language in which a paper is written (English language papers have a better chance of being cited); the type of publication: review articles, methods paper, theoretical papers, and papers in 'fashionable areas' are cited more often (e.g. Cole *et al.* 1978: 224-225; Garfield 1979b: 363-364); the arbitrary selections of references that authors make (Kochen 1974: 77, cited in Smith 1981: 84), eminent scientists are more often cited than others when given a choice (the so-called 'halo effect', Martin and Irvine 1983: 69).

These objections have been summarised by Johnes in the statement that our knowledge about citing habits is insufficient to be able to seriously consider citation analysis as a policy tool (1988: 61)[6]. Furthermore, it has been stated that the validity of citation analyses is low because of the incompleteness of the ISI databases and a high error rate, an argument we will deal with in Section 3.2.

The empirical base of the objections referring to citation behaviour is not very strong since it is mostly anecdotal. In contrast, studies of the usage of cited works in publications (so called citation context analyses) have shown that 90% of the references given are confirmative (Moravcsik and Murugesan 1975; Dieks and Chang 1976). In a detailed study of two highly cited articles, Cozzens (1985) showed how the knowledge claims of the articles are used by other authors. These findings were later confirmed by a citation content analysis, carried out by Amsterdamska and Leydesdorff (1989). 'Bad work' or 'wrong work' is much more often ignored than cited (Meadows 1974: 45; Herbertz and Müller-Hill 1995: 961). But even when

[4] 1. Paying homage to pioneers; 2. Giving credit for related work (homage to peers); 3. Identifying methodology, equipment, etc.; 4. Providing background reading; 5. Correcting one's own work; 6. Correcting the work of others; 7. Criticising previous work; 8. Substantiating claims; 9. Alerting to forthcoming work; 10. Providing leads to poorly disseminated, poorly indexed, or uncited work; 11. Authenticating data and classes of fact – physical constants, etc; 12. Identifying original publications in which an idea or concept was discussed; 13. Identifying original publications or other work describing an eponymic concept or term ; 14. Disclaiming work or ideas of others (negative claims); or 15. Disputing priority claims of others (negative homage)

[5] Baird and Oppenheim (1994: 8) argue that acknowledgements need not to be included as long as there is a strong correlation between citation counts and acknowledgement counts. Unfortunately, this has yet to be tested.

[6] Even Garfield, one of the originators of citation analyses for research evaluation purposes, has recently observed that we know little about uncitedness (Garfield 1998: 75-76).

citation biases occur, they are unproblematic as long as they are randomly distributed (Cole and Cole 1973: 25) and the sample is sufficiently large.

Van Raan (1998: 134-135) forcefully rebuts the argument that citations are meaningless indicators:

> That is statistically only the case if all researchers refer to earlier work completely arbitrarily. But nobody can seriously maintain that the references in, for instance, this paper are totally unreasonably and completely arbitrary ….valid patterns in citations will be detected if a sufficiently large number of papers is used for analysis. Furthermore, it is statistically very improbable that all researchers in a field share the same distinct reference-biases.

The number of citations is a measure of the strength of influence when it is applied to sufficiently large aggregates (Smith 1981: 88; Van Raan 2000). However, a detailed measurement model that links strength of influence to numbers of citations is still missing (Gläser and Laudel 2001: 428-429). Problems of validity were addressed by formulating methodological and technical requirements that have to be fulfilled before citation indicators can be used (see Section 3.3).

3.1.3 Do peer reviews and quantitative assessments agree?

Many studies have been conducted comparing the results of quantitative analyses (mainly citation studies) with peer judgements to demonstrate that indicators are valid measures of research quality and/or impact, but validation attempts have proved difficult. "The fundamental problem in any discussion of the validity of indicators of scientific productivity is the fact that there is no absolute standard of measure of such productivity" (Narin 1976: 82).

Some of the problems of using peer judgements for validation purposes were also discussed in detail by Martin and Irvin (1983). Nederhof claims that since the reliability of peer judgements is weak, the correlations between the peer review results and bibliometric indicators can only be moderately strong (Nederhof 1988: 209). In addition, peer review judgements and bibliometric evaluations are not unrelated – many peers use bibliometric measures in their deliberations, such as the number of publications in top journals (Van Raan 2004: 38). Nevertheless, peer review remains the main criteria against which the validity of quantitative measures is assessed.

A number of early studies were focussed on analyses of individual scientists. Many of these studies found a high correlation between scientists who were judged as 'leading' by peers and their citation rate (Cole and Cole 1967; Cole and Cole 1973; Virgo 1977; Zuckerman 1977). However, Lawani and Bayer (1983) were critical of the focus at the top end (especially with the concentration on Nobel prize winners), rather than on medium levels where it is harder to discriminate. They ran a validity study in which peer ratings in cancer research were correlated with citation rates. For peer ratings, they used the inclusion of articles in a yearbook of cancer research. They found a high correlation between papers judged by peers as very good and (high) citation rates, and also found a high correlation between papers judged by peers as average and (low) citation rates. Although Lawani and Bayer planned to do a qualitative study in order to investigate these deviant cases, their plans were never realised.

A general theme of the literature is that discrimination between research performance in the middle range can prove difficult. Nederhof and Van Raan (1987; 1989) sought to determine if this could be done using bibliometric techniques. Their analysis of cum-laude and non-cum laude doctorates showed that cum-laudes publish more and were more highly cited than non cum-laudes before graduation, but this difference rapidly vanished. Lindsey also found that the differences between articles that attract no citations and articles that attract two or three citations were not substantial and concludes: "Thus, in the heavily populated middle range of the continuum of quality, citation counts are of doubtful utility"(Lindsey 1989: 196).

Porter *et al.* chose a different approach for their validation study, scientists' self-assessment of their work. Sloan Fellows in chemistry were asked to name their three best publications. A comparison with their most highly cited publications showed that only a third of them were also assessed by the scientists as their best work. The scientists commonly named theoretical and empirical work as their best work (79%) and a much smaller number of methodical works (21%), while 48% of the most highly cited publications were methodical papers (Porter *et al.* 1988).

Validity studies that focussed on higher levels of aggregation generally produced more positive results than those focussed on the individual scientist or publication. The most extensive comparative study undertaken is that of Rinia *et al.* (1998). They found positive and significant, though not perfect, correlations between peer review and assessment and a number of citation measures, with the highest correlations being to citation per publication rates, and citation averages normalised to world averages (ibid: 105). This finding was corroborated by the study of Aksnes *et al.* (2004: 37) though their correlations, based on a much smaller study, were weaker. Rinia *et al.* also found that correlations were higher in the case of basic research than in more applied research (1998: 105).

A study that compared expert rankings of American universities with various publication and citation measures found highly positive correlations with the number of publications and with "the citation quality of the university's publications (influence per paper)" (Anderson *et al.* 1978). Oppenheim (1997, 2003) investigated a number of fields and found very high correlations between the peer ratings in the British Research Assessment Exercise (RAE) and the number of citations. Other studies found high correlations with the peer ratings of the RAE but were concerned about the poor connection for a few individual departments (Warner 2000; Smith and Eysenck 2002). The Van Raan and Van Leeuwen study (2002) showed a good correspondence between their bibliometric data and the opinions of heads of departments and senior staff members in a research institute.

The problem of discrimination within the mid-range, identified in studies of individual scientists, is also an issue at higher levels of aggregation. The study of Aksnes *et al.* found a good correspondence if a unit was highly or poorly cited, but no positive correlations at intermediate levels of citations (Aksnes and Taxt 2004: 38).[7]

In non-scientific disciplines, the form of validation stretches beyond assessing journal-based publications and then submitting the findings for peer judgement. In the humanities, Lewison (2001: 10-11) uses surveyed peer opinion alongside bibliometric data in order to define what counts as a 'quality' book in medical history, and obtains a high degree of qualitative agreement between citation outcomes and separately generated peer evaluation. For the humanities and social sciences Luwel *et al.* (1999: 22) rely upon peer judgements to operationalise a concept of academic 'quality' covering a myriad range of standard and non-standard publications, and they later assess the connection between quantitative indicators and peer opinion. They go on to champion a reflexive approach to research evaluation which routinely feeds back peer

[7] Although these results are in accordance with the results of former studies, they must be interpreted with great caution, since the authors admit in the discussion section that their methodological approach incorporates several weaknesses.

judgments about bibliometric findings (Moed *et al.* 2002: 517). Meho and Sonnenwald demonstrate a close relationship between journal-based citation analysis and peer opinion in the form of book reviews and solicited rankings of individuals in the social sciences (2000: 136).

In addition to correlation studies, the validity of citations as an indicator of impact has been tested by citation content analyses. As mentioned in the previous section, these give a causal explanation for the correlations. The validation of bibliometric studies at both the individual and the aggregate level shows a consistent picture. There is a good correlation between bibliometric evaluations and peer judgements, to which there are usually exceptions i.e. few individual cases (publications, researchers, or research organisations) where peer judgements significantly differ from bibliometric evaluations, and which is weaker in the middle of the spectrum.

The discussion of validity must necessarily return to a reflection on the role of quantitative indicators in the assessment of research (Section 2.2.1). The significance of many of the concerns on validity is reduced when the indicators are being used as an aid to peer reviews where differences between values can be interpreted and exceptions can be discussed. They are, however, at the forefront of concerns related to their use in isolation from informed peer input.

3.2 Indexed databases – traps for the unwary

With the exception of relatively simple counts, the data for most bibliometric indicators usually come from databases constructed by someone (or some enterprise) other than those seeking to utilise them. As a consequence, those employing the measures have no quality control over the data collection. They may not be cognisant of the peculiarities of particular databases and the way these can affect the validity of the indicators they are constructing. Of even more import, these databases are often constructed for reasons entirely unrelated to their recent utilisation for evaluation purposes (Wouters 1999). This section canvasses some of the most common pitfalls encountered when using third-party databases.

3.2.1 Coverage of the published literature

A very basic premise underpinning quantitative assessment is that bibliometric indicators are only appropriate where publications are the principal carriers of knowledge (Van Raan 1993: 152). It is essential to determine the proportion of total output that is covered by any database being used for the analysis. Without this knowledge, no judgement about the validity of any indicator constructed from its data can be made.

Since the ISI databases are the most extensively used for citation analyses, much of the discussion has centred on these, though the points are equally relevant to any database. ISI indexes include only a selection of the published literature, focussing primarily on journal articles. ISI claims to index the core journals of all research fields, providing "comprehensive coverage of the world's most important and influential journals" (Thomson ISI 2004). Their coverage of output from most disciplines in the natural sciences is good, however it is less comprehensive for the social sciences and humanities which rely heavily on 'non-source' publications, most notably books, monographs and book chapters, as a means for communicating research. In the applied sciences, conference proceedings and reports play an important role, and the majority of these also fall outside the coverage of ISI (Le Pair 1988; Van Raan and Van Leeuwen 2002: 614). Even within the natural sciences, there are disciplines such as taxonomy, mathematics, and earth sciences that are not as well covered (Moed and Hesselink 1996: 830; Korevaar and Moed 1996).

It was found that journals in small disciplines, locally oriented fields and those not published in English language are under-represented (Nederhof 1988: 203). Another problem is that new journals are only included into the ISI databases after two or three years of their existence (Rinia 2000: 368). This can create problems if newly developing fields are evaluated. Rinia recommends the use more than one database (e.g. SCI and Chemical Abstracts) in order to increase coverage (ibid.).

We have briefly touched here about concerns raised about the coverage of the social sciences, arts and humanities. As this issue is central to the debate on the application of quantitative indicators in these disciplines, in particular bibliometric measures, we go in to this in greater detail in 3.4 below.

Verification of coverage

The lower the level of aggregation, the more essential it is that all publications are included in the evaluation process. The smaller the number of publications, the more impact the skewed nature of citation distributions can have on the calculation of indicators. It has been proposed that at the group level, 99% completeness of publication data is necessary (Moed *et al.* 1985c: 139-140). If one well-cited publication is missing, this can lead to an error rate of up to 25% (Nederhof 1988: 204) or even 50% (Smith 1981: 93). The effect of errors will be reduced as the level of aggregation increases and the number of publications being analysed also increases (Nederhof 1988: 207)

To minimise the chance of major omissions, publication data should by verified by representatives of the evaluated research unit and/or by using the unit's research reports (Nederhof 1988: 204, Moed *et al.* 1995: 394, Van Raan and Van Leeuwen 2002: 615).

3.2.2 Coverage of References

Citation analyses are more difficult to undertake than publication analyses. Technically, they require access to ISI citation databases which are currently the sole provider of citation information[8]. Furthermore, these data must be obtained in a form that allows for automated manipulation. These prerequisites are rarely fulfilled.[9] Consequently, only a small number of groups exist world-wide that have the capacity to conduct citation analyses that extend beyond a few basic measures, particularly those undertaken at higher levels of aggregation where a large numbers of publications are involved.

As mentioned in the previous section, the ISI databases only include selected journals. The citations used in standard bibliometric analyses are the references contained in these selected journals (source journals) to other source journals. Citations to non-source publications, such as books, are also extracted from the source journals, but are rarely used in any analyses. Figure 1 depicts the universe of citation links that exist for researchers and identifies those that are visible in the ISI databases.

[8] The impact of recent forays into this field by other providers, such as Elsevier (Scopus database) has yet to be assessed, though administrators and analysts are watching these developments with interest.

[9] The ISI database 'Web of Science' does not allow the conduct of citation analyses for research evaluation purposes (technically or legally). To our knowledge, only the Leiden group has full access to the world citation data, provided by ISI in a form that allows them to undertake the full range of bibliometric analyses for research evaluations.

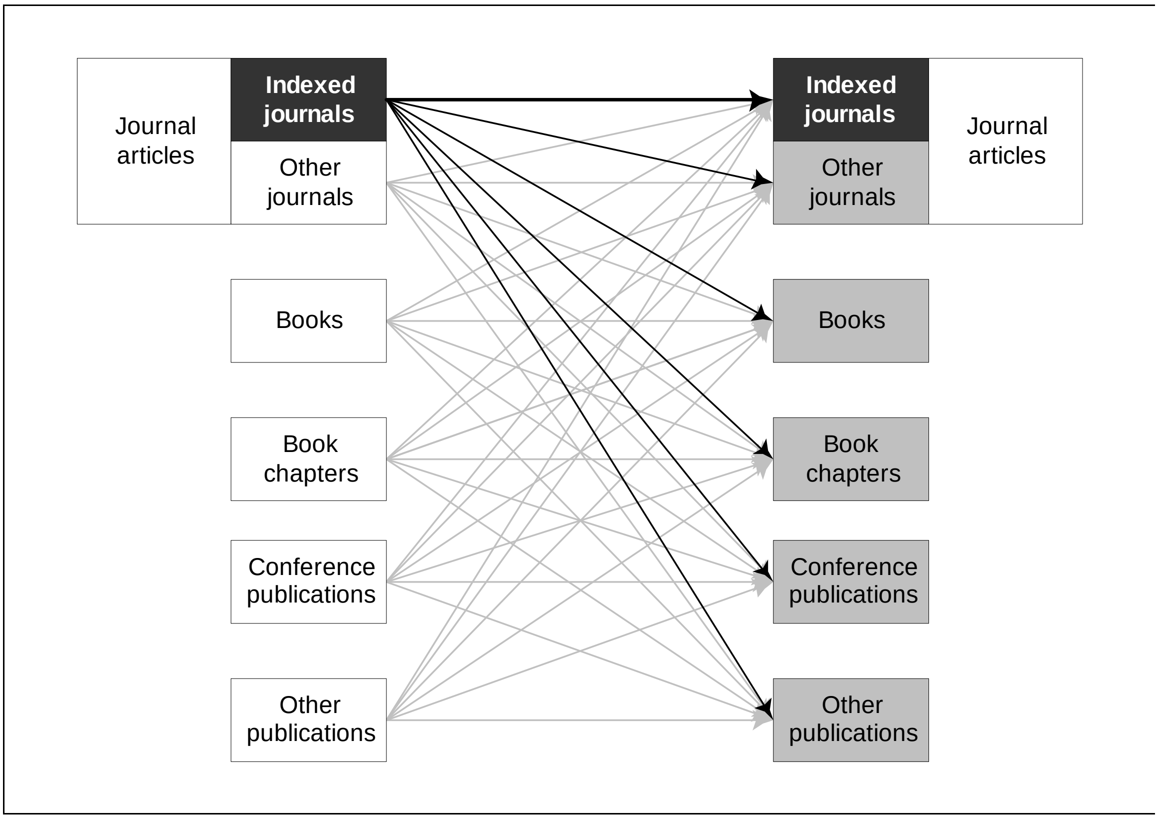

Figure 1: Citations visible in ISI databases.

All references contained in ISI's source journals are recorded in their database, but only those between source journals (the bold solid line) are linked and can be easily used in evaluative studies. Those from indexed journals to non-source items (the bold broken lines) can be identified, but a daunting amount of manipulation is required before they can be used in any analysis. References contained in non-source items, to whatever type of publication (the grey solid lines), are currently unusable in citation analyses.

Citations to non-source items, which could be extremely valuable in the social sciences and humanities, are rarely used in bibliometric analyses as a great deal of supplementary work is needed to compile author details, institution address, etc. Glänzel and Schoepflin automated this process and studied all citations made in 1993 for journals in the Science Citation Index (SCI) and in the Social Sciences Citation Index (SSCI), covering 600,000 source documents and 10,000,000 references. In contrast to SCI citations, they confirmed the dominance of non-journal references in the SSCI (1999: 37).

3.2.3 Errors in databases

To correctly assign publications and citations to the evaluated units, the information on which this assignment is based – addresses, author names, and references themselves – must be accurate. Large bibliographic databases typically require a considerable amount of data 'cleansing' before they can be confidently used. Most studies make little or no attempt to unify address variations below the country level (Moed *et al.* 1995: 390). At the level of departments, address variations can be huge and this requires time-consuming manual editing (Bourke and Butler 1996: 3-6). According to Van Raan, the address purification process is a heavily underestimated problem in bibliometric analyses (Van Raan 1996: 403, 405).

If individuals (or groups of researchers) are the subject of analysis, author names need to be cleaned and unified (Van Raan 1996: 402).[10] Misspellings, maiden names, inclusion or otherwise of second and subsequent initials, common names shared by many researchers (homonyms) are just some of the problems faced in identifying authors (Phelan 1999:131).

The reason for this statement is that a considerable amount of errors still occur in the citing-cited matching process leading to a 'loss' of citations to a specific publication. On average, the number of non-matching references — although they are citation index covered source papers — is about 7% of the citations matched (Moed *et al*. 2002). Frequently occurring non-matching problems relate to publications written by 'consortia' (large groups of authors); to variations and errors in author names particularly — but certainly 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 *et al*. 2002). So if the citation indexes are used for evaluation purposes, all these possible errors have to be corrected as much as possible.

3.3 Bibliometric indicators – technical decisions to be made

After tackling all the general issues relating to the application of quantitative indicators, and assessing the implications of the choice of database to use in constructing them, there are still several additional technical matters to be addressed before any indicators can be applied with confidence.

3.3.1 Types of publications to include

A decision has to be made about what to treat as a research publication. Electronic databases such as the ISI contain many different types of publications — in addition to (original) articles, reviews, notes and proceeding papers, they also contain meeting abstracts, corrections and editorials, letters of correspondence to the editor, etc. This latter group are generally not rated as major modes of communicating and are excluded from analyses (Martin 1994: 30; Moed *et al.* 1995: 394, 398), though there is some debate on whether to include the 'letters' category – CWTS measures include it, while ISI, REPP and others exclude it.

In the case of self-reporting, rather than database extraction, decisions also have to be made about which types of publications to include. Should every possible type of publication be included, or should it be restricted to a sub-set of the most 'important'? One assessment of international influence included journal publications with international referee boards, proceedings of international conferences and books with ISBN numbers (De Bruin *et al.* 1993). The funding formula used in Australia to distribute research funding incorporates four publication types – books, book chapters, refereed journal articles and refereed conference papers (all with quite restrictive definitions), though at one time data was collected for 22 different types of publication (Bourke and Butler 1996: 13).

[10] This is always necessary whatever ISI product is used. Also, the special tools provided by ISI for evaluations (e.g. National Science Indicators or Essential Science Indicators) have not undergone an address and author purification.

In the recent peer-based evaluation of the Australian National University scientists could choose their five best research outputs without being restricted to certain types of output. Beyond the usual types mentioned above, the scientists frequently selected research reports/working papers, reports for external users, encyclopaedia entries and a huge variety of different kinds of creative works (Gläser and Laudel 2005). As expected, these latter types were especially important in the social sciences, arts and humanities. Similarly, for evaluations in linguistics and law a much broader variety of publication types must be included (Luwel *et al.* 1999: 30-31). Of course, this large variety creates further problems, such as how to compare these different types and whether to include arbitrary weights, for which no clear solutions exist.

Luwel *et al.* (1999) include an extensive range of publication types in their study of law and linguistics which extends far beyond the coverage of the standard ISI literature. This study found that 41% of law publications and 55% of linguistics publications were not in the serial literature, and that only 19% of all linguistics publications were in refereed journals (no assessment of this was made in law). What we may term the grey 'enlightenment' and 'non-scholarly' literature is an important output in both cases (1999: 64, 156), although this is largely overlooked by bibliometricans[11].

3.3.2 Selecting the period

Indexed databases generally span many years. ISI indexes contain data that is of sufficient rigour for citation analysis back to 1981. The choice of time frame and citation window are two closely related topics, though they approach the time issue from slightly different perspectives.

Time frame

Moed *et al.* note that the period covered in any analysis will depend on the aims of the study (1985a). They distinguish between short-term and long-term impact. Short term impact focuses on the impact of research in the first few years after results are published. However the choice of time frame to include can be critical, particularly for emerging fields and groups. In one recent study, the bibliometric indicators showed only a moderate impact (moderate citation rate), while the peers who were also assessing the unit stated that it had recently made important contributions to the field. These would have become visible in the citation data in later years (Van Leeuwen *et al.* 2002).

An analysis of long-term impact can provide insights into a group's lasting contribution to their discipline. This necessitates a longer time frame as research groups need time to establish their position. Van Raan claims even periods of five years are too short (1996: 403). Researchers from science fields suggested two generations of PhD students (about eight years) is the absolute minimum (Moed and Van Raan 1988: 185), while others advocate an even longer time period of ten years (Johnes 1988: 63, Van Leeuwen *et al.* 2003: 258). A long time-span has the

[11] A full list of publication types is given in the report of the study. A few illustrative examples are provided here. In law, the field chosen to represent social science, a few of the more uncommon types identified were: teaching course notes; juridical publication for a wide audience; and scholarly edition of Law codes, jurisdiction volumes, bibliographies.

In linguistics, chosen to represent the humanities, they identified an extensive list of potential publication categories, and divided them into scholarly outputs and publications intended for a wider audience. A few more novel examples of the two types are listed here. Scholarly publications: Annotated scholarly version of a text; substantial scholarly contribution of Lexicons; and published scholarly software, or multi-media applications. Publications for a wide audience: article in a newspaper, a trade journal or a non-scholarly periodical; contribution to TV or radio programs; edition of linguistic or literary texts.

additional benefit of reducing random factors and increasing the substantive reasons for being cited (Nederhof 1988: 207).

The decision to focus on short-term rather than long-term impact is often made on pragmatic grounds (Moed and Van Raan, 1988: 181). A shorter time-frame means more recent publications can be incorporated in the analysis. A longer-time frame has less relevance for evaluations that seek to assess current or recent performance.

Citation window

A critical issue for studies of short-term impact (or those that focus on relative performance at a given point in time) is the choice of citation window, that is, the time span in which citations to a publication are counted. In determining the width of the citation window, one faces a trade-off between completeness and being up to date. The wider the citation window, the more comprehensive is the measurement of a publication's total impact. However, the wider the citation window, the more publications must be excluded because they are too young to enable this window (figure 2).

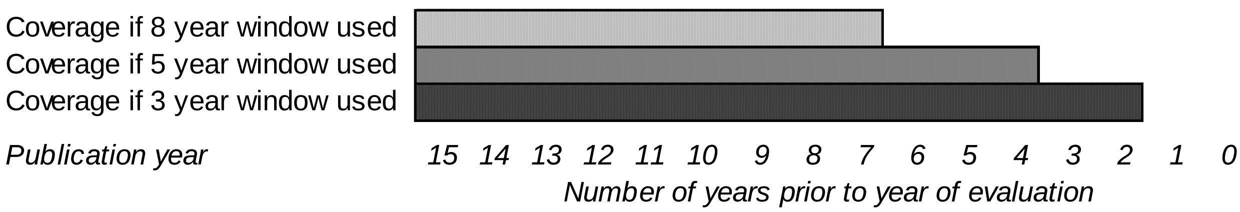

Figure 2: The effect of selecting different citation windows

A common aim is to cover the period up until citations for the publications reach their peak. In many natural science disciplines, this is reached after three years (Moed *et al.* 1985c:136; Van Raan 1993: 154) and in these fields a period of three to five years is common and appropriate. Where this is not the case, a longer period is needed. The time-span until the citation peak is reached can also depend on the type of the work, for example, biological and biomedical experimental work is rapidly cited and citations decline more quickly whereas work in other fields such as classical biology or high-energy physics is much slower in being cited (Moed and Van Raan 1988: 187-188).

There are varying views about the appropriate length of citation windows for the social sciences and humanities. Glänzel chose to use three-year moving citation windows as a middle ground between a longer observation period that he felt was more suited to social science citation practices and the danger of studying obsolete literature (1996a: 293). Nederhof and Van Raan decided upon a three year citation window for economics research which includes self-citations, and a four year citation window when excluding self-citations. Like Glänzel, they maintain that '[f]or policy purposes, the shorter three-year window is more relevant as it allows assessment of more recent publications. However, a longer window involves greater numbers, and thus offers a superior basis for analysis' (Nederhof and Van Raan 1993: 256). Van Raan believes that '[f]or social sciences, this window should be longer than in the natural sciences, and around five to six years' (1998: 4; 2003: 25). The consensus therefore is that the citation window should be between three to six years for the social sciences.

Trend analysis

Whenever trend analyses of citation rates are carried out, it is necessary to ensure that the period in which a publication can attract citations is constant across time. Distortions can occur if older publications have a longer time-period in which to attract citations than more recent ones. There are two ways of addressing this issue:

(i) Fixed-length overlapping windows: Analyses are undertaken using overlapping windows of a set length. Publication years are progressively dropped from and added to the analysis across the period for which data exists. For example, a time series analysis based on five-year windows would plot points for 1981-1985 (i.e. publications appearing between 1981 and 1985, and the citations these publications achieved in the same five years), 1982-86, 1983-87, and so on. This allows data up to the most recent year to be incorporated into the analysis, and has the added benefit of reducing the volatility of yearly citation rates when small units are being analysed (Butler 2001).

(ii) Fixed citable period: Analyses calculate the number of times publications are cited during a given number of years after they appear (Moed *et al.* 1995:394). If the citation window is too large, younger publications cannot be included and the analysis is of little value for science policy because they usually want information about the recent performance of their research organisations (see figure 2 above). If the window is too small, the validity of citation measures is compromised.

3.3.3 Field delineation

The controversy about Martin and Irvine's evaluation of radio astronomy centres has shown how crucial it is to have the correct field delineation (Martin *et al.* 1985; Gouguenheim 1985; Gillmor 1985). Whatever unit of analysis is chosen, the field-dependent publication and citation practices must be taken into account.

The ideal scenario for field of research classification is to assign a classification to each publication. At higher levels of aggregations, where hundreds or thousands of publications are involved, such a task is generally impossible. Three methods have been used for overcoming this problem. One method is to classify research according to the address of the author, that is, trying to assign a field of research code to the author's department or research group (Bourke and Butler 1993:9). However, investigations show that there is an imperfect 'fit' between the field of research denoted by the title of university departments and the range of fields of research housed in them (Bourke and Butler 1998). As a result, most practitioners chose a second option, classifying publications on the basis of the journals that carry them.

This method has it problems, but there is still nothing better available in terms of an automated procedure (e.g. Van Raan 2000: 317; Van Leeuwen *et al.* 2003). While custom journal sets can be created as needed, the journal classification scheme most commonly used is that drawn up by ISI which classifies its journals into about 200 subject categories. Problems exist when categories are relatively broad and cover more than one discipline (Moed and Hesselink 1996: 822). Scientists have expressed concerns that the wide-ranging SCI fields might lead to the under-estimation of the impact of publications from these fields (De Bruin *et al.* 1993: 39). Since a journal often covers more than the discipline, and may be assigned to more than one journal set, it is not precise for single publications (Glänzel 2003: 38-39) and can lead to problems of double-counting in multi-disciplinary analyses (Bourke *et al.* 1999: 15). Some journals such as *Nature* and *Science* cannot be classified to a single field – they are assigned to a 'multidisciplinary sciences' category.

In an attempt to overcome the deficiencies of journal set analysis, more sophisticated methods have been developed, such as cluster analyses based on co-citations (e.g. Small 1977) or filters for key words as well as words from titles, abstracts, or addresses (Noyons and Van Raan 1998; Lewison 1999; Noyons 2001). So far, these methods are only used for mapping research fields and for small-scale evaluation purposes (e.g. Lewison 1998).

Another fundamental technical problem is the fluctuation of ISI's journal sets which change annually by about 5 to 10% — journals are excluded; new journals are added; new categories are created; and journals are assigned to different fields in the ISI classification system. A number of analyses attempted to overcome the problem using 'fixed' journal sets. However this led to problems in interpreting long-term trends (Weingart *et al.* 1988: 397; Van Raan 1993: 163) and most analyses now use 'variable' journal sets.

3.3.4 Identifying the need for reference values

Citation rates vary enormously among different scientific disciplines and in order to construct meaningful indicators, citations need to be standardised to take this into account. This can be achieved by comparing them with (international) reference values for all publications (or journals) from the same discipline. A second reason for normalisation is that the citation rate is increasing worldwide (Van Raan 2000: 317) and trend analyses need to control for this. These pressures have led to the introduction of a class of indicators, of which ISI's Journal Impact Factor is the most commonly used. We term these indicators 'reference variables' and describe them in detail in Section 4.2.1.

Cross-field comparisons are not the only reason for normalising indicators by use of reference values. Problems can also arise in the interpretation of comparative performance within a single discipline: can differences be considered as 'significant' with a certain probability or should they be ascribed to mere chance (Moed *et al.* 1985c: 142). Statistical tests have been developed in order to check if a citation rate differs significantly from the world average (Schubert and Glänzel 1983). However, if one takes into account that a great deal of background knowledge is necessary to interpret the results of bibliometric evaluations, statistical tests might be misleading. If the numbers of publications are small, only large differences are significant (Nederhof and Van Raan 1993: 334-335).

3.3.5 How to handle multiple authorships

A very extensive review of the problem of multiple authorship in bibliometric analysis is covered by Harsanyi (1993). We cover below some of the issues that were addressed there, and in more recent discussions of the issue.

Scientists have stated that their own contribution to a co-authored publication was sometimes very minor. If such papers are highly cited, it would be wrong to credit them with this performance (Laudel 2001: 12; Van Raan and Van Leeuwen 2002: 630; Gläser *et al.* 2004: 30). However, as publications increasingly span the boundaries of the evaluated units, it becomes progressively more difficult to justly attribute the results of collaborative efforts to the units that took part in the collaboration. (Gläser *et al.* 2004: 20). Three main methods are used to cope with multiple authorships:

- Whole counting , i.e. each author (or group, or institution, etc) receives full credit;
- First author counting or straight counting, i.e. only the first author, or the first listed group or institution, is counted. This is used when it is assumed the first author has made the greatest contribution (Cole and Cole 1973: 32-35), or where the database being used only indexes

details of the first author (e.g. *Medline*). It can also be considered as a sampling strategy, though one that should be examined for its accuracy (Verbeek *et al.* 2002);

- Fractional counting: each of x authors (or units, or institutions) receives a score 1/x (Lindsey 1980), the assumption being that each has made equivalent contributions to the research. A variation of fractional counting sees authors being given varying fractions based on their position in the author list. One example is where the last author getting most of the credit, the other authors less. The assumption is that the last author has contributed the most important part of the work (Zuckerman 1968).

While fractional counting has some appeal, the assumption that a co-author's contribution can be revealed by his or her place on the list has been discredited. Co-authorship rules, and rules about the order of co-authors, exist but they are not as simple as previously suggested and they differ from field to field (Harsanyi 1993: 337). Furthermore, these rules may be superseded by local reward practices and ad-hoc negotiations (Laudel 2002: 9-13).

3.3.6 Self-citations

The debate on the validity of citation measures broached the issue of whether self-citations should be included in the calculation of performance measures. A self-citation is defined as "a citation given in a publication for which at least one author (either first author or co-author) is also an author of the cited paper" (Van Raan and Van Leeuwen 2002: 616). The significance of this issue is influenced by factors such as the research field, the type of publication, the age distribution of publications, the size of the aggregation level, the number of articles, and the extent to which articles are cognitively interrelated (Van Leeuwen *et al.* 2001: 197).

There are basically three methods of handling self-citations:

a) *Ignoring them*: This course of action is generally used as it is argued that excluding self-citations is a costly procedure and that self-citations usually indicate that researchers build on their previous work and therefore justifiably cite it (Katz and Hicks 1997: 542). The second justification becomes even more reasonable when the growing number of co-authors, often working worldwide in different laboratories, is taken into consideration.[12] A self-citation then becomes a measure of international impact — an indication that the cited knowledge is used in different research laboratories worldwide.

b) *Omitting them*: If small citation windows are used (3 years or even less), self-citation will be a larger problem because they are more prevalent in the first year after publication. However, self-citations cannot always be omitted because some relative citations indicators do not work anymore, e.g. indicators that use average field citation rates for normalisation. (Aksnes 2003: 242-244). More importantly, an exclusion of these citations as self-citations would lead to an artificial construct of impact. Aksnes points to this problem and suggests only to count first-author self-citations (2003: 244). However, referring to our discussion about multiple author counting (Section 3.3.5), this solution must be rejected.

c) *Noting them*: Rather than exclude self-citations entirely, some analysts advocate calculating the proportion present and using it as a control variable. The significance of self-citation rates typically varies for different fields. For example, Aksnes found a self-citation rate for clinical medicine of 17% and for astrophysics 30% (Aksnes 2003: 242).

[12] An extreme example is the publishing behaviour of high-energy physics communities where practically each member of the scientific community appears as co-author on a publication (Krige 1993: 234).

> If the share of self-citations is particularly high (>50%) or particularly low (<20%), the calculated citation indicators get over- or under-rated (Van Leeuwen *et al.* 2001: 197).

In practice, analysts have little choice other than to use the first option. The technical ability to exclude or identify self-citations requires access to the full citation database, which few analysts other than those operating out of CWTS have.

3.4 The Social Sciences, Arts and Humanities – are standard measures applicable?

Many of the issues raised thus far are taken to apply equally to science, social science and the humanities. This section of the literature review uses the discussion in sections 3.1 to 3.6 as a trigger to reflect upon issues of validity, problems of method and the technical concerns that arise in the literature with specific reference to the social sciences or humanities. The dominant issue is one that is either at the forefront of deliberations or lurks implicitly in the background – whether bibliometric approaches suited to measuring research output or research impact in the experimental sciences can legitimately be applied to the social sciences and humanities. Put another way, the literature asks if ‘standard’ science (ISI journal-based) approaches capture the essence of those disciplines where ‘non-standard’ publications (books and other non-ISI literature) are important or even dominant? These and other salient issues are addressed in the order they were first encountered above.

The literature speculates about why bibliometrics have come to be applied to the social sciences and humanities, and there is consensus that the driving force has been the desire of funding agencies or policy-makers to evaluate the whole research sector. There has been ‘considerable debate’ between academics and research evaluators about the “usefulness of bibliometric indicators as an evaluative tool for the social sciences” (Katz 1999: 1) motivated by a need for “humanities and social sciences to develop methodological tools to assist evaluation agencies or policy-makers in carrying out their tasks, in the same way as current SCI-based methodologies provide supplementary research assessment tools in the natural and life sciences” (Luwel *et al.* 1999: 13). However, while Nederhof *et al.* (1989: 425) earlier acknowledge that “bibliometric tools have proved their usefulness as monitors of developments in the natural and life sciences”, they caution that “evidence on this point is almost completely lacking for the humanities and many of the social science disciplines”, concerns that persist today, as we shall see below.

The vexed question of coverage

Does the database in question cover the typical published output of a field? We saw above (section 3.2.1) that bibliometric indicators are appropriate where the database being analysed adequately covers the publications that are the principal carriers of knowledge. It was noted that the coverage of ISI databases is good for most science disciplines, but it is not so good for the social sciences and humanities due to role of ‘non-source’ publications, most notably books, monographs and book chapters. This is the most widely considered theme in the relevant bibliometrics literature, and one that underpins the very validity of applying science-based measures to the practices of the social sciences and humanities.

Hicks observes that the humanities are dominated by book publishing but there is currently no database that indexes books or their references, and that in the social sciences indexed journal publications sit alongside non-indexed book publishing and other non-source literature. It follows that while bibliometrics are a relatively straightforward means of assessing research in the experimental sciences, we find that “[w]hen challenged to evaluate scholarly work in the social sciences and humanities, we are rudely forced to work outside this comfort zone in a frankly messy set of literature" (2004: 474). She reiterates familiar concerns about literature in the social sciences (Katz 1999: 2) and to a greater extent the humanities (Cronin and La Barre 2004: 94) being comprised of non-source items that citation studies based upon the content of scientific periodicals will miss (Broadus 1971: 238-9; Butler 1988 cited in Hicks 1999: 196; Nederhof *et al.* 1989: 427; Nederhof and Van Raan 1993: 357; Burnhill and Tubby-Hille 1994: 148; Lindholm-Romantschuk and Warner 1996: 390; Cronin *et al.* 1997: 270; Glänzel and Schoepflin 199: 32-33; Hicks 1999: 201; Hicks 2004: 473, 475, 482).

Several studies also point to the existence of a 'grey' literature of scholarly works (extending beyond books, book chapters and non-indexed journals) which is similarly not covered by indexed databases and is often overlooked by bibliometricians. This includes conference proceedings, internal reports, reviews, notes and other unpublished scholarly material (Nederhof and Noyons 1992: 252; Villagrá Rubio 1992: 6; Nederhof and Van Raan 1993: 357; Rosenthal 1994: 86-94). A separate 'enlightenment' or 'non-scholarly' literature aimed at the general public is also identified (Nederhof *et al.* 1989: 424-5), and which Hicks (2004: 489) recognises as the 'fourth literature of social science' in the form of non-scholarly journals directed at non-specialists, newspaper contributions and specialist publications, works that have 'application' and are not likely to be picked up by SSCI – see also section 4.4.3 below where Luwel *et al.* (1999) note the importance of 'other' scholarly and non-scholarly publications in law and linguistics.

Several researchers do not believe that the limitations of indexed database coverage are necessarily problematic for bibliometric study of the social sciences and humanities. For example, Meertens *et al.* (1992: 95) are satisfied that ISI journals "represent the most important (social) science journals throughout the world" and are confident to generalise from their ISI-based findings. There is a widely endorsed view that using the A&HCI and SSCI is unproblematic because non-ISI citations are covered in the bibliographies of indexed journal papers (Endler *et al.* 1978: 1066; Nederhof and Noyons 1992: 255; Burnhill and Tubby-Hille 1994: 148; Van Raan 1998: 5), and ISI journal-based citation patterns are a sufficient surrogate for the citation practices of the social sciences and humanities as a whole. Godin insists that in the case of Canada "bibliometrics is as good a tool for measuring the social sciences as it is for the natural, biomedical and engineering sciences" (2002: 4), and Van Raan claims that we must beware of 'an all too easy acceptance of the persistent characterization of the social sciences (and the humanities) as being "bibliometric inaccessible"' (2003: 24) as 'linguistics and experimental psychology are more and more approaching the publication behaviour of the "hard" sciences' in respect of the central role of international 'core' journals, and the increasing citation of recent work (Van Raan 1998: 3), which makes them more and more accessible to bibliometric analysis.

However, a far greater proportion of bibliometricians hold that indexed database coverage for the social sciences and humanities is problematic. In a study of Australian university bibliographies, Bourke *et al.* found that by excluding non-source items (and due to the under-representation of Australian and regional journals), journal-based indicators in the social sciences and humanities used as a 'surrogate for total publications citation rates will be more misleading than in the sciences' (1996: 54), a sentiment echoed by Glänzel and Schoepflin (1999: 31) and Hicks (2004: 473). Most criticism is aimed specifically at the coverage of the A&HCI and SSCI. For example, Nederhof and Noyons (1992: 255) caution that "focus on article publications alone could thus lead to seriously flawed results", and Hicks maintains that by replicating practices suited to science publication 'bibliometric evaluation produces a distorted picture of social science fields' (2004: 473).[13] The strongest criticism is made with reference to the social sciences in Australia when Royle and Over argue "reaching conclusions about the relative research productivity of individuals or universities, or seeking to compare disciplines or nations, by reliance solely on the database used to compile the ISI source indexes is not a worthwhile exercise" (1994: 86).

The general consensus within the literature is that the application of bibliometrics to the social sciences and humanities remains a thorny issue precisely due to the limitations of indexed database coverage (see Hicks 1999: 196; Katz 1999: 2; Hicks 2004: 477). These concerns are heightened with respect to the relative international 'periphery' of Australian research in the social sciences and humanities. However, bibliometric study of the social sciences and humanities using the A&HCI and SSCI persists, and is conducted even by the strongest critics of ISI coverage. But why is this so? "[T]he bibliometric community must choose one database to use for mainline work, and no database even comes close to the [SSCI] in this competition" (Hicks 1999: 193). The SSCI is therefore "imperfect, yet unrivalled in the database world" (1999: 194).

To sum up, '[n]on-journal publishing is significant in the social sciences. Some have wrestled with this problem; others have acknowledged it' (Hicks 2004: 478). We may add that with respect to the social sciences and humanities, others have ignored it.

Indicators for the creative arts

In spite of being covered by the A&HCI, the creative arts is a vastly neglected area in the bibliometrics literature, but potentially provides the most ground-breaking area of study, challenging orthodox views of what counts as a publication in a non-bibliometric yet quantifiable sense. Strand discusses at length how we may think about defining publication in the case of Australian creative arts, and present the case that 'publication may take many forms, of which the written word is merely one' (1998: 55). He argues that to publish something means to make it publicly known, which is to present 'for public viewing, criticism, analysis and evaluation, an author's ideas, knowledge, stories or information' and in this way '[t]he author's ideas become available for public comments and testing, made available for public use, building on to previous knowledge' (1998: 53-4). This process can take a variety of formats, such as 'the production and distribution of books, reports, papers and such that consist of words, drawings and photographs. In the broader sense, however, publication of a work may entail hearing it, viewing it reading it or experiencing it in other ways' and it follows that a publication may 'be presented in ephemeral and less tangible ways, such as through performance on a stage or in some other public forum' (1998: 54). In this light we could say that the creative performance is the publication.

The wide variety of potential publication formats can be illustrated in several ways. For example, in the case of music, a text is something that can be read, 'whether it is in visual, audio or written forms' (Strand 1998: 54) and an opera 'could be performed on stage, recorded on compact disc and filmed all at the same time' (1998: 55). Strand borrows an example from Webber (1994) and asserts that in architecture the publication is the actual building and not photographs of it or written descriptions of it in journal papers. Strand emphasises that a narrow view of evaluating publication in the creative arts in written terms alone creates anomalies whereby a painter's paper about their own exhibited painting counts as a publication but the

[13] For further bibliometricians' concerns about A&HCI and SSCI coverage see Nederhof and Zwaan (1991: 323-33), Nederhof and Noyons (1992: 255), Burnill and Tubby-Hille (1994: 148), Glänzel (1996a: 292), Cronin *et al.* (1997: 264, 270) and Hicks (1999: 197, 210-2).

painting does not, and a critical paper on a musical composition counts as publication while a performance of the composition, and even the composition itself, do not count as publication.

3.5 Applying patent indicators – a thorny field

There is no agreement in the literature on whether to consider patent-based indicators as bibliometric indicators or not. However, patents exhibit many of the characteristics of publications. They, too, are an outcome of research activities (Weingart and Winterhager 1984: 89) and publish scientific and technological knowledge (the exclusiveness of patents refers to the use rather than distribution of the knowledge they contain). Patents also cite (publications and other patents) and are cited (by publications and other patents). Thus patents can have functions in the knowledge production process which are similar to those of publications. For this reason, and following Weingart and Winterhager (1984), as well as Narin, one of the major developers of patent indicators (Narin 1994; Narin and Hamilton 1996), we regard patents as bibliometric indicators.

It has been mentioned before (Section 3.2.1) that bibliometric indicators based on publications can be a poor reflection of research activities in applied research areas. Patents are regarded as a much better indicator of the output in these disciplines. They are now becoming important even in fields with fuzzy boundaries between basic/applied research (e.g. biosciences).

Patent indicators are often used to measure the economic or innovative strength of a country in a certain area; thus, many analyses are undertaken on the macro-level in a cross-country comparison. Unfortunately, the literature which deals with patent indicators in relation to lower levels of aggregation, particularly for research groups or organisations, is scarce.

The most simple patent indicator is the number of patents. This indicator comes in two variants, either as 'granted patents' or as 'patent applications'. The latter version is a marked departure from publication-based indicators where nobody suggests measuring 'submitted publication manuscripts'. In the case of patents, analysts justify resorting to applications on the grounds that a patent application is a costly and long-term process[14], which implies that the applicant (supported by patent agencies) will be careful in ensuring that new knowledge was in fact produced. This is further encouraged by the patent regulations:

> The applicant must show that the invention is novel, useful, and non-obvious to someone with average expertise in the same industry. To do so, the inventor will cite earlier patents, and explain why the new patent improves on the earlier inventions. ... It is fraud on the patent office not to cite earlier relevant work, and it is also undesirable to cite irrelevant work unnecessarily (Narin *et al.* 2004: 555-556).

Another reason for using patent applications is the huge time-lag between application and award of patents (up to six years), which makes it nearly impossible for research evaluations to use 'granted applications' as an indicator of current or recent performance.

When using patents or patent applications as an indicator, it is important to keep in mind that the connection between the research process and patents is not as immediate as that between the research process and publications. Hornbostel (1999: 68) cautions that patent applications are

[14] For example, the European Patent Office took more than 49 months for about half the patents granted in 2002, and after 70 months the granting procedure was only completed for about 80% of the patents (Hinze and Schmoch 2004: 216).

aiming at a commercial use and that the application process is costly. This brings in intervening factors other that research quality. He recommends scrutinising what exactly what the indicator measures (without going in further detail).

As with publications, the research quality inherent in a patent is not uniform. Patent citation counts have been used to discriminate between patents of higher and lower quality. "The underlying assumption in patent citation analysis is that a highly cited patent (a patent which is referred to by many subsequently issued patents) is likely to contain technological advances of particular importance that has led to numerous subsequent technological improvements." (Narin *et al.* 2004: 556). Patent citation analyses may refer to both citations of other patents (patent-patent bibliometrics) or to citations to scientific publications (patent-paper bibliometrics, Hicks *et al.* 2002: 42-48). With an increasing number of references to scientific publications, the latter analysis has become more important.[15]

Several studies have validated patent citation counts by peer review. For example, it has been shown that patents rated highly by industry's staff were more frequently cited than patents of lower rank (Albert *et al.* 1991) and thus citations to patents can serve as an indicator of the quality of patents. One crucial difference between citing of publications and citing of patents exists. Whereas one can be confident in the former case that the citing actor is the scientist, in the latter case it has been shown that citations are given not only by the inventor, but also by the patent examiner and patent attorneys of the organisation (Vinkler 1994;Meyer 2000).

Another of the characteristics that patents share with publications is that the distribution of their citations is very skewed: a large number of patents is cited only a few times, and only a very small number of patents is cited more than once (Hicks *et al.* 2002: 33).

Hinze and Schmoch lament the weak methodological discussion of patent indicators (in contrast to other bibliometric indicators). This leads to black box approaches and irreconcilable results. They made a first step in the direction of overcoming the major problems: how to cope with the varying practices of different patent offices; changes in practices over time; the identification of what should be counted as a 'publication'; how to avoid double counting; and how to deal with the special problem of cross-country analyses (Hinze and Schmoch 2004). Hinze and Schmoch's description of all the traps into which one can fall highlights the need for extreme caution in conducting patent analyses. This is confirmed by the fact that patent analyses became a specialised area within bibliometrics, suggesting that expert knowledge is necessary to conduct them.

As for publication and citation analyses, large databases are used for patent analyses. While for the former usually one database is used (the Citation Indices, provided by ISI), numerous patent databases exist (mostly provided by national and transnational patent offices).[16] The advantage is that access to these databases is often easier than to publication databases where only few institutions have the data in a form that allows for automated manipulation. Each patent contains classification codes which can be used for field delineation. Depending on the aim of the study, several databases might be used in combination, though this leads to additional problems of data analyses (see Hinze and Schmoch 2004: 222-227). The most frequently used databases are that

[15] So far, it is mainly used for studies of technology transfer from science rather than to measure research performance. Although we observe this large increase of cited publications in patents, the question is whether their number is so high that they should be included in citation studies of publications.

[16] Databases for special fields exist, e.g. *Derwent Biotechnology Abstracts* which contains patent as well as publication data. ISI databases can also be used for patent analyses.

of the United States Patent and Trademark Office (USPTO) and the European Patent Office (EPO).

The application of patent indicators is limited to analysis of the natural and applied sciences. For the social sciences and humanities, no pendant to 'patent' exists.

3.6 Other quantitative indicators

Studies identifying possible performance indicators for research beyond the standard bibliometic measures have been in the public domain for some time. The first extensive Australian review was canvassed in Bourke's review of the existing literature as it related to university assessment (1985). In a lead-up to the introduction of performance-based funding of research by the Australian government, the Australian Research Council funded a study to investigate the feasibility of using particular indicators. Survey responses were obtained from nearly 4000 researchers across a wide range of disciplines. Respondents were asked to select eight indicators from a list of twenty and rank them in order of their relevance as indicators of research performance in their field (NBEET 1993: 74). Unfortunately, though the original aims of the study included a detailed assessment of the favoured indicators, this phase of the study did not proceed.

The only comprehensive study we have been able to identify in the literature that has moved past a listing of possible non-bibliometric indicators and attempted to design and operationalise other quanitative indicators was that undertaken by CWTS in their study of law and linguistics (Luwel *et al.* 1999).

The following sections (3.6.1 to 3.6.5) cover the most common non-bibliometric indicators, though discussion is necessarily short as they are given little space in the reviewed literature.

3.6.1 External funding

External funding – funding obtained from sources other than the host institution – is one of the most common quantitative indicators of research performance.[17] It use as a performance indicator is appealing because the award of a grant is based on peer review (Gillett 1991: 254) and because it measures current rather than past performance (Hornbostel 2001). It seems easy to quantify by counting the number of grants or adding up the funds awarded. Finally, external funding is a performance indicator that is well accepted among scientists. When Australian scientists were asked which indicators are appropriate to measure research performance in their field, external funding was one of the preferred indicators (Tognolini *et al.* 1994).

Despite the widespread use of the indicator, very few studies have been undertaken to examine its validity. Its application rests on the assumption that success in attracting external funds depends only on the quality of a researcher or the research proposal. However, it has been shown that a funding proposal's success can depend on numerous factors that are not linked to quality and cannot even be controlled by scientists. Unless a country's investment in research is adequate, even excellent projects will be rejected. Fields that do not have any promise of applicability for industrial innovation will not be funded by industry nor by agencies that are devoted to the promotion of industrial innovation. For a comprehensive discussion of these and other factors see Laudel (2005b).

[17] It is often assumed that external funding is an input indicator of research (funds that enable research work) and therefore cannot be an output indicator (e.g.Phillimore 1989; Johnes and Johnes 1995). However, the widespread use of external funding as output indicator is based on the peer evaluations that accompany grant applications.

Thus, the current widespread simple use of external funding as a performance indicator needs further consideration and caution must be exercised in its application. Those using it as a measure must be mindful of possible 'Matthew effects': those who get the most external funding are likely to get even more, while others are excluded from the system. Citation indicators point the way to improving the current low validity of the indicator, through normalisation and statistical aggregation. However, some questions remain unanswered (Laudel 2005a).

3.6.2 Research student data

Indicators based on research student data also feature regularly in evaluations. They are part of the Australian formula for the distribution of research funds to universities. They are applied in the British RAE, and in the Dutch evaluations of university research (Geuna and Martin 2001). However, discussions of its use are largely ignored in the reviewed literature. Concern was raised that this indicator might reflect teaching rather than research (Phillimore 1989: 263). This concern is particularly relevant if undergraduate or non-research degrees are included. In one study, scientists from a number of natural science fields regarded the number of dissertations as a poor measure (Jauch and Glueck 1975: 70). The questionnaires used provided no scope for further explanation.

3.6.3 Indicators of esteem

Indicators of esteem encompass forms of acknowledgment by peers. They include prizes and awards; memberships of professional societies; editorial board memberships; membership of grant review panels; invited talks at international conferences; and guest visits. The use of these indicators in the context of assessment have been criticised because they measure past excellence rather than recent research performance (Phillimore 1989: 262). They have also been criticised because scientists believe these indicators penalise younger colleagues (Jauch and Glueck 1975: 69; Wood 1989: 240).

Such indicators are often used by the scientists themselves in their CV's as a sign of quality. In research assessments they serve to inform peer review (e.g. Fries 2003). These indicators usually require detailed background knowledge (about the prestige of prizes, journals, conferences, funding agencies) or further specifications (e.g. about the frequency of referee activities) and are not easily reduced to a simple count. These problems have been raised especially in the case of prizes (Cole and Cole 1967: 383-384; Hornbostel 1997: 208-211).

Awards and honorary degrees "are particularly important for many parts of the social sciences and humanities where bibliometric methods are more difficult to use" (Van Raan 1998: 5). But little research has attempted to operationalise them, a rare example being a comprehensive exploratory study applied to law and linguistics in four Flemish universities (Luwel *et al.* 1999). A detailed discussion of specific esteem indicators will be given in Section 4.4.2.

3.6.4 A new development – Web indicators

A fairly new discipline is webmetrics, in which measures are based on counts of web pages and links between web pages. A distinction is made between 'outlinks' (links *from* a web page) and 'inlinks' (links pointing *to* a web page, Ingwersen and Björneborn 2004: 339). In recent years, studies have been undertaken to investigate whether web indicators can be used for the assessment of research quality (for example it is part of the WISER project – see http://www.wiserweb.org). The basis of this is a formal analogy between a webpage and a scientific article, whereby a web page's outlinks are seen as similar to an articles reference list, and its inlinks as similar to citations of an article. The analogy has led to new terms such as 'sitations' and 'Web Impact Factors'.

Attempts have been made to discover correlations between link counts and traditional research productivity indicators. Most studies to date have found only weak relationships (for an overview see Björneborn 2004, chap. 2.4.2). The notable exception is Thelwall, who conducted two studies that found a correlation between Web-based indicators and indicators of research productivity. In one study, he compared peer review ratings given in the British RAE with one web indicator –the ratio of web pages with links to research-based web pages per researcher – and found a positive correlation (Thelwall 2001). In another study, Smith and Thelwall related inlinks to the research performance of Australian universities, measured by the number of publications and the amount of external funding (Smith and Thelwall 2002).

The whole field of research on web indicators has one gaping hole which is only just starting to be addressed in the literature. While with bibliometrics there has been considerable research on the role of a citation in the process of scientific production, this is conspicuously absent from discussions on webmetrics. There is no measurement model that links the indicators to the properties of research they are supposed to measure. Social actions such as 'creating a website' and 'creating links to websites' are highly ambiguous and poorly understood. They can be (and indeed are) performed by a multitude of actors, among which researchers are only one group. They can also have a variety of purposes, and it is by no means clear that the purposes related to the quality of research (if there are any) are present to an extent that enables statistical analysis. This is a stark contrast to the act of 'citing a publication': we know that by citing another publication, a researcher refers to a certain body of scientific knowledge.

Thelwall has outlined this need for a future research program which is essential before one can think of using web indicators for measuring research quality:

> The first and most fundamental problem of using Web link counts to evaluate online impacts is the issue of validity. Citation counts have been subject to extensive research over the past 30 years to get to the position where experts can use them productively, if great care is taken (Moed *et al.* 2002). Web links need the same kind of research to give evidence of a consistent relationship with quality at any given level of aggregation. This process has begun, but is not complete (Thelwall and Harries 2004).

In the absence of an understanding of the motivations that lead to websites and weblinks, no causal explanation for a correlation between web indicators and research productivity or research quality can be given.

3.6.5 Policy impact

Cave *et al.* produced a report for the UK Economic and Social Research Council (ESRC) which discusses potential performance indicators (PIs) for ESRC-funded research. They examine, but do not operationalise, the possibility of evaluating the relevance of social science research by using 'policy usefulness' as a novel non-bibliometric performance indicator. They believe that 'it should be possible to scrutinise key policy developments and identify any contribution of research and other disciplined enquiry to, for example, patterns of care and prevention in drug abuse or child abuse' (§9.1.5), and that "[t]his could be done by a process of identifying the relevant group of policy makers or practitioners and interrogating them about their knowledge of research results obtained and the benefits derived from it" (1988: iii).

4 Description of quantitative indicators

Throughout this review we have referred separately to bibliometric and non-bibliometric indicators, and will continue to do so in this section. *Bibliometric indicators* are based on the published literature in all its forms – journal articles, monographs, book chapters, conference papers, patent, and the like – and the references these publications contain. *Non-bibliometric measures* focus on other aspects of research performance, such as honours and awards, editorship of journals, keynote addresses, and many other indicators of esteem. They also cover rapidly expanding ventures into web-link analysis. Bibliometric indicators can be further sub-divided into three groups. The simplest indicators are restricted to an analysis of the *publications* produced by a research unit. These measures are primarily restricted to aggregate or size-normalised counts, though some measures have been developed to weight different types of output.

More complex bibliometric indicators are those that analyse the references authors make in their publications to other research, that is *citations*. Citation analysis is less frequently undertaken because it requires access to a database that contains details of the references listed in research publications. The most commonly used databases for this purpose are the indexes maintained by the Institute for Scientific Information (ISI), now part of Thomson Scientific, which include the Science Citation Index (SCI), the Social Sciences Citation Index (SSCI) and the Arts and Humanities Citation Index (A&HCI).

A third group of bibliometric indicators is *structural* indicators. These indicators are not strictly performance measures, but give additional contextual information about the research activity being undertaken. They are used in combination with other indicators.

We also present a group of indicators that we describe as *control variable*. Whenever comparative evaluations are being undertaken between units that are not strictly analogous, additional data is desirable to overcome the existing disparities. The most commonly used indicators of this class are those that are used to normalise for the size of the unit, or for the different citation practices across research fields.

The indicators presented in this section are presented in form of tables which cover the following information:

- *Description:* A brief description of the indicator;
- *Application:* Where applicable, we note whether the indicator is currently used in evaluative studies by our own unit (REPP – the Research Evaluation and Policy Project, at The Australian National University), or by the Centre for Science and Technology Studies (CWTS) at the University of Leiden. We specifically refer to this latter group as it has not only made major contributions to the design of advanced bibliometric indicators, but also regularly applies them in research evaluations at all levels of aggregation. Those measures first described in the 2003 publication of Van Leeuwen *et al.* are designated their 'new' indicators.
- *Calculations*: This section specifies how each quantitative indicator is computed, and the required elements for the calculation;
- *Variations*: Many indicators referred to in the literature are only a variation on the more standard REPP and/or CWTS indicators, and are noted as such rather than listed separately;

- *Issues noted in the literature*: Any aspects of our assessment of indicators in the final phase of our project that have already been addressed in the literature are noted here e.g. comments on validity, fairness, transparency, field-specific issues, etc.
- *Data requirements:* The final issue covered is the source of data for the construction of the indicator.

A general problem is that the indicator descriptions in the literature are patchy. Indicators are usually presented as a rather unconnected list, and only in a few cases are the advantages of one indicator in comparison to other indicators discussed. The authors seldom reveal why they applied one indicator in preference to another. The way an indicator is calculated is sometimes not clearly described and the list of conditions for using an indicator is not comprehensive. We hope to fill these gaps in the course of this project. Our description of bibliometric indicators may not be complete, but it should at least provide an overview about the state of the art.

4.1 Publication Indicators

4.1.1 Generic indicators

Indicator 1. Number of publications	
Description	This indicator covers the published output of research. It is very comprehensive and flexible as it can include a large variety of publication types. It is the most commonly used indicator because it is applicable to nearly all fields and does not rely on external databases.
Calculation	There are three variants of indicator construction: 1) all publication types are counted together for a given time-span: P_{total} = number of books + number of journal articles + number of book chapters etc., 2) different publication types are counted separately: P_{books} = number of books, $P_{articles}$ = number of journal articles $P_{bookchapters}$ = number of book chapters etc. 3) different publication types are weighted and one publication index is constructed: e.g. $P_{total\_weighted}$ = number of books * 5.0 + number of refereed journal articles * 1.0 + number of book chapters * 0.8
Issues noted in the literature	*Publication types*: a decision on which publication types to include must be made beforehand. The debate on this issue centres on what counts as the output of research work. Problems occur if so many different publication types exist that the results become too complex for comparative analyses. One solution is to weight publications according to their value for the scientific community or other recipients. *Weighting*: The weighting given to different publication types is an arbitrary decision. The choice of weights can heavily influence the evaluation results. See the list following this table for some of the methods that have been proposed for weighting publication counts in the social sciences and humanities. *Attribution:* a decision on whether to apply whole counting or fractionation needs to be made. If fractionation is used, the allocation method needs to be determined. *Validity*: simple publication counts are often criticised for their poor validity. They reward 'mass producers' (scientists who publish a lot but who are rarely cited, Cole and Cole 1967) which is not necessarily desirable. This negative effect can be compensated for by the careful selection of publication types, for example, by including only articles in refereed journals, or by a weighting system which gives publications which have undergone a peer review process a higher weight. Additionally, several indicators were constructed to 'qualify' publications, i.e. to improve their validity.
Data requirements	Data must be collected from academics' CVs or the research reports of institutions and research groups. No central databases exist that carry all the necessary information.

As there is less support for citation analysis in the social sciences and the humanities, many methods for weighting publication counts (some of them curious in the extreme) have been proposed as a way of differentiating the output in these disciplines. They include:

- Number of books or proceedings weighted by prestige of publisher (Luwel *et al.* 1999: 28, 77 for law) and (1999: 164-5, 210-212 for linguistics) including weighting calculations. In peer feedback, one linguistics respondent "was afraid that classification would lead to an undue preference for established publishers and would discourage publication in new journals or with young, starting publishers" (1999: 216). See also Nederhof *et al.* (2001: 264).
- Number of journal publications weighted by peer perceptions (Luwel *et al.* 1999: 93-105 for law and 194-197 for linguistics). Not limited to ISI papers. In peer feedback, one linguistics respondent said "A fear is that the present results might be used "against scholars" by eliminating lesser-qualified journals from library collections" (216). See also Nederhof and Zwaan (1991) and Nederhof *et al.* (2001: 264).
- Number of papers classified as scholarly, applied or popularising (Luwel *et al.* 1999: 65-67). Law journals were classified on the basis of survey data relating to substantial contributions, small contributions, publications aimed at a wide audience and 'other' publications in these journals. Calculations (1999: 66). See also Nederhoff *et al.* (1989) for number of 'enlightenment' (rather than scholarly) publications.
- Number of publications in English and other languages Luwel *et al.* (1999: 28).
- Length of publication Luwel *et al.* (1999: 28, 67-69) – measuring publications in terms of the number of written pages produced, which favours doctoral theses and books which we are told constitute the core scholarly publications in law. This focus on length reduces the importance of journal articles (1999: 65). In linguistics monographs are favoured, reducing the importance of refereed journal papers which are in sixth place for scholarly publications behind monographs, substantial contributions to lexicons, edited books, external research reports and inaugural lectures (1999: 161-162).
- Publications with (semi-) identical titles (Luwel *et al.* 1999: 69). Identifies identical publications in different sources. (1999: 69-71 for law and 162-164 for linguistics).
- Number of papers with a colon in the title. (Dillon 1981a; 1981b) found that 72% of papers in social science scholarly journals (psychology, education and literary criticism). had a colon in the title, compared with 13% of non-scholarly journals. Correlation also holds for scholarly productivity, distinction, complexity and scholarly character.

Indicator 2. Number of funded papers	
Description	The theory behind this indicator is that externally funded papers have a higher impact because of the ex-ante evaluation of the projects (Martín-Sempere *et al.* 2002: 153).
Calculation	P_{funded} = total number of publications acknowledging a funding source.
Issues noted in the literature	*Validity:* This indicator assumes that authors acknowledge each external funding source that contributed to the research referred to in a publication. Although studies have shown that, at least in the UK, authors are diligent about funding acknowledgements, the practise is not perfect (Jeschin *et al.* 1995). In addition, fields and types of research (e.g. theoretical research) differ in the extent to which they rely on external funding. Furthermore, how much a research group depends on external funding relates to the extent of recurrent funding. This is another problem with using 'external funding' as a performance indicator (see 3.6.1).
Data requirements	This indicator requires the identification of publications linked with a particular funding source(s). There are two main ways in achieving this: i) Accessing the full text of the publication and extracting details of acknowledgements; or ii) Obtaining publication lists direct from the funding agencies. Either method is very time-consuming.

4.1.2 ISI indicators

Indicator 3. Number of ISI publications	
Description	This is the most basic of the indicators that are based on ISI data. It specifies how many research publications are produced in a given time-period. Usually, only publications which are classified by ISI as articles, letters, notes and reviews are included (e.g. Van Leeuwen *et al.* 2003: 259). Sometimes, other publication databases, for example "Chemical Abstracts" might be used.
Application	Standard REPP and CWTS indicator
Calculation	P = Number of ISI publications, published in the time period x.
Variations	The indicator can also be applied using other searchable databases. Similar indicator exist using specialty indexes such as *Medline*, *Chemical Abstracts*, etc. rather than the ISI indexes.
Issues noted in the literature	*Coverage*: the indicator is only suitable where ISI journals are the main carrier of knowledge. To find out if this condition is satisfied, two auxiliary indicators[18] are constructed (Moed and Hesselink 1996: 826): - P/Pjtot which is the share of ISI publications (P) in the total number journal articles (Pjtot). Since SCI publications are mainly journal articles, it indicates if the SCI covers the journals in which a unit is usually publishing; and - Pjtot/Ptot which is the share of journal articles in relation to the total number of submitted publications. It indicates the relative importance of journals as publication media.[19]
Data requirements	This indicator relies on access to the ISI database. Individual and group analyses can be done through queries using the ISI internet version, Web of Science, though analyses at higher levels of aggregation generally require access to the underlying data.

[18] Both auxiliary indicators also seem to be used as performance indicators by the Leiden group. They measure whether a unit publishes in the most important international journals in its research field(s); based on the assumption that the ISI database covers these journals. If the ratio is high, then a research unit publishes in international journals (Moed *et al.* 1995: 407-408). But the authors also admit, that these indicators are difficult to interpret because you first have to be confident that the ISI database provides a good coverage of the field to be evaluated.

[19] A general threshold value i.e. a percentage seen as sufficient to use SCI journal publications as the main output, was not found in the literature. A study of agricultural sciences found a coverage of 43% of the evaluation unit's publications which was assessed as not sufficient to use only ISI publications (Nederhof *et al.* 1993: 161).

Indicator 4. Number of ISI publications, weighted by journal impact	
Description	Publication counts are weighted on the basis of the impact or prestige of the journals in which they appear. The most commonly used weight is the ISI Journal Impact Factor.
Calculation	P*IF = number of articles weighted by the Journal Impact Factor
Issues noted in the literature	*Application*: It has been claimed that reviewers (e.g. those assessing grant applications) often use the quality of a journal as a surrogate measure for the quality of a scientist's contribution in it (Jauch and Glueck 1975: 74). *Underlying assumption:* To have an article accepted in a top journal, the author's manuscript has to undergo a robust peer review process. *Inappropriate use*: An unacceptable practice is to use journal impact factors in place of actual citation counts, and infer from this method the impact of a set of publications. The correlation between the citations that publications actually receive and the journal impact can be weak (Seglen 1994; Vinkler 1986: 162) due to the skewed nature of the citation distribution. Impact factors – whether ISI's Journal Impact Factor or impact factors calculated by other methods, should only be used as reference values or to weight publication counts.
Data requirements	The indicator only requires knowledge of ISI's journal coverage and access to the Journal Impact Factors for each of the journals, rather than to the underlying data. Both of these are readily accessible.

Indicator 5. Number of publications in top journals	
Description	Publication counts are limited to the top journals of a field (Van Leeuwen *et al.* 2003). The top journals are identified by setting a threshold value based on s special field-normalised journal impact indicator.
Application	CWTS (new indicator)
Calculation	A journal is defined as a 'top journal' if its JFIS > 3 (see Indicator 10 for a detailed description of the JFIS and its calculation). The absolute number of papers and/or the share of papers in top journals can then be calculated.
Variations	Rather than using impact factors, top journals can be identified by a panel of experts, though reaching an agreement on a list of journals is often a difficult task (Australian Academy of Science 1996)
Issues noted in the literature	*Constraints*: To cope with errors stemming from small numbers, two conditions were formulated: - only journals that received a specified minimum number of citations per paper on average (excluding self-citations) are included[20]; - only journals that contain more than 20 publications per year are included. Not all fields contain to the same extent journals with high impact scores. Therefore, no absolute threshold values (like JFIS>3) should be used. *Validity*: Impact factors do not say anything about the quality of individual articles, only about the impact of the journals as a whole. *Pros and cons:* The advantage of this indicator, with top journals identified using impact factors, is that the identification of the top journals in a field can be made 'objectively' without the involvement of experts. However it is necessary to arbitrarily define a threshold value in order to decide which journals are top journals.
Data requirements	ISI data on journals (Journal Citation Reports) is required for the calculation of the JFIS for each journal.

[20] For a study spanning 10 years, Van Leeuwen *et al.* 2003 set the threshold at 10 citations per paper. For analyses covering a shorter period, this threshold would necessarily be lower.

Indicator 6. Distribution of publications over journal impact classes	
Description	It indicates how the publications are distributed in good (highly cited) and less highly cited journals.
Application	CWTS (new indicator)
Calculation	Five impact classes are constructed, using citations and field reference values: JCS = number of citations received by a journal FCS = field-normalised impact factor, i.e. the average number of citations in a certain field (see Indicator 9 for a detailed description) JCS/ FCS = relative impact of a journal compared to the field average Class I: JCS/ FCS = 0 ; i.e. journals with no impact at all, Class II: 0.8 > JCS/ FCS > 0; i.e. journals with an impact up to 80% of the field average, Class III: 1.2 > JCS/ FCS > 0.8, Class IV: 2.0 > JCS/ FCS > 1.2 Class V: JCS/ FCS ≥ 2.0, i.e. at least double the field average (Van Leeuwen *et al.* 2003: 262)
Variations	REPP undertakes analyses using journal impact quartiles, rather than the five CWTS classes. The journal set for a field is divided into four equal groups on the basis of the journal's citation impact. The proportion of the unit's publications that appear in journals allocated to each of these quartiles is then calculated.
Issues noted in the literature	*Other influences:* The relative impact of a journal can be influenced by factors other than quality. For example, a journal focussing on basic research is likely to have a higher impact than one carrying articles reporting applied research (Narin 1976). This issue must be taken into account when interpreting any results.
Data requirements	ISI data on journals (Journal Citation Reports) is required for the calculation of the JCS and FCS for each journal.

4.2 Citation performance indicators

4.2.1 Reference Values

These indicators are usually not used separately in evaluations but they are incorporated into the calculations of other indicators (e.g. impact factors) or they are used as control variables for checking other indicators and interpreting them correctly (e.g. the share of self-citations). The first (Indicator 7) is the ISI Impact Factor which is widely used in research assessment. CWTS has developed three impact measures (Indicators 8 to 10) to use as reference values in the construction of more sophisticated indicators.

Indicator 7. Journal Impact Factor (IF)	
Description	The IF is published by ISI and is one of the most commonly used citation measures. It is a measure of the frequency with which the 'average article' in a journal has been cited in a particular year or period.
Calculation	B_x= Cites in year x to articles published in the proceeding two years C_{x-1-2}= number of articles published in years x-1 and x-2 IF_x= B/C = impact factor in year x
Issues noted in the literature	*Methodological weakness:* The ISI impact factor has been strongly criticised for its methodological weaknesses. The main arguments are: the inaccurate definition of the citable documents – they do not differentiate between the type of publication (review, article, letter etc.); they are based on citation windows too short for nearly all fields of science; different research fields cannot be compared (for a review of the discussion see Glänzel and Moed 2002). *Inappropriateness:* The literature cautions against the use of ISI impact factors for research evaluation. In addition to its methodological weakness, it conceals the difference in article citation rates that result from a highly skewed distribution (Seglen 1997:498).
Data requirements	ISI data on journals (Journal Citation Reports) is required for the calculation of the JCS and FCS for each journal.

Indicator 8. Expected Citation Impact	
Description	This indicator is used as a reference value to compare the citation rates of evaluated units to the average for all publications from the set of journals in which they publish. It takes into account both the type of paper (article, review etc.) as well as the specific years in which the group's papers were published (Moed *et al.* 1995: 399).
Application	REPP, CWTS (JCSm) standard indicator
Calculation	JCS = (number of citations received in the time-span y_1-y for all publications of type T published in year Y in journal j) / number of publications of publication type T in year Y in journal j Generally, a group publishes its papers in several journals rather than one. Therefore, a weighted average is calculated, with the weights determined by the number of papers published in each journal[21]: JCSm = ∑ (JCS) * (number of publications of publication type T in Year Y) / total number of publications
Issues noted in the literature	*Yearly fluctuations*: The average impact of journals may have considerable annual fluctuations (Moed and Van Leeuwen 1996). *Inappropriate use:* The indicator should only be used as a reference value, and never as a substitute for actual citation counts.
Data requirements	ISI data on journals (Journal Citation Reports) is sufficient if 5 year windows or publication year to date calculations are used. If other citation windows (e.g. two or three years) are to be used, then access to the full database is required.

[21] The indicator is similar to the Expected Citation Rate developed by Schubert and colleagues (Schubert, *et al.* 1989). The Leiden group claims its variant is more accurate since it takes account of differences in publication types. This is particularly useful at the level of research groups or departments (Moed *et al.* 1995: 409-410). The Leiden method is also used in REPP analyses, though in their studies the title 'expected citation rate' is retained.

Indicator 9. World Citation Impact	
Description	This is a second reference value used for analysing the citation performance of units. It represents the world average for all publications in journals classified to a particular field.
Application	REPP, CWTS (FCSm)
Calculation	FCS = (number of citations received in the time-span y_1-y_2 by all publications published in year Y in journals classified to field F) / total number of publications in year Y in field F
Variations	*Variation 1:* A unit is usually active in more than one field. If it is necessary to create a single indicator for the unit's total ouvre, rather than analysing each field separately, a weighted average can be calculated. The weights are determined by the number of papers published in each field. Where a journal is classified into more than one field, the calculation is based on the average of all the fields (Van Raan 2000: 317). *Variation 2*: Rather than basing the calculation on data for a single year (Y), the calculation can be done on a window of, say, five years.
Issues noted in literature	*Limitation*: This indicator delineates fields on the basis of journal sets, rather than identifying the field to which each individual publication in the ISI database belongs (a task most analysts find impossible to achieve). The field boundaries are therefore open to question.
Data requirements	ISI data on journals (Journal Citation Reports) is sufficient if 5 year windows or publication year to date calculations are used. If other citation windows (e.g. two or three years) are to be used, then access to the full database is required

Indicator 10. Journal to Field Impact Score (JFIS)	
Description	This is a variant of impact factor that serves as a reference value for comparisons. It describes the impact of a particular journal. Therefore it is a much improved variant of Journal Impact Factors provided by ISI (Van Leeuwen and Moed 2002).
Application	CWTS (new indicator)
Calculation	JI = (sum of citations to all publications of types $T_{1..x}$ in journal j for the period $y_{1...x}$) / (number of publications of types $T_{1..x}$ in journal j for the period $y_{1...x}$) FI = (sum of citations to all publications of types $T_{1..x}$ in journals classified to field F for the period $y_{1...x}$) / (number of publications of types $T_{1..x}$ in journals classified to field F for the period $y_{1...x}$) JFIS = JI/FI
Variations	*Variation 1*: The CWTS version limits the calculation to articles, notes, letters, and reviews, however it can be done using a different selection of document types. *Variation 2*: If data aggregated from related fields is required, it can cater for this by weighting according to the number of fields assigned to the journal.
Issues noted in the literature	*Flexibility*: It is based on flexible and variable citation and publication windows. For example JFIS for the time period 1993-1997 is based on publications published in a certain journal during the time 1993-1997, and on the citations gathered by these publications during the same time-period. Consequently, the citation window is variable, and varies from 5 years for publications published in 1993 to 1 year for publications published in 1997 (ibid.: 251-252). *Inappropriate use:* The indicator should only be used as a reference value, and never as a substitute for actual citation counts.
Data requirements	ISI data on journals (Journal Citation Reports) is sufficient if 5 year windows or publication year to date calculations are used. If other citation windows (e.g. two or three years) are to be used, then access to the full database is required

4.2.2 Control Variables

Indicator 11. Share of self-citations	
Description	This indicator calculates the proportion of total citations that a publication receives from other publications sharing a common author. It is used to determine whether the self-citation rate of the evaluated unit is unusually high or low.
Application	CWTS
Calculation	For a given publication set: %Pnc = total number of self-citations * 100 / total number of citations
Data requirements	Requires access to the full ISI database with details of individual references, rather than just citation counts.

4.2.3 Performance indicators

The following indicators are variants of citation counts that include or exclude self-citations.

Indicator 12. Number of Citations	
Description	The total number of citations received by a set of publications (e.g. Moed *et al.* 1995: 398) over a specified time period.
Application	CWTS (C)
Calculation	C = ∑ number of citations to each publication in years $y_{1\ldots x}$
Issues noted in the literature	*Control variable*: Because self-citations are not excluded, the share of self-citations must be calculated as a control variable.
Data requirements	ISI databases with yearly counts of citations to each publication.

Indicator 13. Number of external citations	
Description	The total number of citations received by a set of publications, excluding author self-citations (i.e. a citation received from a publication of which at least one author is also an author of the cited paper).
Application	CWTS (Cx)
Calculation	Cx = (∑ number of citations to a publication) – (number of self-citations to the publication)
Issues noted in the literature	*Cost:* More costly than ignoring self-citations (Katz and Hicks 1997: 542). *Variation*: Excluding all 'in-house' citations, not just those where at least one author of the linked publication is common. The indicator tallies the number of citations received by a set of publications, but excludes citations from the same institution (e.g. from a neighbouring department). This reflects the impact of a research work outside its own institution (Moed *et al.* 1985c: 136). However, it does not exclude the impact a publication has in the research laboratories of international co-authors.
Data requirements	Full ISI database, including bibliographic details of all citing publications.

Indicator 14. Citations per publication	
Description	This indicator calculates the average citation rate of a publication. The aggregate citation tally is normalised by the number of publications being analysed. This indicator normalises for the size of unit. CWTS usually uses the variant which excludes self-citations.
Application	REPP, CWTS (Cpp)
Calculation	For a given publication set: Cpp = Total number of citations to all publications / number of publications or with self-citations excluded: Cpp_{ex}= (Total number of citations to all publications – total number of self-citations) / number of publications
Issues noted in the literature	*Advantages:* Organisational units of different size can be compared, assuming that larger units produce more publications and hence get more citations in aggregate. *Number threshold:* Because the distribution of citations is highly skewed, this indicator should not be used on very small numbers of publications (Moed *et al.* 1995: 411). *Reference values:* World-wide citation rates are increasing, so when using this indicator in a time series analysis, it must be normalised to other reference values, such as field or journal averages (Van Raan 2000: 317– see below)
Data requirements	For Cpp, a version of the ISI database that contains yearly citation counts to each publication. For Cpp_{ex}, full ISI database with complete reference information for every citation.

Indicator 15. Uncitedness	
Description	Percentage of papers in a given publication set that have not been cited.
Application	CWTS (%Pnc)
Calculation	For a given publication set: %Pnc = number of uncited publications / total number of publications * 100
Issues noted in the literature	*Alternative use*: Other bibliometricians used it as auxiliary measure since its changes over time might heavily influence the normalised citation indicators (Glänzel et al 2003: 200).
Data requirements	A version of the ISI database that contains yearly citation counts to each publication.

Indicator 16. Distribution of publications over field-normalised impact classes [22]	
Description	Due to the skewness of citation distribution, the distribution of publications over impact classes should be analysed to see if the citation per publication rate (particularly a high one) is due to the presence of many well-cited items, or rests on a small number of outstanding contributions.
Application	CWTS (new indicator)
Calculation	Five impact classes are constructed, using citations and field reference values: C = number of citations received by a publication FCS = field-normalised impact factor, i.e. the average number of citations in a certain field (see Indicator 9 for a detailed description) C/ FCS = relative impact of a publication compared to the field average Class I: C/ FCS = 0 ; i.e. publications with no impact at all, Class II: 0.8 > C/ FCS > 0; i.e. publications with an impact up to 80% of the field average, Class III: 1.2 > C/ FCS > 0.8, Class IV: 2.0 > C/ FCS > 1.2 Class V: C/ FCS > 2.0, i.e. at least double the field average (Van Leeuwen *et al.* 2003: 262)
Data requirements	ISI data, including counts of citations to each publication.

[22] A similar approach was chosen by Crouch *et al*. (1986: 247)

Indicator 17. Top 5% most frequently cited publications	
Description	Used to discover if there are outstanding single contributions among a publication set. Because of the high skewness of citation distributions, even on aggregated levels, these contributions 'vanish' if average citation counts are used (Tijssen *et al.* 2002: 386).
Application	REPP, CWTS (new indicator)
Calculation	Number of publications that belong to the 5% most frequently cited papers in a given year in a given field for a given citation window. C_{top5} = citation threshold for a paper to be in the top 5% most highly cited in the field P = total number of publications in the field P_{top5} = number of publications in the field with a citation count ≥ C_{top5} $E(P_{top}) = 0.05 * P_{to\text{-}t1}$ = expected number of publications among the top 5% given the size of the unit (i.e. 5% of the total) $P_{top5}/E(P_{top5})$ = Ratio of the actual and expected number of papers among the 5%
Varaitions	*Variant 1*: The calculation can also be used at different levels – top 1%, top 10%, etc, e.g. Bourke *et al.* 1999: 17). *Variant 2*: REPP calculates the number of papers among the top 1% of Australian papers in a given field (e.g. Butler 2001: 151), rather than among the world's most highly cited. *Variant 3*: A former variant of this indicator was the 'star effect' which estimated what would happen if an important scientist leaves the group by removing this person's publications and citations from the calculation (Nederhof and Van Raan 1993: 356).
Issues noted in the literature	*Interpretation*: $P_{top5}/E(P_{top5}) > 1$ means, the unit had more highly cited papers than expected $P_{top5}/E(P_{top5}) < 1$ means, the unit had less highly cited papers than expected (Van Leeuwen et. al. 2003: 261-266) *Validity*: the assumption is that a unit with outstanding (highly cited) publications performs better than a unit which produces a steady stream of publications that are medium cited (Van Leeuwen *et al.* 2003: 261). *Publication type*: CWTS excludes publications classified as 'letters', which they include in other indicators, because it has a deviant citation pattern compared to articles and reviews.
Data requirements	The standard CWTS indicator requires access to the full ISI database with yearly citation counts to each publication. The REPP variant only requires access to the subset of Australian publications from the ISI database with yearly citation counts to each publication.

The next three indicators (18 to 20) apply the three impact scores that were introduced as auxiliary indicators. They serve as reference values in citation measures.

Indicator 18. Comparison of actual citation rates to journal averages	
Description	It compares the average number of citations to a research unit's oeuvre compared to the mean impact of all publications in the journals in which those publications appear.
Application	REPP, CWTS(Cpp/JCSm)
Calculation	Cpp = actual citation per publication rate (see indicator no. 14 for calculation) JCSm = expected citation impact for the journal set (see indicator no. 8 for calculation) A/E_j = Cpp/ JCSm
Variations	The indicator is similar to the Relative Citation Rate, developed by Schubert *et al.* (1989).
Issues noted in the literature	*Interpretation*: If the ratio is above 1.0, the mean impact of a group's papers exceeds the mean impact of all articles published in the journals in which the particular group published (Moed *et al.* 1995: 399-400). *Validity*: Low impact publications published in low-impact journals might get a similar score as high impact publications published in high-impact journals. Therefore, the field-normalised citation rate (see below) is a better reference value (Van Leeuwen *et al.* 2003: 260).
Data requirements	A version of the ISI database that contains yearly citation counts to each publication

Indicator 19. Comparison of actual citation rates to the field average	
Description	It compares the average number of citations to a research unit's output in a particular field to the citation average for the field as a whole.
Application	REPP, CWTS (Cpp/FCSm)
Calculation	Cpp = actual citation per publication rate for publications in field F (see indicator no. 14 for calculation) FCSm = world citation rate for field F (see indicator no. 8 for calculation) A/E_F = Cpp/ JCSm
Variations	A similar indicator was developed earlier by Vinkler (1986), called the Relative Subfield Citedness
Issues noted in the literature	*Interpretation*: If the ratio is above 1.0, the research unit's oeuvre is cited more frequently than an 'average' publication in the subfield(s) in which the unit is active (Moed *et al.* 1995: 399). *Validity 1*: It is considered to be the most adequate measure for research performance (Van Raan 2000: 317). This indicator should preferably be used, if the definition of the field is acceptable. Otherwise, the journal citation average should be used as reference value (Moed *et al.* 1995: 410). *Validity 2:* The indicator is problematic for interdisciplinary research groups that may be highly cited in their main field, but cited low in other fields (ibid.: 404, Rinia *et al.* 2001: 360).
Data requirements	A version of the ISI database that contains yearly citation counts to each publication.

Indicator 20. Weighted impact	
Description	This indicator weights each citation received by a publication according to the field normalised citation impact of the journal which carries the citing paper.
Application	CWTS (new indicator)
Calculation	The journal impact JFIS is calculated (see above) Weighted impact = ∑(citations * JFIS) / total number of citations
Issues noted in the literature	*Interpretation*: Scientists think, it makes a difference if one's work is cited in well-known high impact journals. (Van Raan and Van Leeuwen 2002: 619; Van Leeuwen *et al.* 2003: 263-264). For each ISI publication the citing publications are identified. *Limitations*: Some document types (e.g. meeting abstracts) and citations to non-ISI journals cannot be included because no JFIS value can be calculated due to small numbers. (ibid.: 273) A second problem is that some citations stem from Non-ISI journals, for which no journal impact can be calculated. *Fairness*: Younger scientists are disadvantaged since it is more difficult for them to get in the top journals (ibid.).
Data requirements	ISI database with complete reference information for every citation.

4.3 Structural indicators characterising the research

The following indicators are not direct performance measures, but give the evaluators additional information about the characteristics of the research undertaken, such as publication strategies, the place of a research unit in its scientific communities, etc. Trend analyses that show if these characteristics change over time are especially interesting. They require peers who are able to interpret this background information.

Indicator 21. Ratio journal impact to subfield impact	
Description	For a given field, the indicator compares the average impact of the ISI journals in which a unit publishes to the average impact of all journals in that field.
Application	REPP, CWTS (JCSm/FCSm)
Calculation	JCSm / FCSm (for calculation of JCSm and FCSm – see indicators 8 and 9)
Issues noted in the literature	*Interpretation*: If the value is above 1.0, then the research unit publishes in journals with above-average impact. (Van Raan and Van Leeuwen 2002: 617; Van Leeuwen *et al.* 2003: 263) The authors do not clearly describe the purpose of this indicator. But it seems to characterise the publication strategy of the research unit, i.e. whether it publishes in leading journals or not. *Variations*: A similar indicator was developed by Schubert and Glänzel (1983) as well as Vinkler (1988).
Data requirements	ISI data on journals (Journal Citation Reports) is sufficient if 5 year windows or publication year to date calculations are used. If other citation windows (e.g. two or three years) are to be used, then access to the full database is required.

Indicator 22. Position in the journal spectrum	
Description	This gives more qualitative information about the journals in which a unit publishes or about the journals from which it is cited.
Application	CWTS
Calculation	Publishing journal list: list of the journal titles in which a research unit published and the number of publications appearing in each journal. Citing journal list: list of the journal titles from which a unit is cited and the number of citations that come from each journal.
Issues raised in the literature	*Interpretation*: Publishing and citing journal lists may be compared: is a group cited by the same journals in which it publishes? (Moed *et al.* 1995: 400). To position citing publications requires direct access to a full citation database.
Data requirements	Requires the full ISI database with complete reference information for every citation.

Indicator 23. Field distribution of publications	
Description	Since a research unit often does research in several research fields, it is informative to know, what their research profile is — what fields are the major focus of their research?
Application	REPP, CWTS
Calculation	All ISI publications are disaggregated into subfields, based on the ISI classification of journals into categories. For each ISI subfield the number of publications and the impact (field-normalised citation rate Cpp/FCS) are calculated (Moed *et al.* 1995: 401-403).
Issues raised in the literature	*Interpretation*: This is useful for the evaluation of interdisciplinary research groups and can be seen as indicator for the extent of interdisciplinarity (Rinia *et al.* 2001, Van Raan and Van Leeuwen 2002: 618). It is also useful in comparative evaluations for determining whether the units being compared have similar research profiles.
Data requirements	The ISI website lists journals by subject category, and this can be used for allocating journals to the desired field classification scheme.

Indicator 24. Field distribution of citations	
Description	This is a field-specific break-down of all publications citing the work of the research unit. It is useful for the evaluation of interdisciplinary research groups. The citing fields can also be compared to the fields in which a unit publishes (Van Raan and Van Leeuwen 2002: 619-621).
Application	CWTS
Calculation	Citations are assigned to ISI subfields. Field-normalised citation rates (Cpp/FCSm) are calculated for each field.
Issues raised in the literature	*Validity:* Interviewed scientists assessed this indicator as very useful, especially if changes over time are analysed (ibid.: 630).
Data requirements	Requires direct access to a full citation database.

Indicator 25. Number of citers	
Description	The indicator measures the dissemination of a scientific work, by calculating the number of different authors who cite it (Dieks and Chang 1976: 262).
Calculation	A_c = total number of authors citing the publication.
Issues noted in the literature	*Interpretation*: The idea is that publications are worth more if they are cited by different research groups rather than repeatedly by the same group. *Variations*: the calculation of international citers or the number of different journals in which the citations occur.
Data requirements	It requires access to the full ISI database with complete reference information for every citation.

Indicator 26. Level of collaboration	
Description	This indicator demonstrates the links a unit has with internal and external groups.
Application	REPP, CWTS
Calculation	Each publication can be classified to one of five levels, using the authors and addresses listed on it: Single author: i.e. no collaboration; Group: involves more than one author from the same research group; Institutional: involves authors from more than one unit at the same institution; National: contain addresses from more than one institution in the same country; International: contain addresses from authors outside the country. For these types of collaboration, publication and citation rates are calculated.
Variations	The differentiation of collaboration levels depends on the amount of 'cleaning' done of the raw address data. The CWTS uses groups as the reference point, and does not distinguish between institutional and national collaboration.
Issues raised in the literature	*Deployment*: It shows how frequently a research group has collaborated with other groups, and how the impact of papers resulting from collaborations compared to non-collaborative papers (Moed *et al.* 1995: 401-403).
Data requirements	It requires access to a database, such as ISI, listing all addresses.

Indicator 27. Country analysis of collaborators and citers	
Description	Country-specific collaboration profiles are constructed, showing which countries play a dominant role.
Application	REPP, CWTS
Calculation	The publications of a research unit are arranged according to the countries they contain in their address information. Field-normalised citation rates (Cpp/FCSm) are calculated.
Variations	While most analyses chose countries as the level of aggregation, it would also be possible to chose a lower level of aggregation and analyse group, institutional, sectoral or geographical patterns of collaboration.
Issues raised in the literature	*Rationale:* From the citations, the profile of citing countries is derived. By comparing different countries, similarities and differences can be found. (Van Raan and Van Leeuwen 2002: 624-625) For each country, citation rates are calculated in order to analyse the impact of a country.
Data requirements	Requires a citation database with the address information.

Indicator 28. Level of Research	
Description	It describes the basic or applied character of the publications. SCI journals are assigned to four categories. This classification system is used to characterise single articles, based on the research orientation of the journals on the basic-applied dimension.
Application	REPP
Calculation	The four levels are: Level 1 – Applied Technology; Level 2 – Engineering-Technological Science; Level 3 – Applied Research; Level 4 – Basic Scientific Research
Issues raised in the literature	*Rational:* For research evaluations, it can provide useful information because if a research unit has a strong applied character the unit is less often cited (Butler 2001: 30). *Validity:* The basic-applied character of a publication is only an approximation. Firstly, because journals are not always easily assignable to these categories, and the single articles might differ again from the journal assignment. Secondly, there is a growing number of research processes which have both basic and applied elements. A problem with the classification system is that it differentiates between different types of applied research (level 1 to 3) but not between different types of basic research (only level 4).
Data requirements	Requires a frequently updated classification system, as provided by Computer Horizons Inc. (CHI).

Indicator 29. Activity Index	
Description	This indicator measures the research profile of the unit being analysed – it identifies the fields or disciplines in which the unit is more or less active than the norm.
Calculation	AI = Activity Index S_u = the share of the given field in the publications of unit S_w = the share of the given field of total world (or national, etc.) publications $AI = S_u / S_w$
Variations	Relative specialisation index (RSI) = (AI – 1)/(AI + 1) The RSI reflects the internal balance among fields within the unit of analysis, as positive RSI values must always be balanced by negative ones (Glanzel 2000: 126).
Issues raised in the literature	The calculation of the benchmark against which activity is judged will depend on the level of aggregation – if the analysis focuses on a nation's output, the appropriate norm would be the distribution of publications across fields for the whole world. If the focus is institutional, the more relevant benchmark may be the distribution of publications across fields for the country, or for a particular sector within that country. The indicator is easily re-defined to take the level of aggregation into account (Glanzel 2004: 49).
Data requirements	This indicator relies on access to comprehensive databases such as ISI to be able to calculate benchmark data.

4.4 Non-bibliometric indicators

Several other quantitative performance measures which are not based on publications have been suggested. The most common ones are described in the next section. These are presented in the same format as publication-based indicators (see Section 4 for an explanation of the format) with the exception of the "Application" details. For non-bibliometric indicators, this section notes any routine use of the indicator in the evaluation of research performance.

4.4.1 Reference variables

The idea of relating output measures to input measures is borrowed from economics. While not a performance measure in their own right, input measures can be used to normalise the different sizes of research units being evaluated. The number of researchers is the most commonly used input indicator. It is often applied to calculate the 'number of publications per researcher' or the 'number of citations per researcher' (e.g. Nederhof and Van Raan 1993).[23]

Indicator 30. Number of researchers	
Description	The number of researchers that were active in the time period under evaluation.
Calculation	The number of researchers is counted (regardless of the time they devote to research) or the number of full time equivalent (FTE) researchers, taking into account part-time or fractional appointments.
Variations	Some applications limit the counts to 'tenured staff (e.g. Aksnes 2003: 35, Schloegl *et al.* 2003).
Issues noted in the literature	The calculation does not take into account the actual time available for research. Also, special roles need to be considered, for example, scientists fulfilling service functions who therefore do not publish to the same extent as other researchers (De Bruin *et al.* 1993: 37-39; Gläser *et al.* 2004). Another problem is, how to adjust for movement of scientists between institutions or units, and how to treat guest scientists (Johnes 1988: 62). Guest scientists should be excluded, if they only came into a research group in order to learn new methods, a quite common behaviour in the sciences (Moed *et al.* 1998: 234, 252).
Data requirements	Access to statistics on research personnel.

[23] Unfortunately, there is a tendency to incorrectly interpret the outcome of these indicators as research efficiency.

Indicator 31. Research time	
Description	This auxiliary indicator is used to weight staff numbers according to the proportion of their time spent on research.
Calculation	T_r = time spent on research / total work time * 100
Issues noted in the literature	The time researchers can devote to research varies according to type of appointment (i.e. whether in a full-time research position or a teaching and research position) and to the other tasks required of their position (Nederhof *et al.* 1993: 159; Casey 1997: Report 3, Chap.4). There is little consensus on the relative weighting that should apply to the different categories. A basic problem is whether to base the calculation on total hours worked or contractual working time (ibid.). In OECD statistics it is assumed that researchers dedicate 40% of their time to research (Moed *et al.* 1998: 234). The study conducted for the Dearing review of higher education in the UK found the proportion varied between 10% and 50% depending on age, level and type of institution (Casey 1997: Report 3, Chap.4).
Data requirements	Access to statistics on research personnel.

Other auxiliary indicators were suggested, for example the number of support staff, which Martin and Irvine argue that they have an influence on the research activity (Martin and Irvine 1983: 78). '*Research expenditure*' is another input indicator, for example measured by an institute's annual budget (Lee 2003) or the proportion of a nation's GDP. However, expenditure has a strong relationship to research capabilities (productivity and competitiveness) rather than actual performance (BIE 1996: Vol 2, 27).

4.4.2 Performance indicators

Indicator 32. External funding	
Description	The amount of research funding attracted from agencies external to the institution.
Application	This indicator is often used in research evaluations, e.g. in the British Research Assessment Exercise or as a component of Australia's funding formula for universities.
Calculation	$ received in a specified period In some instances, the source of funds that can be included may be limited e.g. only nationally competitive grants money can be included.
Issues noted in the literature	It is often suggested by scientists who were asked which indicators are appropriate to measure research performance in their field (e.g. Tognolini *et al.* 1994). The appeal of using external funding as an indicator lies in the fact that the award of a grant is based on peer review (Gillett 1991: 254) and that it measures the current performance rather than past performance (Hornbostel 2001). However, objections have been made against its use on the grounds that grant applications measure only research potential rather than actual achievement (Gillett 1991: 254). Hornbostel claims that it is only appropriate if: - external fund acquisition is common for the field; - grant proposals are reviewed by qualified peers in a competitive system; - success rates are not too low; - there is a mix of different funding sources available; and - adequate infrastructure support is provided (Hornbostel 2001). It is also often claimed that research income is an input rather than an output indicator and that this leads to double counting of research grants and resultant publications (Phillimore 1989: 263; Gillett 1991; Johnes and Johnes 1995: 305). It is necessary to take into account the varying resource requirements of different disciplines (Phillimore 1989: 263; Hornbostel 2001: 525). Although the quality of a proposal and the reputation of a researcher are important prerequisites for a successful grant application, success can depend on several other factors outside the control of by scientists, e.g. the research field a scientist is working in. It seems especially problematic to use it in a comparative manner or to aggregate it across different research fields (Laudel 2004).
Data requirements	Data from funding bodies disaggregated to the level of evaluation.

Indicator 33. Research students data	
Description	Data collected on the research students of a unit.
Calculation	Research student data can be collected in a number of formats: actual numbers of students; student load; completions; throughput rates.
Issues noted in the literature	This indicator is often used. For example, it is part of the Australian formula for the distribution of research funds to universities, it is used in the British RAE and in Dutch evaluations of university research (Geuna and Martin 2001). Discussions about the validity of this indicator were rarely found in the reviewed literature. Concern was raised that this indicator might reflect teaching quality rather than research, particularly if degrees lower than PhDs are included (Phillimore 1989: 263). Scientists from different natural science fields regarded the number of dissertations as a poor measure, though questionnaire design did not allow for further explanation (Jauch and Glueck 1975: 70).
Data requirements	Standardised data from all institutions.

Indicator 34. Keynote addresses	
Description	Invitations to deliver keynote addresses or present refereed papers at major national or international conferences.
Calculation	Count of the number of presentations in a given time period.
Issues noted in the literature	An invitation to present a paper at an international conference implies that a scientist is held in high regard by her peers in the international scientific community (King 1987: 271-272). However, surveyed scientists claimed that there is a bias because of personal networks (Wood 1989: 240).
Data requirements	

Indicator 35. International visits	
Description	Visits to overseas research institutions, or visits by foreign researchers.
Calculation	Count of the number of visits in a given time period.
Issues noted in the literature	The idea is to analyse the number of overseas visits, length of stay and standing of their home research institute, per analysed research unit. Similarly, visits abroad might provide useful data (King 1987: 272). The number of guest researchers was used in evaluations, based on peer review (e.g. Moed *et al.* 1985b: 155).
Data requirements	

Indicator 36. Honours and awards	
Description	International and major national honours and awards awarded by peers in the field.
Calculation	
Issues noted in the literature	Prizes vary in several ways: what is awarded depends on the donor; the criteria are often vague; non-scientific criteria are sometimes included; and the period for which an honour is awarded ranges from a single achievement to the lifework etc. (Hornbostel 1997: 208-211). Honours and professional status indicators are usually used by the scientists themselves in their CV's to demonstrate their reputation. In research evaluations they serve to inform peer review (e.g. Fries 2003). The problem is that these indicators usually require background knowledge about the prestige of prizes, journals, conferences, funding agencies or further specifications (e.g. about the frequency of referee activities) that go beyond a simple counting. Honours and awards can be ranked by expert committee on a five point scale (Luwel *et al.* 1999: 82, 177-8).
Data requirements	

Indicator 37. Election to learned societies	
Description	An elected fellow, or equivalent, of a national or international learned academy or other learned society.
Calculation	
Issues noted in the literature	Election to learned societies, where societies ranked by expert committee on a five point scale (Luwel *et al.* 1999: 84, 179).
Data requirements	

Indicator 38. Editorial board membersip	
Description	An editor or member of an editorial board/committee of a scholarly refereed serial.
Calculation	
Issues noted in the literature	Refereeing activities of journal articles is not seen as a discriminating indicator as it is seen as a standard activity of academic research (Wood 1989: 240). The value of editorial board memberships depend on the status of the journal (Wood 1989). King claims that other factors besides scientific merit play a role. Furthermore, insufficient data limit the applicability of these indicators (King 1987: 272). Includes editor in chief and membership of editorial board.
Data requirements	Data gathered can cover title of periodical or serial work, first year and last year of doing; and can be divided into journals with a referee system and without a referee system (Luwel *et al.* 1999: 32).

Indicator 39. Membership of review committees	
Description	Appointed on the basis of field expertise to be a member or chair of an academic review committee or a scientific advisory board.
Calculation	
Issues noted in the literature	
Data requirements	

Indicator 40. Membership of government bodies	
Description	Invitations to serve on major national and international bodies, such as the OECD, WHO, UNESCO, ARC, NHMRC, etc.
Calculation	
Issues noted in the literature	
Data requirements	

4.4.3 Non-bibliometric indicators for the social sciences and humanities

Due to concerns about the coverage of indexed databases (see section 3.4 above) non-bibliometric indicators have particular significance for the social sciences and humanities, and may be taken as a more appropriate way to measure research performance. It has been noted that Luwel *et al.* (1999) have provided a rare and comprehensive investigative study in this area. Their methodology-based study is confined to four major Flemish universities, and to the case studies of law (for social sciences) and linguistics (for humanities). Questionnaires provide the primary data source for scholarly output by academics, and there is no use of the SSCI or A&HCI. Respondents in the studied institutions and from an international sample were asked to provide data on scholarly committee memberships, scholarly prizes and awards, and on external funding and collaborations (1999: 12). Luwel *et al.* demonstrate that "bibliometrics is much more than conducting citation analyses based on the ISI indexes, as citation data do not play any role in this study" (1999: 13). The law part of this study is replicated in Moed *et al.* (2002).

Potential non-bibliometric indicators to add for the social sciences and humanities include:

- Formal research cooperation outside of one's own university leading to shared project applications/co-publications (Luwel *et al.* 1999: 33).
- 'Research activity not leading to publications' including specialist lectures at conferences and workshops, organising conferences/workshops, editing special issues of journals, activities in learned societies/institutions, providing academic advice (e.g. assessing research applications, manuscript refereeing, membership of expert groups for pilot projects in the area of research evaluation, supervision of PhD thesis, membership of a PhD jury, contributions to dissemination/popularisation of research in the media (Luwel *et al.* 1999: 29).

- Participation in university management and administration such as membership of academic council or faculty council, being a research co-ordinator (Luwel *et al.* 1999: 28).
- Rendering scholarly services to the community, for example, policy preparation research, membership of government advisory bodies, delivering lectures or information for secondary school pupils (Luwel *et al.* 1999: 29).
- Advisory function and membership of an academic committee on the basis of expertise, i.e. 'membership of a policy advisory commission or council in the context of academic societal service' (Luwel *et al.* 1999: 32).
- Membership of assessment committee outside one's own university to assess research proposals or scholars' academic records (Luwel *et al.* 1999: 32).
- Membership of a jury for a research reward (Luwel *et al.* 1999: 32).
- Administrative functions within learned societies (Luwel *et al.* 1999: 32).
- Station or channel of media broadcast (Luwel *et al.* 1999: 30).
- Honorary degrees (Van Raan 1998: 5).

In follow-up surveys, respondents suggested other markers of esteem (Luwel *et al.* 1999: 89, 187) including:

- invited lectures especially at overseas and prestigious universities;
- invitations to contribute to publications;
- invitations to present lectures for professional organisations and firms.

4.4.4 Non-bibliometric indicators for the creative arts

Non-bibliometric indicators for the creative arts have generally been overlooked by the mainstream bibliometrics literature, yet due to the importance of non-written work this is an area ripe for novel investigation. Hattie recounts that in 1987 the Research Committee at the University of Western Australia decided to allocate its research budget for three years according to measures of research productivity, and it included the novel non-bibliometric indicator of "creative work which has a demonstrable research component" (1990, 250).

Section 3.4 above discussed Strand's definition of publication in the creative arts extending far beyond the written word. He notes that publication in the creative arts includes:

- *public performance* for dancers, actors, choreographers, musicians, playwrights;
- *building or manufacture* for architects and designers;
- *written works* for academics, researchers and creative writers;
- *exhibitions* for visual artists, craftspeople; and
- *computer software* for designers, musicians, visual artists and so on (Strand 1998: 55).

5 Concluding remarks

We could extract from the literature a fairly comprehensive list of quantitative indicators, but we could not achieve our aim of discovering the efficacy of their use in the context of research evaluation. Only a limited discussion, focussed on a restricted number of measures, could be found on issues concerning the validity of using quantitative measures to assess research performance, and the methodological and technical concerns in their application. However it was apparent from the literature was that the seriousness of concerns on these issues increases as the level of aggregation decreases and is most acute at the level of individual scientist or publication.

Most discussion in the literature focuses on bibliometric indicators and a wide range of measures have been identified. Publication indicators that merely count the number of publications have been rejected as valid indicators of research performance. Attempts have been made to overcome their limitations by restricting counts to special types of publications (such as the number of review articles or the number of publications in top journals), or by using journal impact factors to weight the raw counts.

Citation indicators are more accepted as performance measures, though it is acknowledged that they address only one aspect of research quality – impact. Simple citation counts are criticised for their failure to take into account field-specific characteristics. More sophisticated measures normalise citation counts by reference variables such as the journal or field averages. Some citation indicators focus on the external influence a work has by excluding citations from the cited research department or organisation. Unfortunately, a number of the more sophisticated and complex indicators, which can overcome many of the concerns of those being assessed, are based on methods which presuppose access to a full citation database which is available to very few practitioners.

Newer developments in indicators by CWTS go in two directions: the detection of ‘scientific excellence’ and the development of indicators which provide additional structural information. The focus on ‘scientific excellence’ is an attempt to take account of the skewed distribution of citations by concentrating on that portion of the publication output that drives the more commonly applied statistics that are based on means.

Bibliometric indicators cannot be used like cooking recipes. Many methodological problems have not yet been satisfactorily resolved, though some partial solutions are offered by the bibliometric community. Considerable expertise is required to conduct bibliometric studies and interpret the results, such as: what share of publications is a sufficient coverage to allow us to use ISI databases? What follows if the self-citation rate is unusually high? etc. This knowledge, accumulated by units such as CWTS or REPP who regularly do evaluations, remains to a great extent implicit. Any attempt to use bibliometric indicators for research assessment should be undertaken with caution and preferably by experienced bibliometricians or by tapping into their know-how.

Structural indicators do not directly measure research performance but give additional information about the unit being assessed. Such information includes: the fields in which a research unit publishes; the fields in which it is cited; characteristics of the collaborators (their institutional affiliation etc.); the basic or applied character of the unit’s research; and the cognitive structure of its research field. This information, combined with publication and citation indicators or as additional information in peer review processes, can be very useful.

Ranking tables where the research units are presented according to their performance, measured by bibliometric or other quantitative indicators, are very popular. The advantage of these tables is that they are easily understandable for science politicians and the public. But the danger is that they imply they are the final result rather than the result of a quantitative analysis of research performance which needs to be interpreted by peers of the analysed research field(s).

While many examples exist of the application of bibliometric measures in the research assessment process, few go beyond a description of the methodology and results to detail how they influenced this process. Typically, we get a list of indicators, the names of the evaluated research units and the bibliometric data, either in tabular or graphical format. The next steps – how these indicators are used by evaluators and an assessment of their value in the process – rarely feature in the analysis. The exceptions to this are some of the studies carried out by CWTS.

Non-bibliometric indicators are largely neglected in the literature, despite some of them being frequently used in evaluations (e.g. 'external funding', number of PhD students) or regularly listed as possible indicators by researchers (e.g. honours and awards).

We have seen that while many issues must be resolved in using bibliometric techniques to evaluate research performance in the experimental sciences, there is a belief that sophisticated techniques may be able to overcome many of the problems. This is not the case for the social sciences and humanities where indexed databases do not cover the typical published output of many fields. The problem of fields that are not well covered by ISI has recently received more attention with the fast growing demand for quantitative indicators. In this respect, exploring more novel indicators of research performance (both bibliometric and non-bibliometric) in the social sciences and humanities may prove to be a fruitful future area of investigation.

References

Aksnes, D. W. 2003. A macro study of self-citation. *Scientometrics* 56: 235-246.

Aksnes, D. W. and R. E. Taxt. 2004. Peer review and bibliometric indicators: a comparative study at a Norwegian university. *Research Evaluation* 13: 33-41.

Albert, M. B., D. Avery, F. Narin and P. McAllister. 1991. Direct validation of citation counts as indicators of industrially important patents. *Research Policy* 20: 251-259.

Amsterdamska, O. and L. Leydesdorff. 1989. Citations: indications of significance? *Scientometrics* 15: 449-471.

Anderson, R. C., F. Narin and M. McAllister. 1978. Publication ratings versus peer ratings of universities. *Journal of the American Society for Information Science* 3: 91-103.

Australian Academy of Science. 1996. *The Impact of Australian Science*. Discussion paper, The Australian Academy of Science, Canberra.

Baird, L. M. and C. Oppenheim. 1994. Do citations matter? *Journal of Information Science* 20: 2-15.

Björneborn, L. 2004. Small-world link structures across an academic web space - a library and information science approach. PhD Thesis. Copenhagen: Royal School of Library and Information Science.

Bourke, P. 1985. *Quality Measures in Universities*. Commonwealth Tertiary Education Commission, Canberra.

Bourke, P. and L. Butler. 1993. *A Crisis for Australian Science?* PIP Monograph Series No. 1, Research School of Social Sciences, ANU. .

Bourke, P. and L. Butler. 1995. The use of bibliometric data in evaluating a research university: issues and measures. Koenig, M. E. D. and A. Bookstein (eds), *Proceedings of the Fifth International Conference of the ISSI*, River Forest, Illinois.

Bourke, P. and L. Butler. 1996. *The Origins, Structure and Uses of the Australian National University Bibliometric Database*. Paper presented at the Citation Database: Its Construction, Management and Applications, Taipei, Taiwan.

Bourke, P. and L. Butler. 1998. Institutions and the map of science: matching university departments and fields of research. *Research Policy* 26, 711-718.

Bourke, P., L. Butler and B. Biglia. 1996. *Monitoring Research in the Periphery: Australia and the ISI Indices*. REPP Monograph Series No. 3, Research School of Social Sciences, ANU.

Bourke, P., L. Butler, and B. Biglia. 1999. *A Bibliometric Analysis of Biological Sciences Research in Australia*. DETYA No. 6307HERC99A. Commonwealth of Australia: Department of Education, Training and Youth Affairs, Higher Education Division.

Broadus, R. N. 1971. The literature of the social sciences: a survey of citation studies. *International Social Science Journal* 23(2): 236-243.

Bud, R. 1985. The case of the disappearing caveat: a critique of Irvine and Martin's methodology. *Social Studies of Science* 15: 548-553.

Bureau of Industry Economics (BIE). 1996. *Science System: International Benchmarking.* Report 1996/2, AGPS, Canberra.

Burnhill, P. and M. E. Tubby-Hille. 1994. On measuring the relation between social science research activity and research publication. *Research Evaluation* 4(3): 130-152.

Butler, L. 1988. Personal communication of unpublished data in D. Hicks, 1999. The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. *Scientometrics* 44(2): 193-215.

Butler, L. 2001. *Monitoring Australia's Scientific Research: Partial indicators of Australia's research performance*. Canberra: Australian Academy of Science.

Butler, L. 2003. Explaining Australia's increased share of ISI publications—the effects of a funding formula based on publication counts. *Research Policy* 32: 143-155.

Butler, L. and M. Visser. 2005. Extending citation analysis to non-source items. *Scientometric,* submitted.

Casey, B. 1997. Academic staff in higher education: their experiences and expectations. R. Dearing, *The National Committee of Enquiry into Higher Education*, Report 3, url: http://www.leeds.ac.uk/educol/ncihe/

Cave, M., M. Kogan, S. Hanney and D. Burningham. 1988. *Performance Indicators for the Economic and Social Research Council*: *Report on Preliminary Project prepared for UK Economic and Social Research Council*.

CHASS (Council for the Humanities, Arts and Social Sciences), forthcoming 2005. *Measures of Quality and Impact of Publicly Funded Humanities, Arts and Social Sciences Research*.

Cole, J. R. and S. Cole. 1973. *Social Stratification in Science*. Chicago: The University of Chicago Press.

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., J. R. Cole and L. Dietrich. 1978. Measuring the cognitive state of scientific disciplines. Y. Elkana, J. Lederberg, R. K. Merton, A. Thackray and H. Zuckerman (eds), *Toward a Metric of Science: The Advent of Science Indicators*. New York: John Wiley & Sons, Inc, 209-251.

Collins, H. M. 1985. The possibilities of science policy. *Social Studies of Science* 15: 554-558.

Cozzens, S. E. 1987. Expert review in evaluating programs. *Science and Public Policy* 14: 71-81.

Cronin, B. and K. La Barre. 2004. Mickey Mouse and Milton: book publishing in the humanities. *Learned Publishing* 17(2):85-98.

Cronin, B., H. Snyder and H. Atkins. 1997. Comparative citation rankings of authors in monographic and journal literature: a study of sociology. *Journal of Documentation* 53(3): 263-273.

Cronin, B., G. McKenzie, L. Rubio, and S.Weaver-Wozniak. 1993. Accounting for influence - acknowledgments in contemporary sociology. *Journal of the American Society for Information Science* 44: 406-412.

Crouch, D., J. Irvine and B.R. Martin. 1986. Bibliometric analysis for science policy: an evaluation of the United Kingdom's research performance in ocean currents and protein crystallography. *Scientometrics* 9: 239-267.

De Bruin, R.E., A. Kint, M. Luwel and H. F. Moed 1993. A study of research evaluation and planning: the Univerisity of Ghent. *Research Evaluation* 3: 25-42.

Dieks, D. and H. Chang, 1976. Differences in impact of scientific publications: some indices derived from a citation analysis. *Social Studies of Science* 6: 247-267.

Dillon, J. T. 1981a. The emergence of the colon: an empirical correlate of scholarship. *American Psychologist* 36(8): 879-884.

Dillon, J. T., 1981b. Functions of the colon: an empirical test of scholarly character. *Educational Research Quarterly* 5: 71-75.

Edge, D. 1979. Quantitative measures of communication in science: a critical review. *History of Science* 17: 102-134.

Egghe, L. R. Rousseau and G. Van Hooydonk. 2000. Methods for accrediting publications to authors or countries: consequences for evaluation studies. *Journal of the American Society for Information Science* 51: 145-157.

Endersby, J. W. 1996. Collaborative research in the social sciences: multiple authorship and publication credit. *Social Science Quarterly* 77: 375-392.

Endler, N. S., Rushton, J. P. and H. L. III. Roediger. 1978. Productivity and scholarly impact (citations) of British, Canadian, and U.S. departments of psychology (1975). *American Psychologist* 33: 1064-1082.

Fries, M. 2003. Evaluation von Forschung und Lehre in Bayern. IHF-*Beiträge zur Hochschulforschung*: 6-31.

Garfield, E. 1965. Can citation indexing be automated? in Mary E. Stevens et al.(eds), *Statistical Association Methods for Mechanized Documentation*, Washington, DC: National Bureau of Standards.

Garfield, E. 1979a. Is citation analysis a legitimate evaluation tool? *Scientometrics* 1: 359-375.

Garfield, E. 1979b. Perspectives on citation analysis of scientists. E. Garfield (ed.), *Citation Indexing - Its Theory and Application in Science, Technology, and Humanities*. New York: John Wiley & Sons, Inc, 240-252.

Garfield, E. 1998. From Citation Indexes to Informetrics: Is the Tail Now Wagging the Dog? *Libri* 48: 67-80.

Geuna, A. and B. R. Martin. 2001. *University Research Evaluation and Funding: An International Comparison*. Brighton: Science Policy Research Unit.

Gilbert, G. N. and S. Woolgar. 1974. Essay review the quantitative study of science: an examination of the literature. *Science Studies* 4: 279-294.

Gillett, R. 1991. Pitfalls in assessing research performance by grant income. *Scientometrics* 22: 253-263.

Gillmor, C. S. 1985. Comments on the paper, “A re-evaluation of the contributions to radio-astronomy of the Nancay Observatory”. *4S Review* 3: 19-21.

Glänzel, W. and U. Schoepflin. 1999. A bibliometric study of reference literature in the sciences and social sciences. *Information Processing and Management* 35: 31-44.

Glänzel, W. 1996a. A bibliometric approach to social sciences. National research performance in 6 selected social science areas, 1990-1992. *Scientometrics* 35(3): 291-307.

Glänzel, W. 1996b. The need for standards in bibliometric research and technology. *Scientometrics* 35: 167-176.

Glanzel, W. 2000. Science in Scandinavia: A bibliometric approach. *Scientometrics* 48(2): 121-150.

Glänzel, W. 2004. *Bibliometrics as a Research Field: A Course on Theory and* Application of Bibliometric Indicators (http://www.norslis.net/2004/Bib_Module_KUL .pdf).

Glänzel, W. and H. F. Moed. 2002. Journal impact measures in bibliometric research. *Scientometrics* 53: 171-193.

Glänzel, W., R. Danell, and O. Persson. 2003. The decline of Swedish neuroscience: decomposing a bibliometric national science indicator. *Scientometrics* 57: 197-213.

Gläser, J. 2004a. Why are the most influential books in Australian sociology not necessarily the most highly cited ones? *Journal of Sociology* 40(3): 261-282.

Gläser, J. 2004b. Produzierende Gemeinschaften: Soziale Ordnung und kollektive Produktion (nicht nur) in scientific communities. Habilitationsschrift, eingereicht bei der FU Berlin. Unpublished Manuscript.

Gläser, J. and G. Laudel. 2001. Integrating scientometric indicators into sociological studies: methodical and methodological problems. *Scientometrics* 52: 411-434.

Gläser, J. and G. Laudel. 2005. Advantages and dangers of ‘remote’ peer evaluation. *Research Evaluation*, submitted.

Gläser, J., G. Laudel, S. Hinze, and L. Butler. 2002. *Impact of Evaluation-based Funding on the Production of Scientific Knowledge: What to Worry About, and How to* Find Out (Expertise für das BMBF), http://www.sciencepolicystudies.de/ ExpGlaeLauHinBut2809.pdf.

Gläser, J., T. H. Spurling and L. Butler. 2004. Intraorganisational evaluation: are there 'least evaluable units'? *Research Evaluation* 13: 19-32.

Godin, B. 2002a. Outline for a history of science measurement. *Science, Technology & Human Values* 27: 3-27.

Godin, B. 2002b. *The Social Sciences in Canada: What Can We Learn From Bibliometrics?* Project on the Management of the Social Sciences, Working Paper no. 1.

Gouguenheim, L., 1985. Comments on the paper, "A re-evaluation of the contributions to radio-astronomy of the Nancay Observatory". *4S Review* 3: 21-23.

Grigg, L. and P. Sheehan. 1989. *Evaluating Research: The Role of Performance Indicators.* Office of the Academic Director of Research, The University of Queensland.

Grupp, H. and M. E. Mogee, 2004. Indicators for national science and technology policy. H Moed, W. Glänzel and U. Schmoch (eds), *Handbook of Quantitative Science and Technology Research*. Dordrecht: Kluwer Academic Publishers, 75-94.

Harsanyi, M. A. 1993. Multiple authors, multiple problems - bibliometrics and the study of scholarly collaboration: a literature review. *Library Information Science Research* 15: 325-354.

Hattie, J. 1990. Performance indicators in education. *Australian Journal of Education* 34(3): 249-276.

Hattie, J., J. Tognolini, K. Adams and P. Curtis. 1991. *An Evaluation of a Model for Allocating Funds Across Departments Within a University using Selected Indicators of Performance*. Canberra: Department of Employment, Education and Training.

Herbertz, H. and B. Müller-Hill. 1995. Quality and efficiency of basic research in molecular biology: a bibliometric analysis of thirteen excellent research institutes. *Research Policy* 24: 959-979.

Hicks, D. 1999. The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. *Scientometrics* 44(2): 193-215.

Hicks, D. 2004. The four literatures of social science. H. Moed, W. Glänzel and U. Schmoch (eds) *Handbook of Quantitative Social Science and Technology Research*. Dordrecht: Kluwer Academic Publishers.

Hicks, D., T. Breitzman, K. Hamilton and F. Narin. 2000. Research excellence and patented innovation. *Science and Public Policy* 27: 310-320.

Hicks, D., P. Kroll, F. Narin, P. Thomas, R. Ruegg, H. Tomizawa, Y. Saitoh and S. Kobayashi. 2002. *Quantitative Methods of Research Evaluation Used by the U.S. Federal*

Government. Second Theory-Oriented Research group, National Institute of Science and Technology Policy (NISTEP), Japan.

Hinze, Sybille, and Ulrich Schmoch, 2004. Opening the Black Box. H. Moed, W. Glänzel and U. Schmoch (eds) *Handbook of Quantitative Social Science and Technology Research*. Dordrecht: Kluwer Academic Publishers, 215-235.

Hornbostel, S. 1997. Wissenschaftsindikatoren. Bewertungen in der Wissenschaft. Opladen: Westdeutscher Verlag.

Hornbostel, S. 1999. Welche Indikatoren zu welchem Zweck. M. Röbbecke and D. Simon(eds.), *Qualitätsföderung durch Evaluation? Input, Throughput, Output. Ziele, Aufgaben und Verfahren im Wandel*. Berlin: Wissenschaftszentrum Berlin, P99-003, 55-72.

Hornbostel, S. 2001. Third party funding of German universities. An indicator of research activity? *Scientometrics* 50: 523-537.

House of Commons Science and Technology Committee. 2004. *Research Assessment Exercise: A Re-assessment*, House of Commons, London.

Ingwersen, P. and L. Björneborn. 2004. Methodological issues of webometric studies. H. Moed, W. Glänzel and U. Schmoch (eds) *Handbook of Quantitative Social Science and Technology Research*. Dordrecht: Kluwer Academic Publishers, 339-370.

Irvine, J. and B. R. Martin. 1985. Evaluating big science: Cern´s past performance and future prospects. *Scientometrics* 7: 281-308.

Irvine, J., B. R. Martin, J. Abraham and T. Peacock. 1987. Assessing basic research: reappraisal and update of an evaluation of four radio astronomy observatories. *Research Policy* 16: 213-227.

Jauch, L. R. and W. F. Glueck. 1975. Evaluation of university professor's research performance. *Management Science* 22: 66-?

Jeschin, D., G. Lewison and J. Anderson. 1995. A bibliometric database for tracking acknowledgements of research funding. M. Koenig and A. Bookstein (eds), *Fifth International Conference of the International Society for Scientometrics and Informetrics*, New Jersey: Learned Information Inc.

Johnes, G. 1988. Research performance indications in the university sector. *Higher Education Quarterly* 42: 54-71.

Johnes, J. and G. Johnes. 1995. Research funding and performance in UK university departments of economics - a frontier analysis. *Economics of Education Review* 14: 301-314.

Katz, J. S. 1999. *Bibliometric Indicators and the Social Sciences.* Report prepared for UK Economic and Social Research Council, www.sussex.ac.uk/Users/sylvank /pubs/ESRC.pdf

Katz, J. S. and D. Hicks. 1997. How much is a collaboration worth? A calibrated bibliometric model. *Scientometrics* 40: 541-554.

King, J. 1987. A review of bibliometric and other science indicators and their role in research evaluation. *Journal of Information Science* 13: 261-276.

Korevaar, J. C. and H. F. Moed, 1996. Validation of bibliometric indicators in the field of mathematics. *Scientometrics* 37: 117-130.

Krige, J. 1993. Some socio-historical aspects of multinational collaborations in high-energy physics at CERN between 1975 and 1985. S. Sorlin (ed.), *Denationalizing Science: The Contexts of International Scientific Practice*, Dordrecht: Kluwer, 233-262.

Laudel, G. 2001. Collaboration, creativity and rewards: why and how scientists collaborate. *International Journal of Technology Management* 22: 762-781.

Laudel, G. 2002. What do we measure by co-authorships? *Research Evaluation* 11: 3-15.

Laudel, G. 2004. The 'quality myth': Promoting and hindering conditions for acquiring research funds. *Higher Education* (submitted).

Laudel, G. 2005a. Is external funding a valid indicator for research performance? *Research Evaluation*, submitted.

Laudel, G. 2005b. The ‘quality myth’: promoting and hindering conditions for acquiring research funds. *Higher Education* (in press).

Lawani, S. M. 1977. Citation analysis and the quality of scientific productivity. *Bioscience* 27: 26-31.

Lawani, S.M. 1986. Some bibliometric correlates of quality in scientific research. *Scientometrics* 9: 13-25.

Lawani, S. M. and A. E. Bayer. 1983. Validity of citation criteria for assessing the influence of scientific publications: new evidence with peer assessment. *Journal of the American Society for Information Science* 34: 59-66.

Le Pair, C. 1988. The citation gap of applicable science. A.J. F. Van Raan (ed.), *Handbook of Quantitative Studies of Science and Technology*. Amsterdam: Elsevier, 537-553.

Lee, C. K. 2003. A scientometric study of the research performance of the Institute of Molecular and Cell Biology in Singapore. *Scientometrics* 56: 95-110.

Lewison, G. 1998. Gastroenterology research in the United Kingdom: funding sources and impact. *Gut* 43: 288-293.

Lewison, G. 1999. The definition and calibration of biomedical subfields. *Scientometrics* 46: 529-537.

Lewison, G. 2001. Evaluation of books as research outputs in history. *Research Evaluation*, 10(2): 89-95.

Lindholm-Romantschuk, Y. and J. Warner. 1996. The role of monographs in scholarly communication: an empirical study of philosophy, sociology and economics. *Journal of Documentation* 52(4): 389-404.

Lindsey, D. 1980. Production and citation measures in the sociology of science: the problem of multiple authorship. *Social Studies of Science* 10: 145-162.

Lindsey, D. 1989. Using citation counts as a measure of quality in science: measuring what´s measurable rather than what´s valid. *Scientometrics* 15: 189-203.

Luwel, M., H. F. Moed, A.J. Nederhof, V. De Samblanx, K. Verbrugghen, L.J. Van der Wurff, 1999. *Towards Indicators of Research Performance in the Social Sciences and Humanities*, Leiden: CWTS.

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

MacRoberts, M. H. and B. R. MacRoberts. 1989. Problems of citation analysis: a critical review. *Journal of the American Society for Information Science* 40: 342-349.

Marginson, S. and M. Considine. 2000. *The Enterprise University: Power, Governance and Reinvention in Australia. Cambridge*, UK: Cambridge University Press.

Martin, B.R. 1994. British science in the 1980's - has the relative decline continued. *Scientometrics* 29: 27-56.

Martin, B. R. 1996. The use of multiple indicators in the assessment of basic research. *Scientometrics* 36: 343-362.

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

Martin, B. R. and J. Irvine. 1985. Evaluating the evaluators - a reply to our critics. *Social Studies of Science* 15: 558-575.

Martin, B. R., J. Irvine, T. Peacock and J. Abraham. 1985. A re-evaluation of the contributions to radio astronomy of the Nancay Observatory. *4S Review* 3: 4-18.

Martín-Sempere, M. J., J. Rey-Rocha and B. Garzón-García. 2002. Assessing quality of domestic, scientific journals in geographically oriented disciplines: scientists' judgements versus citations. *Research Evaluation* 11: 149-154.

Meadows, A. J. 1974. *Communication in Science*. London: Butterworths.

Meerterns, R. W., A. J. Nederhof and H. A. M. Wilke. 1992. Social psychological research in the Netherlands, 1980-1988. *European Journal of Social Psychology* 22: 93-100.

Meyer, M. 2000. Does science push technology? Patents citing scientific literature. *Research Policy* 29: 409-434.

Moed, H.F. and A. F. J. Van Raan, 1988. Indicators of research performance: applications in university research policy. A. J. F. Van Raan (ed.), *Handbook of Quantitative Studies of Science and Technology*. North-Holland: Elsevier, 177-192.

Moed, H.F. and M. Vriens. 1989. Possible inaccuracies occurring in citation analysis. *Journal of Information Science* 15: 95-107.

Moed, H.F. and F. Th. Hesselink, 1996. The publication output and impact of academic chemistry research in the Netherlands during the 1980s: bibliometric analyses and policy implications. *Research Policy* 25: 819-836.

Moed, H.F. and Th. N. Van Leeuwen. 1996. Impact factors can mislead. *Nature* 381: 186.

Meho, L. I. and D.H. Sonnewald. 2000. Citation ranking versus peer evaluation of senior faculty research performance: a case study of Kurdish scholarship. *Journal of the American Society for Information Science* 51(2): 123-138.

Moed, H.F., R.E. De Bruin and Th. N. Van Leeuwen, 1995. New bibliometric tools for the assessment of national research performance: database description, overview of indicators and first applications. *Scientometrics* 33: 381-422.

Moed, F., M. Luwel and A. J. Nederhof, 2002. Towards performance indicators in the humanities. *Library Trends* 50(3): 498-520.

Moed, Henk F., W. Glänzel and U. Schmoch, 2004. Editor's Introduction. H. Moed, W. Glänzel and U. Schmoch (eds) *Handbook of Quantitative Social Science and Technology Research*. Dordrecht: Kluwer Academic Publishers, 1-15.

Moed, H.F., J.M. Burger, J.G. Frankfort and A. F. J. Van Raan. 1985a. The application of bibliometric indicators: important field- and time-dependent factors to be considered. *Scientometrics* 8: 177-203.

Moed, H.F., W.J.M. Burger, J.G. Frankfort, and A. F. J. Van Raan, 1985b. A comparative study of bibliometric past performance analysis and peer judgement. *Scientometrics* 8: 149-159.

Moed, H.F., W.J.M. Burger, J.G. Frankfort, and A.F.J. Van Raan, 1985c. The use of bibliometric data for the measurement of university research performance. *Research Policy* 14: 131-149.

Moed, H. F., M. Luwel, J. A. Houben, E. Spruyt, and H. Van den Berghe, 1998. The effects of changes in the funding structure of the Flemish universities on their research capacity, productivity and impact during the 1980's and early 1990's. *Scientometrics* 43: 231-255.

Mojon-Azzi, S. M., X. Y. Jiang, U. Wagner and D. S. Mojon. 2003. Journals: redundant publications are bad news - publishing the same work twice is unethical and casts doubt on the integrity of research. *Nature* 421: 209-209.

Moravcsik, M. J. 1973. Measures of scientific growth. *Research Policy* 2: 266-275.

Moravcsik, M. J. and P. Murugesan. 1975. Some results on the function and quality of citations. *Social Studies of Science* 5: 86-92.

Nagpaul, P. S. and S. Roy, 2003. Constructing a multi-objective measure of research performance. *Scientometrics* 56: 383-402.

Narin, F. 1976. *Evaluative Bibliometrics: The Use of Publication and Citation Analysis in the Evaluation of Scientific Activity*. Cherry Hill, N.J.: Computer Horizons Inc.

Narin, F. 1994. Patent bibliometrics. *Scientometrics* 30: 147-155.

Narin, F. and K. S. Hamilton. 1996. Bibliometric performance measures. *Scientometrics* 36: 293-310.

Narin, F., T. Breitzman and P. Thomas. 2004. Using patent citation indicators to manage a stock portfolio. H. Moed, W. Glänzel and U. Schmoch (eds) *Handbook of Quantitative Social Science and Technology Research*. Dordrecht: Kluwer Academic Publishers, 553-568.

NBEET (National Board of Employment, Education and Training). 1993. *Research Performance Indicators Survey*. Commissioned Report No. 21. Canberra: National Board of Employment, Education and Training.

Nederhof, A.J. 1988. The validity and reliability of evaluation of scholarly performance. Anthony J. F. Van Raan (ed.), *Handbook of Quantitative Studies of Science and Technology*. North-Holland: Elsevier, 193-228.

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

Nederhof, A. J. and R. A. Zwaan., 1991. Quality judgements of journals as indicators of research performance in the humanities and the social and behavioural sciences. *Journal of the American Society for Information Science* 42(5): 332-340.

Nederhof, A. J. and E. C. M. Noyons. 1992. International comparison of departments' research performance in the humanities. *Journal of the American Society for Information Science* 43(3): 249-256.

Nederhof, A.J., and A.F.J. Van Raan, 1993. A Bibliometric Analysis of Six Economics Research Groups: A Comparison with Peer Review. Research Policy 22: 353-368.

Nederhof, A. J., M. Luwel and H.F. Moed. 2001. Assessing the quality of scholarly journals in linguistics: an alternative to citation-based journal impact factors. *Scientometrics* 51(1): 241-265.

Nederhof, A. J., R. A. Zwaan, R. E. De Bruin and P. J. Dekker. 1989. Assessing the usefulness of bibliometric indicators for the humanities and the social and behavioural sciences: a comparative study. *Scientometrics* 15(5-6): 423-435.

Nederhof, A. J., R. F. Meijer, H. F. Moed and A. F. J. Van Raan. 1993. Research performance indicators for university departments - a study of an agricultural university. *Scientometrics* 27: 157-178.

Nicol, M., L. Butler, and K. Henadeerage. 2004. *Using Citation Analysis to Analyse Track Record Scores given by Assessors of Grant Applications*, abstract for the 8th Science and Technology Indicators Conference, September 23-25 2004, Leiden.

Noyons, E. 2001. Bibliometric mapping of science in a science policy context. *Scientometrics* 50: 83-98.

Noyons, E. C. M. and A. F. J. Van Raan. 1998. Monitoring scientific developments from a dynamic perspective: Self-organized structuring to map neural network research. *Journal of the American Society for Information Science* 49: 68-81.

Oppenheim, C. 1997. The correlation between citation counts and the 1992 research assessment exercise ratings for British research in genetics, anatomy and archaeology. *Journal of Documentation* 53: 477-487.

Oppenheim, C. and M. Norris. 2003. Citation counts and the research assessment exercise V: archaeology and the 2001 RAE , *Journal of Documentation* 56(6): 709-730

Oromaner, M. 1985. The ortega hypothesis and influential articles in American sociology. *Scientometrics* 7: 3-10.

Perrin, B. 1998. Effective use and misuse of performance measurement. *American Journal of Evaluation* 19: 367-379.

Phelan, T. J. 1999. Issues for citation analysis. *Scientometrics* 45:131.

Phillimore, A.J. 1989. University research performance indicators in practice: the university grants committee's evaluation of British universitites, 1985-86. *Research Policy* 18: 255-271.

Porter, A. L., D. E. Chubin, and X.-Y. Jin. 1988. Citations and scientific progress: comparing bibliometric measures with scientist judgements. *Scientometrics* 13: 103-124.

Print, M. and J. Hattie. 1997. Measuring quality in universities: an approach to weighting research productivity. *Higher Education* 33: 453-469.

Rinia, E. J., T. N. Van Leeuwen, H. G. Van Vuren and A. F. J. Van Raan, 2001. Influence of interdisciplinarity on peer-review and bibliometric evaluations in physics research. *Research Policy* 30: 357-361.

Rinia, E. J., T.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.

Rinia, E. J. 2000. Scientometric studies and their role in research policy of two research councils in the Netherlands. *Scientometrics* 47: 363-378.

Roberts, G. 2003. *Review of Research Assessment: Report by Sir Gareth Roberts to the UK Funding Bodies*, issued for consultation May 2003, http://www.ra-review.ac.uk/reports/roberts.asp

Rosenthal, M. C. 1994. The fugitive literature. H. Cooper and L. V. Hedges (eds) *The Handbook of Research Synthesis*. New York: Russell Sage Foundation Publications.

Royle, P and R. Over. 1994. The use of bibliometric indicators to measure the research productivity of Australian academies. *Australian Academic and Research Libraries* 25(2): 77-88.

Salzarulo, L. and M. von Ins, 2001. Bias, structure and quality in citation indexing. *Scientometrics* 50: 289-299.

Schloegl, C., J. Gorraiz, C. Bart, and M. Bargmann. 2003. Evaluating two Austrian university departments: lessons learned. *Scientometrics* 56: 287-299.

Schubert, A. and W. Glänzel. 1983. Statistical reliability of comparisons based on the citation impact of scientific publications. *Scientometrics* 5: 59-74.

Schubert, A., W. Glänzel and T. Braun. 1989. Scientometric datafiles - a comprehensive set of indicators on 2649 journals and 96 countries in all major science fields and subfield, 1981-1985. *Scientometrics* 16: 3-478.

Seglen, P. O. 1992. The skewness of science. *Journal of the American Society for Information Science* 43: 628-638.

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

Seglen, P. O. 1997. Why the impact factor of journals should not be used for evaluating research. *BMJ* 314: 498-502.

SJTU, Shanghai Jiao Tong University. 2004. Academic Ranking of World Universities. Accessed 31 January 2005 at: http://ed.sjtu.edu.cn/rank/methodology.htm

Small, H.G. 1977. A co-citation model of a scientific specialty: a longitudinal study of collagen research. *Social Studies of Science* 7: 139-166.

Small, H.G. and D. Crane. 1979. Specialties and disciplines in science and social science: an examination of their structure using citation indices. *Scientometrics* 1(5-6): 445-461.

Smith, A. and M. Thelwall. 2002. Web impact factors for Australasian universities. *Scientometrics* 54: 363-380.

Smith, A. and M. Eysenck. 2002. The correlation between RAE ratings and citation counts in psychology. Unpublished Manuscript.

Smith, L. C. 1981. Citation analysis. *Library Trends* 30.

Snizek, W.E., 1987. In search of influence: the testing of the ortega hypothesis. *Scientometrics* 12: 311-314.

Steele, C. 2004. 'Ranking mania reflects distortion of priorities', in *The Australian Higher Education Supplement*, November 10 2004.

Strand, D. 1998. *Research in the Creative Arts*. Canberra: Department of Employment, Education Training and Youth Affairs.

Thelwall, M. 2001. Extracting macroscopic information from web links. *Journal of the American Society for Information Science and Technology* 52: 1157-1168.

Thelwall, M. and G. Harries. 2004. Do the web sites of higher rated scholars have significantly more online impact? *Journal of the American Society for Information Science and Technology* 55: 149-159.

Thomson ISI, 2004. *The ISI® Database: The Journal Selection Process*, http://www.isinet.com/essays/selectionofmaterialforcoverage/199701.html/

Tijssen, R.J. W., 2003. Scoreboards of research excellence. *Research Evaluation* 12: 91-103.

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: 381-397.

Tognolini, J., K. Adams and J. Hattie. 1994. A methodology to choose performance indicators of research attainment in universitites. *Australian Journal of Education* 38: 105-117.

Toutkoushian, R. K., S. R. Porter, C. Danielson and P. R. Hollis, 2003. Using publications counts to measure an institution's research productivity. *Research In Higher Education* 44: 121-148.

Van Leeuwen, Th. N. and H. F. Moed. 2002. Development and application of journal impact measures in the Dutch science system. *Scientometrics* 53: 249-266.

Van Leeuwen, T. N., L. J. Van der Wurff and A. F. J. Van Raan. 2001. The use of combined bibliometric methods in research funding policy. *Research Evaluation* 10: 195-201.

Van Leeuwen, Th. N., M. S. Visser, H. F. Moed and A. J. Nederhof. 2002. *Bibliometric Profiles of Academic Chemistry Research in the Netherlands 1991-2000*. Centre for Science and Technology Studies (CWTS).

Van Leeuwen, Th. N., M. S. Visser, H. F. Moed , T. J. Nederhof and A. F. J. Van Raan, 2003. The Holy Grail of science policy: Exploring and combining bibliometric tools in search of scientific excellence. *Scientometrics* 57: 257-280.

Van Raan, A.F.J. (ed.). 1988. *Handbook of Quantitative Studies of Science and Technology*. Amsterdam, The Netherlands: Elsevier Science Publishers B.V.

Van Raan, A. F. J. 1993. Advanced bibliometric methods to assess research performance and scientific development: basic principles and recent practical applications. *Research Evaluation* 3: 151-166.

Van Raan, A. J. F. 1996. Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics* 36: 397-420.

Van Raan, A. J. F. 1998. In matters of quantitative studies of science the fault of theorists is offering too little and asking too much - comments on: theories of citation? *Scientometrics* 43: 129-139.

Van Raan, A. F. J. 1998. Assessment of social sciences: the use of advanced bibliometric methods as a necessary complement of peer review. *Research Evaluation* 7(1): 1-6.

Van Raan, A. J. F. 2000. The Pandora's Box of Citation Analysis: Measuring Scientific Excellence - The Last Evil? B. Cronin and H. Barsky Atkins (eds.), The Web of Knowledge. Medford, New Jersey: Information Today, Inc, 301-319.

Van Raan, A. F. J. 2003. The use of bibliometric analysis in research performance assessment and monitoring of interdisciplinary scientific developments. *Technikfolgenabschätzung: Theroie und Praxis* 12(1): 20-29.

Van Raan, A. J. F. 2004. Measuring science. H. Moed, W. Glänzel and U. Schmoch (eds) *Handbook of Quantitative Social Science and Technology Research*. Dordrecht: Kluwer Academic Publishers, 19-50.

Van Raan, A. J. F. 2005. Fatal attraction: conceptual and methodological problems in the ranking of universities by bibliometric methods, *Scientometrics* 62(1): 133-143.

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

Verbeek A., K. Debackere, M. Luwel and E. Zimmerman. 2002. Measuring progress and evolution in science and technology I: the multiple uses of bibliometric indicators, *International Journal of Management Reviews* 4: 179-211.

Villagrá Rubio, A. 1992. Scientific production of Spanish universities in the fields of social science and language. *Scientometrics* 24 (1): 3-19.

Vinkler, P. 1986. Evaluation of some methods for the relative assessment of scientific publications. *Scientometrics* 10: 157-177.

Vinkler, P. 1988. An attempt of surveying and classifying bibliometric indicators for scientometric purposes. *Scientometrics* 13: 239-259.

Vinkler, P. 1994. The origin and features of information referenced in pharmaceutical patents. *Scientometrics* 30: 283-302.

Virgo, J. A., 1977. Statistical procedure for evaluating importance of scientific papers. *Library Quarterly* 47: 415-430.

Warner, J. 2000. A critical review of the application of citation studies to the research assessment exercises. *Journal of Information Science* 26: 453-460.

Webber, P. 1994. Creative arts as research. C. Cusack (ed.) *Creative Arts and Research*, Centre of Creative Arts: University of Sydney.

Weingart, P. 2003. Evaluation of research performance: the danger of numbers. Paper presented at the 2nd Conference of the Forschungszentrum Jülich GmbH Zentralbibliothek "*Bibliometric Analysis in Science and Research: Applications, Benefits and Limitations*", Jülich, 5-7 November.

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

Weingart, P. and M. Winterhager. 1984. *Die Vermessung der Forschung. Theorie und Praxis der Wissenschaftsindikatoren*. Frankfurt: Campus.

Weingart, P., R. Sehringer and M. Winterhager. 1988. Bibliometric indicators for assessing strength and weaknesses of West German science. A. J. F. Van Raan (ed.), *Handbook of Quantitative Studies of Science and Technology*. North-Holland: Elsevier, 391-430.

Wood, F. Q. 1989. Assessing research performance of university academic staff: measures and problems. *Higher Education Research and Development* 8: 237-250.

Wooding, S. and J. Grant. 2003. *Assessing Research: The Researchers' View*, Santa Monica, Calif.: RAND.

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

Wouters, P. F. 1999. The Citation Culture. PhD thesis, University of Amsterdam.

Zuckerman, H. A. 1968. Patterns of name ordering among authors of scientific papers: a study of social symbolism and its ambiguity. *American Journal of Sociology* 74: 276-291.

Zuckerman, Harriet, 1977. *Scientific Elite: Nobel Laureates in the United States*. New York: Free Press.