Organisation de Coopération et de Développement Économiques
Organisation for Economic Co-operation and Development

04-Jun-2010

English - Or. English

DIRECTORATE FOR SCIENCE, TECHNOLOGY AND INDUSTRY
COMMITTEE FOR SCIENTIFIC AND TECHNOLOGICAL POLICY

Performance indicators used in performance-based research funding systems

Professor Hanne Foss Hansen (University of Copenhagen)

Monday 21 June 2010

Delegates will find attached a paper commissioned by the OECD Secretariat (Directorate for Science, Technology and Industry) for the OECD-Norway Workshop on Performance-Based Funding of Public Research in Tertiary Education Institutions. It was prepared by Professor Hanne Foss Hansen (University of Copenhagen) and will be presented under Item 4 of the draft agenda for the workshop [DSTI/STP/RIHR/A(2010)1]. It discusses the indicators used in performance-based funding systems for research. The paper is for discussion at the workshop.

Sarah BOX: Tel: (+33 1) 45 24 18 69; Fax: (+33 1) 44 30 62 64; Email: sarah.box@oecd.org
Ester BASRI: Tel: (+33 1) 45 24 96 24; Fax: (+33 1) 44 30 62 64; Email: ester.basri@oecd.org

JT03284866

DSTI/STP/RIHR(2010)4 Unclassified
English - Or. English

PERFORMANCE INDICATORS USED IN PERFORMANCE-BASED RESEARCH FUNDING SYSTEMS

PERFORMANCE-BASED FUNDING FOR PUBLIC RESEARCH IN TERTIARY EDUCATION INSTITUTIONS: WORK BLOCK B[1]

Hanne Foss Hansen
Department of Political Science
University of Copenhagen
hfh@ifs.ku.dk

1. Introduction

1. In recent years still more governments have developed performance-based funding systems in tertiary education in relation to both education and research activities (Geuna & Martin, 2003; Whitley & Gläser, 2007; Frölich, 2008). This paper analyses and discusses performance indicators used in performance-based research funding systems (PRFS) introduced by governments to allocate funds for public research to tertiary education institutions.

2. In addition to this introduction the paper holds three sections. Section two gives an overview of the variety of indicators. Departing from a brief overall presentation of performance indicators as discussed in public management literature, the discussion explores different types of research indicators such as ratings by peer review panels, indicators reflecting institutions abilities to attract external funding, as well as results indicators (*e.g.* numbers of publications, citations and patents). Section three holds an analysis of how national performance-based funding systems are constructed, *e.g.* which indicators are used, how indicators are weighted, which data sources are used and whether systems differentiate in the use of indicators across fields etc. Finally section four holds a discussion on consistency in performance measurement, *e.g.* how the use of quantitative indicators compares to peer review processes in terms of capturing performance. The paper is concluded by pointing to knowledge gaps in the form of problems where further analysis usefully could be carried out.

3. The focus in the paper is on indicators used in funding systems based on ex-post evaluation. Foresight methods and other strategies for identifying knowledge requirements are not explored. Also in

[1] This paper on performance indicators constitutes work block B in the OECD's Working Party on Research Institutions and Human Resources project "Performance-based funding for public research in tertiary education institutions". Besides block B, the project includes a work block A (DSTI/STP/RIHR(2010)3) concerning models of performance-based funding systems and a work block C (DSTI/STP/RIHR(2010)6) concerning country experiences in the use of performance-based funding systems. It is not possible to give an overview of indicators used in performance-based funding systems without touching upon different models of systems. There is thus some overlap between this paper and the block A paper written by Diana Hicks. We have however tried to avoid unnecessary overlap.

focus are government funding formulas for institutions. This has two implications. First, that funding systems anchored in individual contracts between governmental agencies and institutions are not explored. Second, that governmental project and program funding as well as research council project and program funding are not explored, neither funding formulas used within institutions. Further as the focus is on funding systems, national research evaluation systems not directly linked to funding are not explored.

4. To analyze the use of indicators in performance-based research funding systems can be compared to shooting at a fast moving target. Systems are steadily redesigned. Awareness of which system versions are discussed is very important.

2. The concept of performance and the rich world of indicators

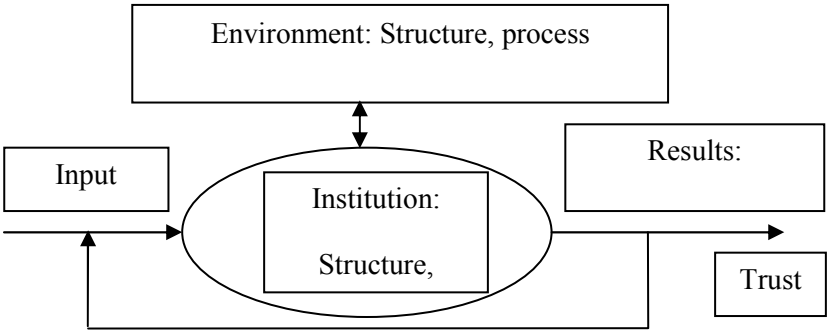
5. Performance-based management has become common in the public and non-for-profit sectors. There is a variety of terms used in the literature on this topic. Besides performance-based management we talk about results management, result-based management, managing for results, managing for outcomes, outcome-focused management and performance management. The intent behind this movement is to measure what results are brought about by institutions and to use that information to help better manage public funds and to better report on the use of those funds (Mayne, 2010).

6. The concept of performance is not unitary and do not have an unambiguous meaning. Rather performance must be viewed as information about achievements of varying significance to different stakeholders (Bouckaert and Halligan, 2008). In PRFSs the focus is on organizational performance. Performance is viewed as activity and results and as creating societal value through knowledge production. The underlying idea of PRFSs is that tertiary education institutions may lose sight of the results they intend to achieve if they are not held accountable the hard way by coupling performance to resource allocation (Talbot, 2005). Or to put it the other way around, PRFSs constitute incentives thought to improve performance. In addition PRFSs are anchored in a belief in the possibilities of defining and measuring research (and to some extent also research linked) performance. As will be shown beneath this is challenging in practice. First of all, there are very many different indicators. Secondly, indicators are proxies and in many respects our knowledge about the uncertainty of the proxies is inadequate.

Measuring performance

7. How can we conceptualize performance and especially research performance? Figure 1 presents a simple systemic framework illustrating the complexity of organizational performance.

Figure 1: A systemic framework for organizational performance



8. According to the systemic framework illustrated in Figure 1, tertiary education institutions are viewed as systems of knowledge production transforming input (funding and other resources) to output (*e.g.* publications and Ph.D. graduates). Knowledge production processes take place in an organizational context constituted by both intra-organizational structure (*e.g.* research groups and departments) and inter-organizational environmental structures (*e.g.* disciplinary and interdisciplinary networks).

9. There is an ongoing discussion in the academic community about how to define quality in research outputs. Quality is a multi-dimensional phenomenon including aspects such as originality, solidity and informativity (see for example Hemlin, 1991 and Seglen, 2009). Further outputs are not an end in themselves. Results in the form of effect (sometimes also termed outcome or impact) and benefits are crucial goals. On these dimensions there are many stakeholders. One is the scholarly community and the contribution research makes to the advancement of scientific-scholarly knowledge. Others are different groups in society and the contribution research makes to educational, economical, environmental, cultural and other dimensions of societal development. Research contributions are not always received positively. New knowledge may be critical, provocative or even (considered) harmful. The concept of research performance is not only multi-dimensional and ambiguous. It is may also be charged with conflict.

10. There is a span between output on the one side and effects not to say benefits on the other. First of all there is a span in time, for example, from knowledge produced to knowledge published and knowledge used. Secondly there are disconnections. For example knowledge may be produced but not published. Also knowledge may be published but not used. And there is probably an even greater span between these dimensions and the trust of citizens in research institutions. Trust is necessary to maintain public funding in the long run. PRFSs not only aim at creating incentives for and assuring productivity and effectiveness. They may also play an accountability role aiming at assuring trust.

11. In PRFSs research performance is measured by indicators. There are three main categories of research indicators: 1) first order indicators directly aiming at measuring research performance by focusing on measuring input, processes, structure and/or results; 2) second order indicators defined as easy to go summarizing indexes aiming at offering simple measures for effect (*e.g.* journal impact factor and the H index); and 3) third order indicators brought up by peer review panels rating for example departments.

12. First and second order indicators, also referred to as metrics, may be used directly and mechanically in funding systems but they may also be used as part of the input to peer review processes,

and thus be input to the production of third order indicators. In the following the content, potentials and limitations in the three categories of indicators will be discussed. In relation to the rich world of first order indicators details about the indicators are presented in tables in order not to make the text very technical.

First order indicators

13. One way to make a typology of first order research indicators is to distinguish between indicators concerning input, process, structure and results, which again can be divided into output and effects. In the following, important examples of such indicators are discussed. Primary sources are Cave *et al.* (1991), Hansen & Jørgensen (1995), Dolan (2007), Hansen (2009) and European Commission (2010).

14. Important input indicators are presented in Table 1.

Table 1: Input indicators

Input indicators	Potentials	Limitations
External funding	Ability to attract external funding may be measured as the amount of external research income, perhaps income per full time equivalent research active staff, and/or as the number and percentage of competitive grants won from selected sources (peer reviewed external research income, international versus national funding, grants from government including research councils versus grant from industry or foundations). External funding indicators tell us something about institutions competitiveness on funding markets and when defined as peer reviewed external funding include and aspect of quality.	Competitiveness does not fully coincide with quality. Reputation and networks also play roles. In addition levels of external funding vary greatly across scientific fields, disciplines and research areas. Differences in levels of external funding combined with differences across institutions in profiles, *e.g.* whether institutions have a large medical faculty with good possibilities of attracting external funding or a large faculty of arts with less good possibilities, severely limit possibilities for fair cross-institutional comparison.
Recruitment of PhD students and academic staff	Indicators related to the ability to attract students and staff tell us something about institutions competitiveness on labour markets and graduates and applicants assessment of how attractive the research environments are. Depending on purpose the number or share of highly qualified and/or international candidates can be counted.	Patterns for applying university posts are influenced by many other factors than how attractive research institutions are experienced in the scholarly community. Research institutions compete not only with other research institutions but also with other career paths.

15. Indicators related to research institutions ability to attract input in the form of funding from the environment tell us something about their competitiveness on funding markets. If measured as the ability to attract peer reviewed external funding indicators in addition tell us something about institutions reputation and performance so far as well as about the scientific quality and relevance of their research plans. Input indicators do not however fully coincide with quality and performance. Networks for example play a role. In relation to indicators of external funding possibilities for making comparisons across scientific fields are limited because levels of external funding differ across fields.

16. Indicators related to the ability to attract PhD students and staff tell us something about institutions competitiveness on labour markets as well as about graduates and applicants assessment of how attractive the research environments are. Application patterns are, however, influenced also by other factors such as for example how attractive other career paths are.

17. Important process indicators are presented in Table 2.

Table 2: Process indicators

Process indicators	Potentials	Limitations
Seminar and conference activity	Number of arranged seminars and conferences as well as number of external conference participations can be indicators of research intensity.	Conference activity may reflect research tourism.
Invited keynotes	Counting the number of invited keynote addresses given at national and international conferences may be used as a proxy for quality, impact and peer-esteem.	Invited keynote addresses may reflect networks rather than quality. No agreed equivalences applying internationally. No possibilities to compare across disciplines.
International visiting research appointments	Counting the number of visiting appointments may be used as a proxy for peer-esteem.	Visiting appointments may reflect networks rather than peer-esteem. No agreed equivalences applying internationally. No possibilities to compare across disciplines.

18. Process indicators may focus on in-house activities such as the number of arranged seminars and conferences as well as visiting international distinguished academic guests or on out-house activities such as abroad conference participation and invited keynotes and other lectures.

19. Important structure indicators are presented in Table 3.

Table 3: Structure indicators

Structure indicators	Potentials	Limitations
Research active staff	Size of research active staff is often regarded as an indicator of research capability and intensity. The indicator may be sophisticated by measuring shares of highly research active staff, e. g. by setting threshold levels of performance for a specific period and counting the number of academics at different levels.	No clear definitions of threshold levels.
Number of PhD students	Number of PhD students is also an indicator of research capability and intensity as active research cultures attract students. Can be measured as mean number per full time equivalent research active staff.	Disciplines may give different priority to PhD activity.
Research collaborations and partnership	As research is increasingly conducted in collaborative teams a count of national and international collaborations with other research institutions can be an indicator of research involvement and scale of activity. Research collaborations may be assessed on the degree to which they result in different types of co-publication (e. g. national, international, interdisciplinary).	Collaboration is many things. Collaboration may be loose or intensive and mutually binding. Collaboration may involve different institutions, *e.g.* university-university, university-external stakeholder. Collaboration and publication patterns differ across fields.
Reputation and esteem	Positions as journal editors, membership of editorial boards and scientific committees and fellowship of learned academies are often regarded as indicators of the extent to which researcher's opinions are highly	May reflect networks rather than recognition. No agreed equivalences applying internationally. No possibilities to compare across disciplines.

	regarded by the academic community.	
Research infrastructure and facilities	Research laboratory facilities, library facilities (books and electronic journal access), computing facilities, support staff, general working conditions etc.	Many indicators in one. No easy access to valid comparable data.

20. Structure indicators may focus on internal aspects such as the share of research active academics and research infrastructure or on external aspects such as collaborations and partnership, reputation and esteem. The possibilities of comparing across disciplines and fields are limited due to differences in facilities and collaboration patterns.

21. Important results indicators are presented in table 4 concerning output indicators and table 5 concerning effect indicators.

Table 4: Output indicators

Result indicators: Output	**Potentials**	**Limitations**
Publications	Publishing is vital for progress in research. If counted per full-time equivalent academic staff cross-institutional comparison may be supported. Depending on purpose selected types of publication can be counted, *e.g.* percentage of journal articles published in high-ranked journals.	Emphasis on quantity and productivity. Different disciplines produce different types of outputs (journal articles, books, proceedings etc.). Rating and ranking of journals is not an unambiguous task.
Non-bibliographical outputs	In some fields non-bibliographical outputs such as artworks, music or performances are important.	Due to the heterogeneous character these outputs are not easily measured.
Number of PhD graduates and completion rates for graduates.	New generations of researchers are vital for continuing progress in research. Counting of PhD graduates may be supplemented by a measure of the share of PhD graduates finishing in due time putting focus on process effectiveness in PhD programs.	Disciplines may give different priority to PhD activity and rates of completion may differ across disciplines. Recruitment as well as external employability may affect through-put.
Public outreach	Measures can be developed for the visibility of researchers in society, *e.g.* in the media.	Media visibility may be very loosely coupled to research activities.

22. Output may be measured by counting publications, non-bibliographical outputs, PhD graduates and different kinds of public outreach. Publishing is vital for progress in research but publication patterns differ across fields. Normally publications are counted in groups and most often focus is on peer-reviewed publications as these are experienced as including an aspect of quality. In some fields *e.g.* the social sciences and the humanities, a distinction between the shares of national respectively international publications may be relevant. Often journal articles are counted in groups reflecting the ranking of journals. In some fields there is high agreement on journal rankings in others only limited agreement.

Table 5: Effect indicators

Result indicators: Effect	Potentials	Limitations
Citations	Citations tell us something about scholarly impact and visibility. Databases such as Web of Science, Scopus and Google Scholar make citation counting possible.	Citations do not fully coincide with research quality. Not all disciplines and research areas are equally well covered in citation indexes. Especially the humanities and parts of the social sciences and engineering are not well covered.
Number of awards and prizes	Indicator of research quality and impact.	No agreed equivalences applying internationally except for Nobel prizes. No possibilities to compare across disciplines.
Employability of PhD graduates	Industry and governmental employment of PhD graduates can be an indicator of the quality of the graduates and the contribution of research to industry and society.	Employability is sensitive to other factors, *e.g.* regional or national economy. Career paths differ across disciplines.
Knowledge transfer and commercialisation of research-generated intellectual property (IP)	Measure of the extent of income created through patents, licences or start ups. Important link between IP, commercialization and economic benefits.	Patents are a poor indicator of IP and sensitive to both discipline and national context.
End-user Esteem	Commissioned reports, consultancy and external contracts are measures of the willingness of external stakeholders to pay for and use research. Such measures, *e.g.* counted as the amount and percentage of funding from end-users (*e.g.* industry, professions, government, community) are thus indicators of the anticipated contribution to innovation.	Different opportunities for different disciplines. Networks influence funding possibilities.

23. Effect may be measured by indicators related to citation counting, awards, employability of PhD graduates, commercialization activities as well as end-user esteem.

24. Citation count is an effect indicator in that it indicates how often publications and therefore, how often researchers are cited by other researchers. Since researchers cite each other for a variety of reasons, it is, however, debatable what citation counts actually measure. Applicability, visibility and impact are central aspects. To the extent that researchers cite each other because they are building on other researchers' ideas and results, there is a quality dimension to citation counts. But how large is this quality dimension? Citation behaviour can also be argumentative (*i.e.* selective as support for the researcher's own viewpoint). It can be used to disagree or to flatter, just as it can be based on a desire to show insight to a subject area (Seglen, 1994b). The intention behind citation counts is often to measure quality, but the information derived from them is more about communication structures and professional networks.

25. In the academic community there is an interesting ongoing discussion on the use and misuse of citation statistics. In 2008 the Joint IMU/ICIAM/IMS-Committee on Quantitative Assessment of Research[2] published a report the limitations of citation statistics and how better to use them. The background for

[2] IMU = International Mathematical Union, ICIAM = International Council of Industrial and Applied Mathematics, IMS = Institute of Mathematical Statistics.

writing the report was the observation that the drive for more transparency and accountability in the academic world has created a culture of numbers. The committee wrote: "Unable to measure quality (the ultimate goal), decision-makers replace quality by numbers that they can measure. This trend calls for comment from those who professionally deal with numbers – mathematicians and statisticians" (Adler *et al.*, 2008). The advice of the committee besides the need for consulting statisticians when practicing citation analyses is to scant attention to how uncertainty affects analysis and how analysis may reasonable be interpreted. The report has been commented by other experts agreeing on the common misuse of citation data but also pointing at ways to make meaningful analysis first and foremost in relation to identifying and comparing the performance of comparable groups of scientist (Lehmann *et al.*, 2009).

26. In addition, there are many challenges as regards technical measurement associated with citation counts. There are competing general databases. The most important are Thompson Reuters ISI Web of Science[3], Scopus[4] and Google Scholar[5]. Transparency in the coverage of the databases and the criteria that form the basis for the material that is included or excluded leave something to be desired. Due to the differences in the degree of coverage of the databases, searches on different databases often give very different results. Add to this the facts that the databases are prone to errors and that the degree of coverage varies from one research area to another. To derive and interpret citation data requires, therefore, a combination of bibliometric skills and specialist scholarly skills in the specific research field. The majority of publication and citation analyses composed by bibliometric experts have used Thompson Reuters ISI Web of Science, which is the oldest database in this field.

27. Since both publication and citation patterns vary considerably between research fields, the opportunities for comparison are severely limited. It is, therefore, recommended only to compare "like with like". If, for example, we wish to make a citation analysis related to research achievement at a university, we will have to normalize our data by, for example, calculating the average number of citations per article relative to the world average for individual research fields. By this means it is possible to show which subjects have greater or lesser impact than we could expect.

28. In the social sciences and the humanities citation counting meets special problems. Journal articles are less important in many disciplines within these scientific fields and citation analyses therefore produce partial pictures of performance (for the social sciences, problems and possibilities in citation analyses are thoroughly discussed in Hicks, 2006).

29. Indicators of knowledge transfer and commercialization have acquired additional interest in recent years as research and innovation policies increasingly have become integrated. Such indicators are for example concerned with licences and start-ups but may also be related to collaborative research, consultancy activities, networks as well as employability and employer satisfaction with PhD graduates (for an overview see Library House, 2007).

Second order indicators

30. As a reaction to the methodological challenges related to conducting citation analyses several index numbers have been developed that can easily be accessed in the databases. It can be tempting to make use of these, but they should be used with great caution. Unfortunately, this is not always the case.

---

3 http://apps.isiknowledge.com/UA_GeneralSearch_input.do?product=UA&search_mode=GeneralSearch&SID=P1@cMnn@MaHa@Hh4P3o&preferencesSaved=

4 http://info.scopus.com/etc/citationtracker/

5 http://scholar.google.com/intl/en/scholar/about.html

31. Two central index numbers are the so-called Journal Impact Factor (JIF) and the H index. JIF is a figure that gives the average number of citations achieved by articles published in a given journal within a given period – the so-called citation window. JIF is a key figure that says something about the characteristics of a journal. Experience tells us that there are substantial differences in the number of citations to individual articles (Seglen, 1994a). Even in journals with high JIFs, articles are to be found with no or only a few citations. For this reason, JIF should not be used mechanically for the ranking of researchers or research groups, for example.

32. The H-index is defined as the number of articles a researcher has published that have a citation figure equal to or higher than H. An H-index of 20 signifies, therefore, that a researcher has among his/her publications 20 that have each been cited at least 20 times. The H-index has been developed in recognition of the limitations of other citation measurements. A second-rate researcher can have a high total number of citations because he/she has published a "big hit" article with other researchers. To achieve a high H-index demands, however, continuous achievement at a high level over a period of years. But this also means that the use of the H-index only makes sense after 12-15 years of research. In addition, the H-index varies according to number of years of employment, subject and collaboration patterns.[6] The H-index does not then solve the problem mentioned above as regards comparison (Leuwen, 2008). In addition its reliability in general has been questioned and the mean number of citations per paper has been characterized as the superior indicator (Lehmann *et al.*, 2006).

Third order indicators

33. Third order indicators are as mentioned brought up by peer review panels rating, for example, departments. The term "peer review" is used to characterize research evaluation in which recognized researchers and experts are used as evaluators. Using concepts taken from the general literature of evaluation, peer review can be characterized as a collegial or professional evaluation model (Vedung, 1997; Hansen, 2005). The fundamental idea is that members of a profession are trusted to evaluate other members' activity and results on the basis of the profession's own quality criteria. Using peer review to produce indicators in PRFSs therefore built field differentiation into the system at the same time as all fields are treated alike in the evaluation process.

34. There is a whole range of forms of peer review (see, for example, Hansen and Borum, 1999; OECD, 2008). In this context it is useful to distinguish between classic peer review, modified peer review and informed peer review.

35. Classic peer review is an important mechanism for quality control and resource distribution at the micro-level in the research community. Through classic peer review, recognized researchers assess the scientific quality of manuscripts for articles, books and dissertations, and the qualifications of applicants to research posts are scrutinized. Classic peer review is also used in research council systems to assess whether applicants are eligible for support. Classic peer review is linked to clear decision-making situations. Judgements are provided as to whether "products" are worthy of support or publication and as to whether the applicant has the correct qualifications.

36. Classic peer review that takes place on the basis of the reading of research production is relational in the sense that the assessment is made in a context. A dissertation is assessed, for example, in relation to the research area to which it seeks to contribute, just as an applicant to a research post is assessed in relation to the job description that gives a level and a profile. It is most often the case that the

[6] To compensate for some of these problems other indexes have been proposed. The m-index dividing the h-index by the number of years since first paper is meant to compensate junior scientist. The g-index is meant to compensate for extraordinarily high citation counts (see Adler *et al.*, 2008).

process includes a form of cross-control of the assessment being made. On the one hand, there may be a number of peers either operating in parallel or on a panel. On the other, there are one or more "supreme judges". The assessment of a manuscript for an article is passed on to the editor, who reaches a decision. An assessment of an applicant is passed on for a decision to be made by management.

37. There is overall agreement that peer review is a solid method for the evaluation of scientific quality at the micro-level. This does not, however, mean that the method is infallible. There are differences between peers, and there is a degree of uncertainty associated with what in the literature is currently known as "the luck of the reviewer draw". In addition, studies have pointed out that in some contexts there are biases. Bias can be a matter of the "Matthew effect" in the sense that "to those who have, more shall be given" but bias can also be a matter of systematic unfair treatment or even discrimination on the grounds of gender, age, race or institutional attachment. Networks may make up for discrimination (Wennerås & Wold, 1997).

38. Over the course of time, other forms of peer review have been developed. After a tentative start in the 1970s, modified peer review has become a commonly used method e. g. in some types of PRFSs. As in classic peer review, modified peer review is also characterized by the fact that recognized researchers act as evaluators. But the task and the object for evaluation have changed. What is now in focus is the scientific quality of the production of the research organization. Modified peer review is most commonly organized as panel work. The panel members together have to cover a larger research area, and for that reason, each individual panel member is a specialist in subsidiary areas within the field that is to be covered. The basis for assessment most often includes selected publications, but other material such as lists of publications, statistics, annual reports and self-evaluations may constitute important background material. When modified peer review is supported by first and second order indicators we may speak of informed peer review.

Summing up on indicators

39. As shown above there is a rich world of research indicators. Research indicators are not objective measures of performance. They are proxies and the knowledge on the ambiguity of most of the proxies is limited. For example the knowledge on how networks shape the measures of institutional competitiveness on external peer reviewed funding markets as well as measures of end-user esteem is limited.

40. The rich world of indicators seems to be steadily expanding. Indicator producers are creative actors. There are several reasons why the world of research indicators expands.

41. One reason has to do with the ambiguity of indicators. As indicators are proxies characterized by both potentials as well as limitations the strength of indicators are constantly debated. This seems to give rise to a dynamic of ongoing attempts to mend existing indicators compensating their weaknesses by developing new ones with other (weaknesses).

42. Another reason has to do with the development of research policy. Back in history research policy was so to speak built into other policy fields first and foremost higher educational policy but also sector policies. But after the Second World War research policy increasingly has developed into an independent policy field. Even though research and teaching is still tightly interwoven at tertiary education institutions, institutions to a large extent are met with separate and independent policy streams related to higher education and research respectively. The development of more independent research policy fields at both the international as well as national levels has given rise to the development of research indicators. In recent years the development of research indicators has been furthered also by the integration of innovation policy into research policy. This development has given rise to new types of indicators related to knowledge transfer and commercialization.

43. Because of differences across disciplines and research areas as well as differences in institutional profiles great care should be taken in using indicators in comparisons. Nevertheless the aim of PRFSs precisely is comparison. Lets us look at how PRFSs have been constructed historical and which indicators are used in systems currently in use.

3. The construction of national PRFSs

44. Above, the variety of indicators has been discussed. In this section focus is on how PRFSs deal with indicators. First, a brief presentation of the historical development of PRFSs is given. Next the analysis of indicators in PRFSs currently in use is presented.

45. As will be shown PRFSs historically have been either third order indicator models anchored in peer review as the main principle or first order indicator models anchored in monitoring input and output as the main principle. The analysis of indicators in PRFSs currently in use will explore whether this is still the case or whether PRFSs these years are changing character.

The historical background

46. Historically two different types of PRFSs have been in use each of them anchored in different indicators. One type PRFS has been anchored in the use of third order indicators using a peer rating model. Another type has been anchored in the use of first order indicators monitoring research institutions input and output. Two countries have been pioneers in relation to developing these two types of systems. Figure 6 provides an overview of these two systems (for a more thorough comparison see Hicks, 2009).

Table 6: National performance-based funding system: Historical background

	Third order indicator model (Britain)	First order indicator model monitoring input and output (Australia)
Organisation responsible	Higher Education Funding Council for England (HEFCE) among others.	Australian Government: Department for Education, Employment and Workplace Relations
Object of evaluation	Departments (staff actively involved in research hand in publications)	Institutions
Method	Peer review resulting in departmental rating. Peer panel structure as well as rating scales have varied. Rating is subsequently turned into distribution of funding.	Indicators resulting in distribution of funding
Frequency	Exercise conducted 1986, 1989, 1992, 1996, 2001 & 2008.	Annual cycle

47. The third order indicator model was developed in England in 1986. The aim was to maintain research excellence by introducing selectivity in funding allocation in an era where the higher educational system was expanding. In 2001 the system was adopted also in Scotland, Wales and Northern Ireland. The British system called the "Research Assessment Exercise" (RAE) [7] is based on a large number of peer panels (in 2008 all in all 67 panels) each assessing and rating the quality of research at all departments in a

[7] Until 1992 the system was conducted under the heading "research selectivity exercises". The RAE system was conducted for the last time in 2008. A new system based upon a combination of peer review and research indicators is being developed.

discipline or given research area. The assessment of the quality of research is among other things based on publications handed in by academic staff. The sixth and last assessment round of RAE was conducted in 2008 and will inform research funding in 2009-2010. A new system called "Research Excellence Framework" (REF) is being developed (see below).

48. The first order indicator model monitoring input and output was developed in Australia in 1990. The system, which is still in use while a new system is being developed (see below) monitors four indicators: *i)* institutions' ability to attract research grants in competition (input), *ii)* number of publications (output), *iii)* number of Masters and PhD students (stock) and *iv)* number of Masters and PhD students finishing on time (output and through put). The system has been applied uniformly across all research areas using a common list of the types of grants and publications that count (Gläser and Laudel, 2007).

49. In 2005-2007, a second generation system, called the Research Quality Framework (RQF) was developed. RQF was a RAE-like system but included in addition assessments by end-users of the impact of research on the economy and society. The RQF became controversial as it was experienced to lack transparency and having very high implementation costs. When a new Government took over in late 2007 RQF was abandoned ever before it was implemented. A third generation system, called Excellence in Research for Australia (ERA) is now being developed. ERA is planned to use first order indicators as input to third order peer panel assessments.

50. The two systems the third order indicator model and the first order indicator model monitoring input and output have inspired other countries. RAE-like systems have been developed in Hong Kong and New Zealand and been proposed but not realized for example in Finland, Australia, Sweden and Denmark. First order indicator models monitoring input and output have been developed in for example Norway and Denmark.

51. These years a third model seems to gain ground. We could call this a first order indicator model monitoring effect. This model which it anchored in the idea of counting citations is being developed in the UK and in Sweden (elaborated below). The background for moving in this direction seems to be a wish for developing "stronger" and more "objective" PRFSs which focus not only on outputs but expand further into the chain of results focusing on effects. As discussed above it is however not obvious that effect indicators are stronger and more objective.

52. Historically the Flanders region of Belgium was in fact the first region to experiment with citation counts. In 2003 the Flanders replaced a funding formula formerly based on student numbers by a formula called the BOF-key weighting student numbers with publication as well as citation counts based on Web of Science data (Debackere and Glänzel, 2004). The BOF-key has been developed through the years giving more weight to publication and citation counting. For 2010 publications and citations are each weighted 17%. The Flanders model differentiates from other country models as it takes into account differences across different disciplines using the journal impact factor.[8]

53. The Flanders has experienced difficulties not being able to apply the system uniformly across all research areas. Especially subfields within the social sciences and the humanities, such as *e.g.* law and literature, have proved difficult to include. This is part of the background for the fact that Flanders these years seems to be moving towards a more output oriented model at least in the field of social sciences and humanities (see below).

54. In the following the indicators used in the PRFSs currently in use will be further explored.

[8] I have not been able to find accessible information about how this is exactly done.

First order indicator model: monitoring input and output

55. Table 7 gives an overview over countries currently using PRFSs anchored in first order indicators monitoring input and output as the main principle.

Table 7: First order indicator PRFSs monitoring input and output currently in use

Country	PRFS	Indicators	Weighting	Data sources	Differentiating
Norway	PRFS was established in 2005 where a research component partly performance-based was introduced as part of the overall funding system. The research component, in 2009 distributing 16% of total resources, is divided into a strategic part (*e.g.* scientific equipment) and a results-based part, called RBO.	RBO has four components: 1) Publications counted (adjustments for publication form, level and share of authorship), 2) PhD graduates 3) Ability to attract external fun-ding from the Norwegian Research Council and 4) EU.	Available sources do not clearly describe how the RBO components are weighted. As the publication component allocates just below 2% of the total budget, the components seem to be equally weighted.	National. A national database of publications has been developed.	No differentiation. The economic value of publication counts is equal across fields. Statistics indicate that different areas are treated fair as they get relatively similar impact in the research component.
Denmark	From 2010 increases in block grants for research at universities have to some extent been allocated across institutions using a PRFS.	Four: -ability to attract external funding, -publications counted (adjustments for publication form, level and share of authorship), -PhD graduates produced, -share of educational resources (which are also allocated on a performance criteria).	From 2012 components are weighted: -external funding 20% -publications 25% -PhD graduates 10%. -education share 45%.	National. A national database of publications has been developed.	The publication component is constructed in such a way that it does not alter the relative share of resources between the humanities, social science, natural/technical science and medical science. Resources are allocated conservatively across these four fields and hereafter allocated across institutions using publication counts. The economic value of publication counts thus differs across fields.

56. In 2005, Norway implemented a performance based funding model combining output indicators (counting of publications and PhD graduates) with an input indicator (external funding).

57. The publication indicator is based on complete data for the scientific publication output (in journals, series and books) at the level of institutions (Sivertsen, 2006 and 2009). The aim of the indicator is to measure and stimulate research activity. The data for the indicator are produced by institutions and included in a national database. The database makes comparable measurement possible as publications are weighted into publication points. Publications are weighted according to publications channels as well as level of quality. According to quality publication channels are divided into two levels, a normal level and a high level. The high level, which may not account for more than 20% of the world's publications in each field of research, includes leading journals and publishing houses. The list of high level journals and publishing houses has been produced by a large number of groups of peers and it is revised annually.

58. The publication indicator is used for allocating a smaller part of the total direct funding of research in combination with measures of PhD graduates produced and ability to attract external funding. The model is applied uniformly across all areas of research.

59. The Norwegian model has been of inspiration for the development of the Danish one. In Denmark allocation of annually increases in resources for block funding of research have for some years been based on a combination of input indicators (external funding and share of educational resources, also allocated on a performance criteria) and output indicators (PhD graduates). From 2010 this model is added a publication component. The aim of adding the publication component is to encourage researchers to publish in the most acknowledged scientific journals and to strengthen the quality of research. The publication component is very alike the Norwegian publication indicator. A national database is established and publications are divided into publication forms and levels according to lists of journals and publishing houses made by peer groups.

60. In defiance of similarities the Norwegian and the Danish models are implemented very differently probably resulting in different consequences. In Denmark the PRFS as mentioned is used annually to allocate the annual increase in block funding for research. It is a political decision each year how large this amount of resources are. Parts of increases in resources are generated by cutting back existing mostly historically budgeted block grants. There is a fear in the university sector that increases in resources risk drying out in the coming years due to the economic crisis. If this scenario turns out to become real the PRFS which in the present version has marginal importance may, if not re-designed, loose its direct importance.

61. The Norwegian PRFS annually reallocates a fixed volume of total block grants. As outputs related to both publications and PhD graduates have increased the income per output unit has decreased. This development has made the Ministry of Education and Research worry about whether incentives in the system are reduced across time (Kunnskapsdepartementet, 2010).

62. In Denmark the system is constructed in such a way as to not change the share of resources distributed to respectively the humanities, the social sciences, the natural/technical sciences and the medical sciences. The implication of this is that the economic value of publication counts differs across fields which again may differentiate the incentive effect of the system. In Norway there is no differentiation across fields but the experience so far shows that there is no noticeable reallocation across scientific fields.

63. The Norwegian model also seems to be of inspiration for the Flanders which has initiated a project developing a bibliographical database for the social sciences and the humanities. The database include different types of research outputs, including journal articles, books as author, books a editor, chapters in books as well as articles in proceedings. The database is planned to be one of the output indicators in the Flemish government's future research funding formula for the universities form 2012 where the BOF-key is to be re-negotiated.

First order indicator model: monitoring effect

64. Table 8 gives an overview over the Swedish PRFS in which an effect indicator is an important component.

Table 8: A first order indicator PRFS monitoring effects currently in use

Country	PRFS	Indicators	Weighting	Data sources	Differentiating
Sweden	Since 2009	Two: -bibliometric publication and citation counting indicator, -external funding (all external funding sources are treated with equal weight).	Bibliometrics and external funding are equally important indicators.	ISI Web of Science, publication and citation counts are field normalized.	Scientific fields are given different weights reflecting that they differ in propensity to score on citations as well as external funding.

65. In Sweden, a White Paper published in 2007 proposed to develop a performance-based model (SOU 2007). The aim was to allocate resources according to performance and quality aiming at stimulating quality. The proposal was to allocate the total amount of general government university funds across institutions on the basis of quality assessments of research (50%), measures of field normalized citations (20%), ability to attract external resources (20%), the share of PhDs among staff (5%) and the number of female professors (5%). As there at that time were only quality assessments done in a few tertiary higher education institutions it was proposed in the short run to allocate 50% of the available resources on the basis of the four indicators related to citations, external funding and staff.

66. In 2009 it was decided to introduce a modified system allocating resources on the basis of publication and citation counting as well as external funding (Carlsson, 2009; Sandström and Sandström, 2009). The staff elements including the gender balance were thus put aside. In the Swedish model, inspired by the at that time British plans to replace the RAE with a system producing robust UK-wide indicators of research excellence for all disciplines (see beneath), the bibliometric indicator based on Web of Science is weighted equally with the measures of external funding. An important aim has been to develop a model which is able to treat all research areas in the same process. In order to meet the challenges of discipline differences and differences in coverage in Web of Science, publication and citation counts are field normalized and publications from the social sciences and the humanities carry considerable more weight than publications in other areas. The result has been that the complexity of the model is high, while transparency is low except for bibliometric experts.

67. The Swedish model has turned out to be so controversial that the Swedish Research Council in 2009 urged the Ministry to suspend the model (Vetenskapsrådet, 2009). This has not happened but inquiries and consultations are going on as to how to proceed in the future.

68. The British experiences also show that developing PRFSs monitoring effects is not an easy task. In Britain the Higher Education Funding Council (HEFCE) has for some time worked with the development of a second generation PRFS, called the Research Excellence Framework (REF). Back in 2006 the government announced that a new system should replace RAE of 2008. At that time the idea was

to produce robust UK-wide indicators of research excellence for all disciplines. The plan was to produce the full set of indicators for the science-based disciplines during 2009 influencing funding allocations from 2010-11. For the arts and social sciences the plan was to phase the new system gradually in for some time continuing using peer review (Higher Education Funding Council, 2007).

69. Observers have characterized the idea as "a move away from the old "subjective" approach to RAEs towards more "objective" methods based on publication counts and citation measures to gauge quality and impact, plus statistical counts of external research income and postgraduate student activity" (Elzinga 2008). Experienced research policy advisors have expressed scepticism towards the idea and warned about the myth of a "trust in numbers" (Nowotny 2007).

70. During the development process HEFCE has asked bibliometric experts at the Centre for Science and Technology Studies (CWTS) at Leiden University for advice concerning the element of citation measures. A report published in 2008 shows that not only the arts and social sciences are partially covered in Web of Science database. Also parts of technical science and computer science are not well covered (Moed *et al.*, 2008).

71. Critical discussions and methodological challenges have forced HEFCE to modify and to some extent roll back plans. The new British PRFS will still be organized with peer panels. The number of panels will however be reduced. And the same goes for the number of publications handed in by academic staff. Also panels will make greater use of quantitative indicators including citation counts where possible. Panels will be asked to rate departments weighting research quality 60%, wider impact of research 25% and vitality of the research environment 15%. A pilot exercise is currently going on. Decisions on the configurations of panels and the methods for assessing impact have not been taken. It seems as if Britain is moving towards an informed peer review model with a component based on an effect indicator.

Third order indicator PRFSs currently in use

72. Table 9 gives an overview over countries currently using PRFSs anchored in third order indicators as the main principle. Notice that the Australian model is not fully implemented but under development.

Table 9: Third order indicator PRFSs currently in use

Country	PRFS	Indicators	Weighting	Data sources	Differentiating
Australia	Excellence in Research for Australia (ERA): Peer panels relying on discipline-appropriate indicators	Discipline-appropriate indicators in four categories[9]: -Research quality (ranked outlets, citation analysis, ERA peer review, peer-reviewed research income), -Research volume and activity (outputs, income), -Research application (research commercialization income), -Recognition (esteem measures).	Not yet decided how to link to funding.	Being developed.	The Australian model is highly differentiating across fields as indicators are developed discipline-appropriate.
Hong Kong	RAE inspired system	Assessment of quality of recent performance through assessment of active research staff in cost centres.	Not relevant.	Basic research products, primarily publications.	
Poland	Effectiveness indicator for research units	Units are assessed in five categories. Category 1 units have an effectiveness indicator that is more than 30% above the average of the homogenous unit, category 5 units have less than 70% of the average.	Complex system of weights of very many underlying scores.	Annual unit questionnaire on both research and practical applications of research.	Differentiating across 19 categories of homogenous units across three categories of homogenous fields: *1)* Humanities, social sciences and arts, *2)* Exact and engineering sciences, *3)* Life sciences.

73. Both the Australian and the Polish third indicator model is a pure informed peer review model. Peers are not required to read publications but will rely solely on discipline-appropriate indicators and information. The Australian indicators are planned to capture both research activity and intensity through measures of research income (input), PhD completions and publications (output), research quality through citation analysis (effect, impact) as well applied research and translation of outcomes.

74. The Polish information collected through questionnaires to research units include input information (*e.g.* finances), process and structure information (*e.g.* participation in international research projects and infrastructure) and output and effect information (*e.g.* publications, patents and copyrights).

75. Both models are characterized as a third indicator model because they are organized with peer panels as the focal point. But it may be discussed whether the room for expert opinion in the models is so restricted as to place peers in a primarily administrative role.

[9] Australian Government, Australian Research Council (2009: 7).

76. In addition to the above mentioned countries, Spain has a national third order indicator evaluation system called the sexenio because it is performed every 6 years (Rodriguez-Navarro, 2009). The Spanish system evaluates the research outputs of tenured professors establishing a salary bonus for each period positively assessed. As I understand it, this is not a funding systems allocating funds to tertiary institutions and thus not a PRFS in the OECD understanding of PRFSs.

Mixed indicator PRFSs

77. We also find countries not anchoring PRFSs in one of the three main principles but instead mixing principles. Italy and New Zealand are examples of this.

78. In Italy it was decided in 2009 to allocate 7% of block funding to the universities on a performance base. 2/3 of this concerned grants for research. Three indicators were used: *i)* Peer review ratings carried out in 2001-2003 and published in 2006, weighing 50%, *ii)* Ability to attract EU funding, weighing 30% and *iii)* Share of government competitive grants, weighing 20%.

79. New Zealand has had a PRFS since 2001. Three indicators have been used: *i)* Peer review inspired by RAE but assessing research performance of staff not departments as such, weighing 60%, *ii)* number of graduates, weighing 25% and *iii)* ability to attract external funds, weighing 15%. The peer review is periodically. It has been carried through in 2003 and 2006 and is planned to take place also in 2012. The two other indicators are measured yearly. The funding period is the calendar year.

Summarizing trends in the use of indicators in PRFSs

80. The analysis of how PRFSs deal with indicators has revealed the following development dynamics and trends.

81. First, comparing the rich world of indicators (section 2) and the analysis of PRFSs (this section) shows that PRFSs use first order indicators especially input and results indicators, as well as third order indicators. Second order indicators (JIF, H-index) are seldom directly used but they may be informally used in peer review processes and may thus influence third order indicators. The Flanders model is an exception as it takes JIF into account trying to correct for differences across disciplines.

82. Second, within the first order indicator types, input and results indicators are the ones mainly used. Process and structure indicators are used in Poland and have been suggested *e.g.* in Sweden but apart from this they are seldom directly used.

83. Third, within the results indicator types, output (publication counting) and effect (citation counting) indicators are the ones mainly used. Indicators related to outreach, commercialization and end-user esteem are seldom used. Indicators used in PRFSs overall are mainly what could be termed academic community indicators. Poland is using both academic community and societal indicators and other countries have been discussing the possibilities of including non-academic community indicators, but hitherto these have only to a limited extent been brought into use. There are probably several reasons to this. Clear non-academic community indicators are not easily developed and probably they are experienced less legitimate in the academic community.

84. Fourth, it seems that still more indicators are brought into use across time. Output indicators systematically counting publications are developed and effect indicators are increasingly brought into systems both as a stepping stone for informed peer review and as effect monitoring in the form of citation counting.

85. Fifth, even though still more indicators are brought into use across time, the number of indicators seem often to be reduced from the phases of talk about how to construct a PRFS to the phases of enacting the PRFS.

86. Sixth, third order indicator systems anchored in peer review are developed from modified peer review systems to informed peer review systems. This may strengthen systems making them more transparent and fair. However it seems to go along with reducing the numbers of peer panels probably with the consequence of reducing the peer coverage of research fields and maybe developing the peer review process into a more mechanical process.

87. Seventh, as the use of indicators changes also data sources and the ways of handling differentiation across fields change. This is summarized in Table 10.

Table10: Data sources and field differentiation

Model	First order indicator model monitoring publications	First order indicator model monitoring citations	Third order indicator model
Data sources	National databases (self reporting, validating)	International citation databases (buy in)	Made up by departments on request in each assessment round
Differentiating across fields	Handled by peers grouping journals and publishing houses in order to produce comparable publication points. Enacted both without field differentiation (*e.g.* in Norway) and with field differentiation (*e.g.* in Denmark).	Necessary as citation counts are not suited for several fields, *e.g.* most of the humanities, several subfields within social science and some also within the technical sciences.	Handled by peer panels translating their qualitative assessment into a rating

88. Finally, as overall country arguments for introducing PRFSs as such are very alike, related to maintaining and promoting excellence implicitly or explicitly related to competitiveness towards other countries, country arguments for choices at the more specific model level differ. At this level models are played off against each other often it seems with rather weak documentation. Arguments at this level are related to the costs of running systems, the degree of transparency in systems and the fairness levelled of against the wish for developing one system that fits all research fields.

89. At national levels the political pressure for introducing and maintaining PRFSs seems to initiate more micro level research political power struggles between different actors advocating for different models.

4. Consistency in performance measurement?

90. In the general literature on indicators and performance as well as in the specific one related to research institutions there are discussions of the characteristics of good indicators. The argument is that if indicators do not meet the criteria of good indicators they are less useful. Table 11 summarizes three proposals of such criteria.

Table 11: Characteristics of good indicators

Indicators in general: The CREAM test (Kusek & Rist, 2004)	Indicators in general (Mayne, 2010)	Research indicators (European Commission, 2010)
Good indicators are: -Clear (precise and unambiguous) -Relevant (appropriate to the subject at hand) -Economic (available at a reasonable cost) -Adequate (Provide a sufficient basis to assess performance) -Monitorable (Amenable to independent validation)	Good indicators are: -Relevant -Available (timeliness) -Understandable (clarity, transparency) -Reliable -Credible	Good research indicators are: -Fit-for purpose -Verifiable -Fair -Appropriate -Capable of facilitating comparisons across disciplines and institutions

91. As shown there are some common features but also some differences in the proposed criteria. First of all there are common features regarding what can be termed the methodological strength of indicators. Good indicators have to be relevant, reliable, credible and verifiable. Secondly, good indicators have to match the purpose of using them. They have to be clear, adequate, fit-for-purpose and as for research indicators especially capable of facilitating comparison. Thirdly, they have to, at least to some extent, be accepted and trusted, to be understandable and fair. Finally there are some more "technical" criteria. Good indicators have to be economic, monitorable and available at the right point in time.

92. The analyses above have shown that these criteria are challenging in relation to the development of PRFSs. Far from all indicators used are clear. In addition research institutions as well as disciplines and research fields are diverse. No single indicator is capable of capturing complexity. If complexity is to be captured adequately several indicators need to be used. But then complexity and costs increase.

93. Also fair comparison across diversity is challenging. If effect indicators are used, as for example in the Swedish PRFS, normalizing of data and differentiated weighing becomes necessary. But then systems become very complex, transparency is weakened, systems risk being less understandable for persons with limited "technical" skills and the incentives may be reduced. Using indicators in PRFSs is thus faced with dilemmas.

94. Peer review has been used to bridge diversity and translate qualitative assessments into ratings possible to use for allocating funding. Peer review is however both cost-intensive and not infallible. It is therefore interesting to explore whether this indicator is consistent with other indicators. Some studies have contributed to knowledge on this question.

95. As a follow-up of the 2001 RAE assessment an analysis of political science was carried through (Butler and McAllister, 2007). The analysis included not only citations received by articles in journals indexed by Web of Science. Rather it included citations to all publications submitted to the RAE. The analysis showed that the mean number of citations a work attracts significantly improves the RAE outcome for a department. The conclusion thus was that citations are an important indicator of research quality as judged by peer evaluation through the RAE. The analysis however also showed that the second important predictor of outcome for a department – slightly less than half as important as citations – was having a member on the RAE panel.

96. In 2007 the Higher Education Funding Council for England commissioned the Centre for Science and Technology Studies at Leiden University to carry through a study exploring technical issues in

developing the REF (Moed et al., 2008). As part of the study a more comprehensive analysis was made of the correlation between the 2001 RAE rating and a normalized citation analysis of the papers submitted to the RAE by departments in 8 subject groups covering clinical medicine, health sciences, subjects allied to health, biological sciences, physical sciences, engineering and computer science, mathematics as well as social sciences and humanities. Overall the analysis revealed that there seems to be a correlation between citation analysis and peer review as the normalized citation impact of departments increased with increasing rating. The analysis however also revealed exceptions. In engineering and computer science departments with RAE rating levels 2, 3a, 3b, 4 and 5 had similar normalized citation impacts whereas only the citation impact of departments with RAE rating level 5* substantially exceeded that of departments with other ratings. A similar pattern, although with higher impact levels, was found in clinical medicine.

97. Also the Italian evaluation has been follow-up by an analysis of the correlation between peer review scores and both article citation and journal impact factors (Franceschet and Costantini, 2009). The conclusions are in line with the ones mentioned above, the higher the peer assessment on a paper, the higher the number of citations that the paper and the publishing journal receive. However the strength of the correlation varies across disciplines and it depends also on the discipline internal coverage of the used bibliometric database. The higher the coverage of the discipline, the higher is the reliability of citation measures. But in addition there are examples of papers receiving positive peer judgments but very few if any citations as well as papers obtaining poor peer judgments but receiving significant numbers of citations. It is worth noticing that during the peer review process peers had very limited knowledge about article citations as citations due to time span were not yet received. They had however access to the impact factors of the journals. In this perspective it is not surprising that the analysis also revealed a correlation between peer review scores and journal impact factor.

Knowledge gaps

98. The lessons learned from the studies mentioned above indicate that there is still a need analytically to dig deeper into the correlation between peer review produced third order indicators and first and second order indicators. More generally there is a knowledge gap in relation to our understanding of how peer review processes in PRFSs are carried out. We have some knowledge about peer review processes in other types of evaluation systems, in grant peer review and in assessment of interdisciplinary research but we have very limited insight into peer review processes in PRFS contexts.[10]

99. Overall there is a knowledge gap on the development processes and the dynamics of PRFSs. This paper (and the block A and C papers by Diana Hicks and Linda Butler – see DSTI/STP/RIHR(2010)3 and DSTI/STP/RIHR(2010)6) show that PRFSs are rapidly spreading and developing across borders these years as well as further developed within borders in new system generations. It would be interesting to follow up this OECD initiative in the coming years in a thoroughly monitoring and systematically carried through comparative analyses of PRFSs. Such analyses should be carried out by experts at arm's length of giving advice on and being responsible for the development of PRFSs.

100. In the general literature on performance measurement through indicators there is an interesting discussion on the performance paradox defined as the weak correlation between performance indicators and performance itself (van Thiel and Leuw, 2002). The phenomenon of the performance paradox is reported to be caused by the tendency to performance indicators to run down over time. The deterioration process of indicators is described to be caused by four processes called positive learning (performance improves but indicators lose sensitivity to detect bad performance), perverse learning (performance is

[10] An interesting overview on the knowledge of peer review is available in Langfeldt, 2001, and a special theme on peer review of interdisciplinary research in Research Evaluation, volume 15, number 1, 2006.

reported to go up but this is due to manipulated assessments), selection (differences in performance is reduced due to the replacement of poor performers with better performers) and suppression (differences in performance is ignored).

101. It would be too lengthy in this context to go into details about the dynamics of these processes. The point solely is that there is a knowledge gap related to whether and how indicators in PRFSs run down over time. Do PRFS indicators cause positive learning in tertiary education institutions and national research systems or do they cause perverse learning, selection and/or suppression? And do dynamics differ across different types of PRFSs?

102. The performance paradox thinking sets a stage where studies of PRFSs analyze systems design, indicators used and their development, destiny or fortune. As mentioned, PRFSs may be seen as systems that constitute incentives to improve research performance. An important question is however whether and how these incentives influence the behaviour of academics whom have classically been reported as being motivated by more intrinsic values. Another approach to PRFS studies is to focus analyses on academic staff behaviour and the importance of the context of PRFSs to this. Important questions in this respect are: Do publication counting systems such as the Norwegian and the Danish ones increase publication performance (more and better publications) or do they result in researchers maximizing publication activity through recycling? Do citation counting systems such as the Swedish one increase research quality or do they advance citation circles?

103. In short we need more knowledge on how contexts and actor strategies shape PRFSs and on how PRFSs shape actor strategies and behaviours as well as on how these dynamics evolve across several generations of PRFSs.

REFERENCES

Adler, R.; J. Ewing & P. Taylor (2008): Citation Statistics. Joint Committee on Quantitative Assessment of Research, (http://www.mathunion.org/fileadmin/IMU/Report/CitationStatistics.pdf).

Australian Government, Autralian Research Council (2009): *ERA 2010 Submission Guidelines. Excellence in Research for Australia*. Canberra: Commonwealth of Australia.

Bouckaert, G. & J. Halligan (2008): *Managing Performance. International Comparisons.* London: Routledge.

Butler, L. and I. McAllister (2007): *Metrics of Peer Review? Evaluating the 2001 UK Research Assessments Exercise in Political Science*. Australian National University.

Carlsson, H. (2009): Allocation of Research Funds Using Bibliometric Indicators – Asset and Challenge to Swedish Higher Education Sector in *InfoTrend*, 64, 4 (82-88).

Cave, M.; S. Hanney & M . Kogang (1991): *The Use of Performance Indicators in Higher Education*. London: Higher Education Policy Series 3.

Debackere, K and W. Glänzel (2004): Using a bibliometric approach to support research policy making: The case of the Flemish BOF-key in *Scientometrics*, Vol. 59, No. 2 (253-276).

Dolan, C. (2007): *Feasibility Study: the Evaluation and Benchmarking of Humanities Research in Europe.* HERA.

Elzinga, A. (2007): *Evidence-based science policy and the systematic miscounting of performance in the humanities.* Paper given at workshop on evidence-based practice, University of Gothenburg, 19-20 May 2008 (http

European Commission (2010): *Assessing Europe's University-Based Research. Expert Group on Assessment of University-Based Research*. Brussels.

Franceschet, M. & A. Costantini (2009): *The first Italian research assessment exercise: a bibliometric perspective.* Preprint Submitted to Research Policy December 14

Frölich, N. (2008): *The Politics of steering by numbers. Debating performance-based funding I Europe.* Oslo: NIFU-STEP.

Geuna, A. & B. R. Martin (2003): University Research Evaluation and Funding: An International Comparison in *Minerva*, 41, 277-304.

Gläser, J. & G. Laudel (2007): Evaluation without Evaluators: The Impact of Funding Formulae on Australian University Research in R.Whitley & J. Gläser (eds.): *The Changing Governance of the Sciences*. Dordrecht: Springer.

Hansen, Hanne Foss (2005): Choosing Evaluation models. A discussion on Evaluation Design in *Evaluation*, vol. 11 (4), 447-462.

Hansen, H. F. (2009): *Research Evaluation: Methods, Practice and Experience*. Copenhagen: Danish Agency for Science, Technology and Innovation.

Hansen, H. F. & F. Borum (1999): The Construction and Standardization of Evaluation: The Case of the Danish University Sector in *Evaluation*, vol. 5, 3.

Hansen, H. F. & B. H. Jørgensen (1995): *Styring af forskning: Kan forskningsindikatorer anvendes?* Copenhagen: Samfundslitteratur.

Hemlin, S. (1991): *Quality in Science. Researchers' Conceptions and Judgements.* Gothenburg: University of Gothenburg.

Hicks, D. (2006): The Dangers of Partial Bibliometric Evaluation in the Social Sceinces in *Economia Politica*, XXIII, no. 2, 145-162.

Higher Education Funding Council (2007): (http://www.hefce.ac.uk/pubs/circlets/2007/c106_07/)

Higher Education Funding Council for England (2008): *Understanding institutional performance. Advice to the Secretary of State for Innovation, Universities and Skills.* Bristol: HEFCE.

Hicks, D. (2009): Evolving Regimes of Multi-university Research Evaluation in *Higher Education*, 57, 4, 393-404.

Kunnskabsdepartementet (2010): *Tilstandsrapport for UH-institusjoner 2010*. Oslo.

Kusek, J. Z. & R. C. Rist (2004): *Ten Steps to a Result-Based Monitoring and Evaluation System.* Washington, D. C.: The World Bank.

Langfeldt, L. (2001): *Decision-making in expert panels evaluating research. Constraints, processes and bias.* Oslo: The Faculty of Social Sciences.

Lehmann, S., B. E. Lautrup & A. Jackson (2006): Measures for measures in *Nature,* vol. 444, 21/28 December, 1003-1004.

Lehmann, S., B. E. Lautrup & A. Jackson (2009): Comment: Citation Statistics in *Statistical Science,* vol. 24, no. 1, 17-20.

Library House (2007): *Metrics for the Evaluation of Knowledge Transfer Activities at Universities.* Cambridge.

Leuwen, T. van (2008): Testing the validity of the Hirsh-index for research assessment purposes in *Research Evaluation*, 17 (2), 157-160.

Mayne, J. (2010): Results Management: Can Results Evidence Gain a Foothold in the Public Sector? In O. Rieper, F. L. Leuw & T. Ling (eds.): *The Evidence Book. Concepts, generation, and Use of Evidence.* New Brunswick: Transaction Publishers.

Moed, H. F.; M. S. Visser and R. S. Buter (2008): *Development of Bibliometric Indicators of Research Quality.* Centre for Science and Technology Studies (CWTS), Leiden University.

Nowotny, H. (2007): How many policy rooms are there? Evidence-based and other kinds of science policies in *Science, Technology and Human Values,* 32 (4), 479-490.

OECD (2008): *The role of expert review in the evaluation of science and technology: Issues and suggestions for advanced practices.* Paris.

Rodriguez-Navarro, A. (2009): Sound Research, Unimportant Discoveries: Research, Universities and Formal Evaluation of Research in Spain in *Journal of the American Society for Information Science and Technology*, 60 (9), 1845-1858.

Sandström, U. & E. Sandström (2009): The field factor: toward a metric for academic institutions in *Research Evaluation*, 18 (3) 243-250.

Seglen, P. O. (1994a): Causal Relationship between Article Citedness and Journal Impact in *Journal of American Society for Information Science*. 45 (1), 1-11.

Seglen, P. O. (1994b): *Siteringer og tidsskrift-impakt som kvalitetsmål for forskning.* Oslo: Det Norske Radium Hospital.

Seglen, P. O. (2009): Er tidsskrifts-renommé og artikkeltelling adekvate mål for vitenskapelig kvalitet og kvantitet i Ø. Østerud (red.): *Hvordan måle vitenskap?* Oslo: Novus Forlag (39-70).

Sivertsen, G. (2006): A bibliometric model for performance based budgeting og research institutions in *Book of Abstracts*. 9th International Science and Technology Indicators Conference, Leuven, Belgium, 7-9 September 2006, 133-135.

Sivertsen, G. (2009): *A performance indicator based on complete data for the scientific publication output at research institutions.* Oslo.

SOU 2007:81. *Resurser för kvalitet.* Stockholm: Utbildningsdepartementet.

Talbot, C. (2007): Performance Management in E. Ferlie; L. E. Lynn & Christpher Pollitt: *The Oxford Handbok of Public Management*. Oxford: Oxford University Press (491-517).

Van Thiel, S. and F. L. Leeuw (2002): The Performance Paradox in the Public Sector in *Public Performance & Management Review*, vol. 25, no. 3, 267-281.

Vedung, E.: *Public Policy and Program Evaluation.* New Brunswick: Transaction Publishers, 2007.

Vetenskapsrådet (2009): *Bibliometrisk indicator som underlag för medelsfördelning. Svar på uppdrag enligt regeringsbeslut* U2009/322/F (2009-01-29) till Vetenskapsrådet. 2009--5-27. Stockholm.

Wennerås, C. & A. Wold (1997): Nepotism and sexism in peer-review in *Nature*, vol. 387/22.

Whitley, R. & J. Gläser, eds. (2007): *The Changing Governance of the Sciences*. Dordrecht: Springer.