

Royal Netherlands Academy of Arts and Sciences

Quality assurance in scientific research

From SEP to CEP: Balancing fairness and simplicity

Academy Committee for Quality Assurance

AMSTERDAM, JUNE 2008

© 2008 Royal Netherlands Academy of Arts and Sciences

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photo-copying, recording or otherwise, without the prior written permission of the publisher.

Kloveniersburgwal 29, 1011 JV Amsterdam

P.O. Box 19121, 1000 GC Amsterdam, the Netherlands


T +31 20 551 0700

F +31 20 620 4941

E knaw@bureau.knaw.nl

www.knaw.nl

ISBN 978-90-6984-554-8

The paper in this publication meet the requirements of  iso-norm 9706 (1994) for permanence.

Contents

Summary 6

1. Is quality assurance too intensive? 7
2. Recent trends in the Netherlands and elsewhere 10
3. Research assessment: an analysis 13
4. Balancing fairness and simplicity 18
5. Recommendation: from SEP to CEP 20

Reference materials 23

Appendix 1 – Survey of assessment practices 25

Appendix 2 – Peer review and bibliometrics: advantages and disadvantages 27

Appendix 3 – Membership and task of the Committee for Quality Assurance 29

Summary

The science system has become much more complex and more demanding for researchers over the past decades. There has been an increasing contextual influence, from government, industry and society at large. This development has consequences for evaluation systems, not those regarding the individual researcher or groups and projects, but also concerning the system as a whole.

The Committee for Quality Assurance of the Royal Netherlands Academy of Arts and Sciences has issued this report after analysing the state of the art of the evaluation of scientific research. The Committee came to the conclusion that the current régime of evaluations in the Netherlands, but also in other countries, is perceived by many as too complex and too burdensome. Moreover, most systems lack the flexibility necessary to judge the quality of the variety of disciplines and their dynamic development in multi-, inter-, and transdisciplinary endeavors.

With regard to the Dutch system, the Committee recommends that there are two pivotal issues that need improvement.

- Simplification of procedures, both in requirements and in sheer number. There are too many, non standardised and partly overlapping evaluations. On top of that, there is a lack of mutual attunement.
- Fairness. The system should be able to do justice to the variety of disciplines and developments in research; and it should reflect the differences in research practices.

At the end of this report, the committee suggests a simplification of the current national evaluation system in the Netherlands, the so-called Standard Evaluation Protocol (SEP). The Committee proposes a concise list of ten evaluation categories that together should be appropriate to meet the two major concerns mentioned above. For each category, the Committee advises a minimum number of indicators and / or a maximum amount of text, in order not to overburden the research group or the evaluation committee.

The SEP will be renewed in 2009, and the Royal Netherlands Academy hopes that the suggestions made in this report will be taken into account when designing the new SEP.

1. Is quality assurance too intensive?

1.1 Introduction

Assessing performance is becoming increasingly important in many areas of government policy. Science and scholarship is no exception, either in the Netherlands or elsewhere. That applies both to the system as a whole, in particular the methods that are used to influence it, and to the individual higher education and research institutions. It is therefore not surprising that when one considers the issue of assessment and quality assurance one is soon confronted by the question of whether this is not too much of a good thing, and whether assessment is carried out as efficiently and effectively as it might be. Those are the core questions dealt with by the Committee for Quality Assurance of the Royal Netherlands Academy of Arts and Sciences (KNAW). Over the past year, the committee has analysed the way scientific and scholarly research is assessed, basing its work partly on a study of the relevant literature and partly on its own experience. This report presents the committee's findings. The committee has found that many people consider the current assessment system to be too complex and onerous – despite it being the specific intention of the Standard Evaluation Protocol (SEP) to reduce the burden – while at the same time assessments are not flexible enough, meaning that some specialist areas cannot be properly assessed.

1.2 The various dimensions of quality

The overall aim of quality assurance and assessment is to enable the research system to function effectively in the light of scientific and social objectives. Those objectives, broadly speaking, are for research to be of the highest scientific quality and the maximum social usefulness and relevance. One needs to remember, however, that the research system comprises a large number of actors whose needs are different and who are not all pursuing the same goals. Government, for example, generally requires data that will allow decisions to be made at national level on financing, priorities, and perhaps reallocation. Researchers themselves want to know how they are doing compared to the competition (including internationally). Local administrators are concerned about whether their institute is fulfilling its aims and whether some groups are achieving better results than others. External sponsors are interested in whether the goals for which they have given support are in fact being achieved. Assessment is therefore seldom a one dimensional matter: it almost always implies 'multi-target' or 'multi-criteria' assessment. Quality consequently involves a lot of different dimensions. Everything would be very simple if all those different dimensions were to ultimately coincide, but that is not very likely.

1.3 Complexity of quality assurance puts extra pressure on researchers

In the past few decades, the system of science and scholarship has become increasingly complex, thus placing much greater demands on researchers. This makes itself felt primarily through the need to access an ever-increasing number of external sources of funding. In addition to the Netherlands Organisation for Scientific Research (NWO) and industry, there are many other external sources whose importance has increased in recent decades, for example government ministries, the EU, and societal funds. Twenty years ago, the maximum proportion of funding for scientific/scholarly research from external sources was on average no more than 20 to 25%; that figure is now about 10% higher (A. Versleijen (ed.) 2007). In some research fields, the percentage is lower but there are some where it is more than 50% and even as high as 80%. This growing external influence also has consequences for assessment and quality assurance. That regards not only the assessment of the quality of individual researchers, research groups, or projects but also of the system as a whole. In addition, it involves not only intrinsic quality but also social relevance. External financiers therefore account for an increasing proportion of research financing and each of them has its own expectations as regards results. Quality assurance consequently takes the form of assessment of the mission and effectiveness of institutions and organisations, together with ‘bottleneck analysis’ (Arnold 2004; Merckx, Van den Besselaar 2008). The bottlenecks concerned here can be identified in various components of the system, for example the types of financing, programming, the organisation of research and careers, researcher training, and also in the assessment system itself. The system of science and scholarship is therefore an increasingly ‘assessment-intensive’ one. Assessment plays a major role in many respects, whether one is dealing with scheduling and programming, allocation of funds, selection of researchers and research groups, the organisation of research, research results, etc. In fact, assessment takes place a great deal – too much according to many researchers – and in many different ways, as well as according to a variety of methods. All this leads to increased pressure on researchers at the cost of their primary research tasks. On the other hand, the committee believes that since the introduction of national assessments (starting with the Conditional Financing system of the mid-1980s) the quality of research has increased, with assessments certainly playing a role. This primarily concerns quality in the sense of how Dutch research compares internationally, together with the transparency of performance.

1.4 Tendency towards simplification and avoidance of duplication

Unfortunately, there is no single magic method for assessing research and answering all questions, although there is a trend towards more systematic assessment at national level. In the Netherlands, the SEP system was introduced in 2003 but the UK has had its Research Assessment Exercise much longer. Australia has the Research Quality Framework, and France the Comité National d’Evaluation de Recherche (CNER).

The similarity between these systems is to be found in their desire to simplify research assessment (in the sense of reducing the administrative burden) and to make

it more effective (by in some way viewing the results from a national perspective).

In order to arrive at a properly substantiated judgment on assessment and assessment pressure, one needs to consider the following questions (as a minimum):

- Why is assessment necessary? What is its intention and what use is made of the results?
- The purpose of assessment. Is the idea to arrive at a judgment on the scientific/scholarly quality of the research concerned (after it has taken place), the promise for the future (before it has taken place), or its impact on society and its social relevance?
- The aggregation level. Are we assessing an individual, a group, a faculty, or an institute/institution? Or are we perhaps assessing an inter-university research school?
- The criteria. Are these relevant and suited to the purpose?
- The method. What measuring instruments and indicators are used? Are they valid and reliable?
- The consequences: What effects does the assessment have on the scientific/scholarly system as a whole or on specific disciplines? Does the assessment process lead to the desired improvements and are the results used in policy-making?

These questions, in various forms, were considered by the Quality Assurance Committee in the course of the past year.

We will propose in this report a systematic approach to the issue of assessment. We will first broaden our horizons by taking a brief look abroad and by considering the current situation in the Netherlands, where the Standard Evaluation Protocol (SEP) is used. We will then look at a number of methodological aspects of assessment and finally consider the most pressing problems and potential solutions.

2. Recent trends in the Netherlands and elsewhere

2.1 Introduction

Although the trend towards more intensive assessment (see above) is everywhere apparent, there are differences between countries as regards both their approach and the use they make of the results of assessment. Umbrella systems of assessment at the national level – such as that of the Dutch Standard Evaluation Protocol (SEP) – are also to be found in a number of other countries, for example the United Kingdom and Australia. In other countries, for example France, Germany, and the United States, there is a wider range of assessment systems. Many countries, however, do not have anything in the way of a highly developed assessment system; that is the case in a number of eastern European countries, for example, and also in Ireland, which has only recently begun intensive investment in scientific/scholarly research.

2.2 Assessment systems outside the Netherlands

In the United States, the dominant form of assessment is the ranking of universities, which has a major influence on an institution's ability to attract students, researchers, and research funding. There are also a number of federal agencies which assess what is done with the funding they provide, whether the specific research is a task of government, and whether the programmes concerned are achieving their scientific and social objectives (Michelson 2006). This type of assessment, and above all political decision-making on research budgets, produces major fluctuations in the research budgets of the various agencies and therefore in the various research topics.

In the United Kingdom, the Research Assessment Exercise (RAE) has taken a variety of forms since it was introduced in 1989. It is now in its fifth phase and a sixth phase is due to follow in 2008. According to Barker (2007), one important effect of the RAE has been the redistribution of research funding to a small number of universities that excel in the traditional academic disciplines, with this being at the expense of multidisciplinary and applied research. In Barker's view, the RAE has also produced a 'transfer market' for high-quality researchers, with a considerable number of sub-top research groups disappearing. In the course of discussion of the sixth phase of the RAE, in which 'research impact' would seem to be the main criterion, some critics have argued that the system has gone too far, mainly because regional innovation systems have deteriorated due to certain types of research and certain educational programmes disappearing. It is also difficult for new research groups that have not yet had any clear impact to comply with the requirements of the new RAE.¹

¹ During her presentation at the ERiC expert meeting at the Royal Netherlands Academy of Arts and Sciences in Amsterdam on 9 November 2007, Claire Donovan added that the new RAE defines 'research impact' as 'maximising the economic impact of research'. (Donovan is a researcher with the Research Evaluation and Policy Project (REPP) at the Australian National University.)

In Australia, a new Research Quality Framework has been introduced recently; its intention is to assess all publicly financed research (5 billion dollars a year) over the course of the next two years. This system assesses not only the scientific quality of research but also the ‘benefits to the wider community’. The RQF is extremely wide ranging and has specific protocols for a large number of fields intended to reflect what is customary in those fields in terms of output and interaction with society. The Australian government is also making efforts to improve access to research results for users. One way of doing this is to include not only scientists/scholars on assessment panels but also users.²

2.3 European trends

In Europe, there is also a trend at EU level towards a greater intensity of assessment. This is, for one thing, simply because there is an increasing amount of money to distribute (the Seventh Framework Programme for Research and Technological Development (FP7) has almost twice as much money available as FP6) but also because there is great political pressure to solve the ‘knowledge paradox’ – whether or not this actually exists – as is shown by the extremely ambitious Treaty of Lisbon (2000). At Lisbon, European leaders adopted the aim of making the EU ‘the most dynamic and competitive knowledge-based economy in the world’ by 2010. This will require a 3% rate of economic growth and 20 million new jobs. Science and technology will need to play a central role in all this. In this context, research must not only be of high quality but must also contribute more to the knowledge-based economy and to producing solutions to social problems such as those concerning the environment, migration, and water, Godin 2005). A number of assessment networks are – or will be – operating that regularly discuss European trends in this field; these include the RTD Evaluation Network (from FP6), the ESF Member Organisation Forum on Evaluation, and the ALLEA Working Group on Evaluation.³ The first of these has already been in existence for about ten years and is primarily intended to bring about the exchange of information; the others were set up more recently. The Academy is a member of all these networks; in ALLEA, it plays a leading role.

2.4 The Netherlands

Research assessment and quality assurance are generally well developed in the Netherlands and this country is often looked to as a positive example.⁴ After a critical review of the incumbent system in the 1990s⁵, the SEP – used everywhere since 2003 – was developed by the Association of Universities in the Netherlands (VSNU), the Netherlands Organisation for Scientific Research (NWO), and the Academy. The SEP has four assessment criteria: quality, productivity, relevance, and vitality.

² http://www.dest.gov.au/sectors/research_sector/policies_issues_reviews/key_issues/research_quality_framework/default.htm

³ All European Academies (ALLEA) is a federation of 53 academies of sciences and humanities in 40 European countries. The Working Group on Evaluation is an initiative by the Royal Netherlands Academy of Arts and Sciences and a number of other academies.

⁴ See Oostveen et al. 2007. We have made use of this report in a number of places.

⁵ Kwaliteit Verplicht, (2001), report by KNAW, VSNU and NWO.

It consequently aims at wide-ranging assessment – i.e. including social relevance and management aspects – of relevant research units, with institutions being free to choose the units that are to be subject to assessment. This has led to a whole range of units open to assessment, varying from whole disciplines to faculties (and parts of faculties), research institutes, and research schools. These units present themselves in the form of a self-assessment report in which they describe the features of their research, as well as any special local circumstances. The boards of the institutions appoint the assessment committees, which have an international makeup. More than the previous systems, the SEP and the self-assessment focus on bringing about a dialogue involving the assessor and the institution – as opposed to a detached assessment – and at improving the quality of the research concerned. The relevant literature distinguishes between a ‘jury’ model of assessment and a ‘coach’ model.⁶

Although broadly accepted and utilised, this system has once again become the object of criticism. Partly as a result of the critical review by the ‘Meta-Evaluation Committee’ *Meta-evaluatiecommissie* (MEC, set up by the Academy, the VSNU, and NWO to keep close track of implementation of the SEP), initial appreciation of the system has to some extent given way to dissatisfaction, one of the reasons being the fact that it is hard to unveil what institutions actually do with the results of assessments. University administrators would seem to consider it inexpedient for various reasons to be transparent about this matter.⁷ When questioned, members of the committee said that little was in fact done with the results of assessment in the current context, either because financial or employment law offered no scope for this, or because of a lack of decisiveness.

A second reason for concern is that subsidising bodies regularly circumvent the SEP and make use of protocols of their own that differ from those of the SEP. This leads to an accumulation of procedures that differ somewhat from one another and thus to additional heavy pressure on researchers, who can consequently devote less time to the primary process of research.

⁶ In these evaluation methods the focus is more on ‘learning effects’ than on passing judgments, see for example Spaapen et al. 2007, paragraph 1.3).

⁷ It should be noted, however, that the Academy does publish a dossier on each institute assessment on its website, including the actual assessment report and administrators’ comments.

3. Research assessment: an analysis

3.1 Introduction

We will examine a number of aspects of quality assurance that play a role in the current discussion of intensification. We will first look at what is actually assessed in the Netherlands so as to get an idea of the ‘accumulation’ of different assessments. After that, we will look at a number of methodological aspects of assessment and finally consider the differences in assessment between disciplines.

3.2 Accumulation of assessments

To a certain extent, an accumulation of assessments is inevitable because assessment and quality assurance are a routine component of research practice. Selecting staff, promoting employees, selecting project proposals, reviewing papers, etc. are all ‘standard’ forms of assessment. There are also types of assessment involving a more important policy-related component, for example accreditation (and re-accreditation) of research schools and assessments according to the SEP protocol. Where such assessments are concerned, the question often arises as to whether they are in fact necessary if the various functions of the research system operate effectively. Is it not the case, for example, that if projects are selected effectively by research sponsors (for example the Academy, the NWO, the Ministry of Education, Culture and Science, the Ministry of Economic Affairs, other government departments, the EU, charities, etc.), then it will become apparent ‘of its own accord’ just which research groups are in fact the good ones? At the same time, there has been a lot of criticism of project selection methods, for example that there is systematically preferential treatment of certain disciplines or of certain theoretical or methodological aspects of certain disciplines. Moreover, certain types of research (for example standardised experimental research) allow articles to be published much quicker than other types, thus boosting the likelihood of high marks when they are assessed.

The level of aggregation and the timeframe also play a role in assessments:

- an aggregation level ranging from *micro* level (individual researchers or research groups) to *macro* level (universities, advisory structure);
- assessment *ex ante* (research plans: project and programme proposals; research portfolio; foresight studies), or *ex post* (research results: projects and programmes, manuscripts, PhDs; external reviews);
- with either a *short-term* horizon (financing for a group) or a *long-term* horizon (decisions on investments in research infrastructures or the composition of the Dutch research portfolio that will determine the agenda for the next ten years).

As pointed out in the previous section, the intensity of assessment is increasing. This applies not only to keeping track annually of production and performance indicators (‘monitoring’) but also to self-assessment, midterm assessment, etc.. The accumula-

tion of assessments is also a problem because it has not been possible to standardise the way in which the various assessments provide information. Constantly needing to provide slightly different information for various different assessments is an extra burden on researchers. Given all this, it would appear that complaints regarding the accumulation of assessments are well founded. In Appendix 1, we list the various types of assessment, perhaps not exhaustively.

3.3 The most frequently used methods: peer review and bibliometrics

In practice, assessment is generally carried out by a committee made up of peers, experts, stakeholders, or a mix of these. The most frequent method of determining quality is peer review. In a number of fields (natural sciences and life sciences), bibliometrics is normally also used. Recently, however, the quantitative approach has been increasingly used in the humanities and social sciences too. All these methods focus primarily on scientific quality, but there is an increasing interest in methods that can establish the social quality and impact of research and thus support multi-dimensional assessment/self-assessment (RMW 2002; SWR/RGW 2005; Spaapen, Dijkstra & Wamelink 2007; Laredo & Mustar 2001; Van Ark 2007).⁸

Peer review has greater support in the world of science than bibliometric methods and a lot of research has been done on how peer review procedures operate. The advantages of peer review were summed up as follows at a recent OECD conference: 'It is a relatively quick, low cost, fast-to-apply, well known, widely accepted and versatile evaluation method that can be used to answer a variety of evaluation questions throughout the project performance cycle as well as in other applications.'⁹ However, peer review is also the object of criticism, for one thing because it is said to not always be disinterested, and to be to some extent conservative, so that innovative research is not always rewarded. It is also said to be difficult to properly assess multidisciplinary or interdisciplinary research by means of peer review. We summarise the main advantages and disadvantages in Appendix 1.

Bibliometric methods have the advantage of "objectifying" matters and being relatively transparent. The growth of the research system is also making it increasingly difficult to acquire a proper overview of more than a very small portion of the research system and even of one's own discipline. Some people therefore consider indicators based on scientometric studies to be preferable. In principle, such indicators can cover a large number of aspects, for example not just articles and citations but also numbers of PhDs, research funding awarded, etc. They are, however, confronted with legitimacy problems, in some fields more than in others, although not simply along a dividing line separating the natural and life sciences from the humanities

⁸ A symposium on methods of social quality assessment was organised by the Academy, NWO, VSNU, Quality Assurance Netherlands Universities (QANU), the Consultative Committee of Sector Councils for Research and Development (COS), and the Netherlands Association of Universities of Applied Sciences (HBO-raad). It became clear at the symposium that a large number of methods are being developed but that there is a lack of mutual communication and practical experience. International interest is also growing. On 9 November 2007, the Academy, working with the same organisations, organised an international expert meeting on the development of methods focusing on the social quality of science/scholarship.

⁹ Working Party on Innovation and Technology Policy, DSTI/STP/TIP(2007)13, 22 October 2007

and social sciences. Here too, there are a large number of problems (summarised in Appendix 2).

It is often suggested that the solution is to combine peer review and bibliometrics. That would not necessarily be a good solution because the results of both types of assessment do not always correlate well (Aksnes & Taxt 2004). In some fields, project selection hardly correlates at all with bibliometric indicators (Van den Besselaar & Leydesdorff 2007). Allocation of larger grants does lead to the selected research groups being more successful in terms of acquiring research funding than groups that have not been successful and have been rejected in the final round, but it leads to hardly any more publications in top journals or citations.

3.4 Some major problems in assessment

Despite the problems referred to above (and dealt with in Appendix 1), peer review is for the moment an effective and widely accepted method of assessment. The committee does not wish to deny that. At the same time, however, we wish to be open-minded regarding the potential weaknesses of peer review, and the same applies to bibliometrics. Every type of assessment also needs to take account of a number of general problems that arise when one attempts to determine the quality and impact of scientific/scholarly research. From an abstract point of view, most assessment criteria agree at the various different aggregation levels: quality, productivity, dynamism, numbers of PhDs, ability to attract researchers, students, and research funding, social benefits, etc. Emphases vary, however, and it is also the case that the lower the aggregation level, the more difficult it is to validly operationalise the criteria and ensure that they are reliable. The main problems include:

- Experts in many disciplines often believe that they know intuitively whether an individual or an article is of high quality, whereas it may be difficult to establish objectively whether that is in fact the case because there is no consensus regarding the indicators.
- The pressure to publish and the role of publication figures in assessment (the ‘publish or perish’ phenomenon) would appear to have gradually gone too far. Some people would say that this leads to overproduction of scientific/scholarly periodicals and separate articles; to a lack of understanding of implicit assumptions in research; to the impossibility of breaking through pointless traditions; and to the conviction that ‘theoretical’ is the same thing as ‘vague’. Additionally, the emphasis on publication is at the expense of other tasks that researchers have, namely teaching and providing services to society (including, for example, important historical collections or medical tasks).¹⁰ It should be noted that some countries now make use of ‘esteem indicators’ to assess a researcher’s other activities, for example prizes won or membership of important committees, editorial boards, etc. (Barker 2007). Such ‘indicators of esteem’ are also frequently used in the Netherlands as one of the criteria for quality but these criteria are as yet not really very specific.

¹⁰ Moed (2005) has determined that scientists have indeed been more productive in recent decades but mainly because of the increase in the phenomenon of co-publication. Productivity per full-time-equivalent has not risen, however.

- Attribution: how can one indicate in a worthwhile manner what the contribution is of the unit that is assessed compared to others with which there has been essential collaboration (Glaeser et al. 2004). Studies have shown, for example, that papers with international co-authors have a higher profile – i.e. are cited more frequently – but it is at the same time less clear just what citations actually measure (for example the quality of the research or that of the network).
- Determining the social benefits of research is difficult at micro-level because innovations are often based on a large quantity of scientific results and the input of other expertise, while the time that intervenes between the actual research and its application is both uncertain and variable. That is a greater problem for assessment at micro-level than at system level.¹¹
- Better data are available regarding the system as a whole than at certain lower levels of aggregation. At individual level, for example, the quality of the data is often only mediocre (10% of the ISI records include errors such as wrongly spelled names of persons and institutes; ensuring consistency in the names of individuals and institutions is a major problem). Competing databases (Google, some major American universities) sometimes produce strikingly different results, certainly at the level of individuals.
- Another trend that we wish to mention (without going into it in detail) is the involvement of stakeholders in the assessment process. A great deal has been written in the literature on innovation regarding ‘early user involvement’, i.e. the need to involve users at an early stage in developments in scientific research. More generally, certainly where research is used to tackle social problems, there is growing interest in the role of stakeholders in research and in assessing research.¹²

3.5 Differences between disciplines

There are major differences between disciplines regarding publication and communication practices and also regarding interaction with actual practice within the community. There is consequently also a reasonably broad consensus that those differences should be allowed to produce differences in assessment methods. Differentiation is sometimes even desirable within a particular discipline. Chemistry researchers publish differently and via other media than chemical technologists. The committee has considered a whole range of relevant differences between – and within – disciplines. In the case of the most important of those disciplines (summarised below), studies need to be carried out to determine the potential consequences for the structure of the assessment system. It should be noted that a number of disciplines are already taking action in this regard, for example medicine, social sciences, humanities, and technology.

¹¹ There have been many new developments recently, both in the Netherlands and elsewhere, regarding the criteria and indicators for social relevance and impact. Information can be found at www.ERIC-project.nl

¹² Guba & Lincoln distinguish four periods in the development of evaluation, referring to them as ‘generations’. They place the involvement of stakeholders in what they refer to as the fourth generation of evaluation (1989). In the three preceding generations, the focus gradually shifted from measuring individual performance to describing and assessing the individual’s environment (the mission of the institute) and towards the broader question of whether good work has been done. In the view of Guba and Lincoln, stakeholders with an interest in the research should be actively involved in assessing it.

- Disciplines differ regarding the extent to which they involve group activity. Some are highly individualistic, while others make use of teams both large and small. That this has consequences for assessment is shown, for example, by the discussion regarding the SEP. That protocol focuses primarily on the level of the group, making it less suitable in its current form for more individualistic disciplines.
- Disciplines also differ with respect to the value of criteria and indicators. In a discipline featuring a great deal of international competition, publications in high-impact journals are decisive but in a discipline with a technical or local focus other criteria may apply. In some disciplines, it is books that are the dominant method for presenting research, while in others it is technical artefacts.¹³
- An increasing amount of research is multidisciplinary, interdisciplinary, or transdisciplinary, and consequently difficult to assess by means of traditional indicators. It is also no simple matter to put together effective committees for such fields. One consequence of this is that it is less easy to specify what constitutes the ‘top’ in such a field.

All these differences should have an effect on how assessment procedures are structured. The question, however, is whether it is in fact efficient to have a separate set of procedures for each field.¹⁴ There is constant pressure from government and other administrative echelons for all research to adopt as far as possible the same – and preferably the simplest – yardstick. In the view of government, the SEP, for example, provides insufficient opportunity for this and there is a desire for changes to be made to the new SEP so as to make possible such things as countrywide comparison of different fields. This disparity is due to the fact that different actors may have different objectives as regards assessment. From the point of view of allocating funds, an assessment involves different requirements – and therefore different criteria and indicators – to when one is concerned with monitoring and improving quality. One of the major challenges in the near future will be to try to reconcile these two basic approaches within a single system. The present SEP runs until 2009. Consultations on a new version commenced in late 2007, with the present analysis of how the SEP works naturally playing a major role. One important issue is the extent to which the results of all these different assessments give policymakers and politicians a sufficient understanding of the quality and effectiveness of the research system. One possible solution to the disparity – each field with its own procedure or a single procedure for all fields – would be to make use of a ‘basic SEP’, with a restricted number of questions being added for each discipline or research question. Such a ‘basic SEP’ would need to be significantly simpler than the current protocol, in line with the desire of researchers to reduce the burden of assessment (see the following section).

¹³ In astronautics or architecture, for example, technical artefacts are far more important than published articles. Such differences can also be found within disciplines. In psychology, for example, a more qualitative movement focuses on books as the publication medium, whereas another school, that of cognitive psychology, focuses on periodicals.

¹⁴ It should be noted that this is precisely what is done in Australia by means of the Research Quality Framework.

4 Balancing fairness and simplicity

4.1 Introduction

Efforts have been made in recent years, both in the Netherlands and elsewhere, to find a solution to the above problems, in particular those of the burden of assessment and the application of criteria and indicators in the different disciplines. The intention of the SEP was to make other types of assessment superfluous – which it has not succeeded in doing – and to provide a certain flexibility for the units and disciplines that are being assessed so that they could take their own research practices into account during assessment. The second of these two objectives has not been very successfully achieved, but progress has been made. This final section of our report makes proposals for improving the current system, in particular as regards the above two issues.

4.2 Towards a fairer and simpler system

For the present, the Quality Assurance Committee will take two of the theme areas dealt with which it believes require systematic attention and will propose a solution.

- *Simplicity.* Assessments are too complex and not sufficiently standardised. The various assessments are also not coordinated sufficiently. This is sometimes referred to as the ‘accumulation’ of assessments, a problem that is made worse by the fact that those being assessed are required to provide information that is different in each case (even if only slightly so). This was the most frequently mentioned problem within the committee, and presumably elsewhere.¹⁵ It was in fact already a major concern for the Minister of Education, Culture and Science when the SEP was introduced. The specific purpose of the SEP was to put an end to the multiplicity and diversity of assessments. The protocol explicitly states that it must be possible to use the data from an assessment for three years in other cases. In practice, however, this seems not to be effective. The committee considers that research, researchers, and research groups are still assessed very frequently, and that the value of the various assessments in relation to one another is at the very least unclear.
- *Fairness.* Different disciplines and specialist fields differ considerably in their research practice and consequently assessment practice. Those differences need to be expressed in some way or other in the way assessment takes place. The SEP makes that possible to a certain extent, but its main focus is on a more traditional assessment of research, in which certain criteria and indicators are more important than others (for example an emphasis on citations). Areas that focus not only on scholarly practice but also emphatically on societal practices are consequently at a disadvantage. Furthermore, the main unit assessed by the

¹⁵ This appeared also in a recent (not yet published) survey of members of ALLEA, the European organisation of academies.

SEP is the research institute and the research group. In some fields, however, it is not the group that needs to be assessed but the individual.

Possible solution. The committee recommends simplifying the current system where possible, and make it fairer with respect to disciplinary variation. The challenge is in finding the right balance between the two.

The SEP should basically be the only system applying, but it should have sufficient flexibility to serve a variety of objectives. That flexibility can be achieved by deciding on a core SEP, a Core Evaluation Protocol : CEP. Such a core protocol should have sufficient flexibility to accommodate other relevant tasks of researchers, such as the training of PhD students at research schools, or the focus on technological, socio-economic and cultural aspects. Funds and other bodies providing funding would then be required to coordinate their assessments with this system as far as possible. The number of criteria in the present SEP could be reduced from four to three – quality, productivity, and relevance – and a more precise specification should be established for each of those three criteria, indicating (in the form of a limited number of indicators or indications) what they actually mean. In the case of relevance, the explicit indicator ‘earning capacity’ could be added, both regarding material and staff contribution. The fourth SEP criterion, ‘vitality and feasibility’, is certainly important but is difficult to specify for an external review committee conducting a short term site visit. It in fact concerns the question of whether the research group is properly managed and has an effective and vital structure (for example a good balance between inflow and outflow). This factor can be more effectively assessed during regular discussions within the institution concerned between administrators and research coordinators.

The SEP therefore needs to be made more concrete and simpler by having a core SEP with three rather than four criteria and compact reporting as regards the self-assessment (for example a maximum number of words per component). Lessons can be learned from assessment practices elsewhere. One example might be the method of assessment at Radboud University Nijmegen: 40% based on earn-back capacity and 60% based on a points system for articles, with articles in the top (S)SCI category earning more points than those in other periodicals. Such systems have also been developed elsewhere, for example by the Research School for Resource Studies for Development (CERES) and the Discipline Consultative Body for Law.

5. Recommendation: from SEP to CEP

In the view of the Committee, more intensive assessment over the past three decades was initially effective. The quality of research has improved in three respects: in general, the Netherlands scores well internationally, low-quality research has been weeded out, and the assessment system has become more transparent. The “Meta-Evaluation Committee” [*Meta-evaluatiecommissie*] considered the present national system to be reasonable to good; however, it seems for the moment not to be reducing the burden of assessment, despite that being its intended objective. In part this is because different financiers set different requirements; it also has to do with the fact that assessment organisations cannot always avoid overlap (or do not wish to do so). As we have seen, this difficulty could be largely resolved by applying a core protocol with sufficient flexibility to do justice to the differences between various disciplines. Harmonisation should also be encouraged on the side of the financiers. The Committee considers that there are two main questions that need to be tackled on the road to a new ‘core SEP’ (i.e. a ‘CEP’):

- A. What does such a CEP need to comprise (as a minimum)?
- B. Are addenda necessary in addition to a CEP, for example as regards social or technological relevance?

A. Core Evaluation Protocol: CEP

The Core Evaluation Protocol (CEP) consists of an amended text of the current SEP, aimed at simplification and fairness for all disciplines. The main changes that we refer to here concern three (rather than four) main criteria: quality, productivity, and relevance. In addition, there will be a much simplified self-assessment, comprising only the information that is absolutely necessary. The thick packages of documentation that are sometimes submitted to assessment committees contain a great deal of information that is never actually read. Where publications are concerned, it is much more interesting, for example, for a review committee to see a number of core publications rather than a complete list of publications for the last five years. A maximum length should also be set for the various components of the self-assessment.

The box below sets out what the Committee believes the self-assessment report should include:

SECTIONS OF THE SELF-ASSESSMENT REPORT

1. Objective of the research [maximum: half a page]
2. Composition of the group, based on two indications: total number of employees in each job category (including external PhD candidates) and overview of the various sources of financing (internal or external). No further details (if necessary, these can be found in annual reports etc.) [maximum: 1 page]
3. Research environment and embedding, national and international positioning, number of guest researchers (with or without their own funds) [maximum: 1 page]
4. Quality:
 - a. 3-5 key publications per group/subgroup
 - b. 3-5 most significant results/highlights relevant to the discipline, per group/subgroup
 - c. number of articles in top 10% of publications relevant to the discipline; ditto for top 25%
 - d. 3-5 most important books or chapters of books, insofar as relevant
5. Output:
 - a. number of articles in refereed periodicals
 - b. number of books, chapters of books
 - c. number of completed PhDs and number of PhDs 'on the way'
6. Earning capacity for competitive funds, national and international
7. Academic reputation for each research coordinator (prizes, invitations to address major conferences, conference organisation activities, editorships, membership of academies [maximum: 1 page])
8. Valorisation in the broad sense: socio-cultural relevance and/or technical or economic impact [maximum: 1 page] (see also B. below)
9. Feasibility of the proposal or programme, available infrastructure and methodology [maximum: 1 page]
10. Vision for the future, including opportunities and threats [maximum: 1 page]

B. Addenda?

After lengthy discussion, the Committee's answer to the question of whether addenda should be used to add specific information to the CEP is that they should not. Addenda might be used for such things as social relevance or technological impact. The Committee considers this information to be highly relevant, but also that the idea of a succinct self-assessment is too important to sacrifice to extra addenda. The Committee also considers that this kind of extra information can perfectly well be provided as part of the above list of criteria, for example points 1, 3, 8, and 10.

The Committee would like to close by making the following remarks about the broader relevance of research. In order for assessment to be fairer, it should also consider researchers' activities other than those focusing purely on the scientific/scholarly community. Much research today takes place in the context of social issues. Public-private partnerships are on the rise, resulting in increased pressure to "valorise", not merely in an economic sense but also in the sense of contributing to social, cultural, political, and welfare processes. This point should preferably be clarified by means of indicators/indications that are supported by the discipline. In the case of technical disciplines, one can consider such things as patents or collaboration with industry; in medical/biomedical research: clinical applications or protocols; in the humanities; exhibitions; in the social sciences: contributions to educational innovation. These are just examples. As has already been pointed out, KNAW, NWO, and VSNU have a joint project aimed at developing indicators for valorisation in various disciplines (www.eric-project.nl). The project is also intended to promote consensus within disciplines as regards assessment of the social quality of science and scholarship. The website also provides information about what is now a broad range of valorisation methods that are being utilised and/or tested both nationally and internationally.

Reference materials

- Adviesraad voor het Wetenschaps en Technologiebeleid, *Alfa en gamma stralen*. AWT: The Hague, 2007. [Advisory Council for Science and Technology Policy, *Alpha and Gamma Radiate*, AWT, The Hague, 2007]
- Arnold, Erik, Evaluating research and innovation policy: a systems world needs systems evaluations. *Research Evaluation* 13 (2004) 3-17.
- Aksnes, Dag W., and Randi Elisabeth Taxt, Peer review and bibliometric indicators: a comparative study at a Norwegian university. *Research Evaluation* 13 (2004) 33-41.
- Barker, Katherine, The UK Research Assessment Exercise: the evolution of a national research evaluation system, *Research Evaluation* 16 (1) March 2007 3-12.
- Commissie Dynamisering, *Investeren in dynamiek*, eindrapport, deel 1, April 2006. [Committee Dynamisering, *Investing in dynamism*, final report, part 1, The Hague, April 2006]
- Dietz, Ton, Het CERES-systeem van prestatiemeting”, [The CERES-system of assessment of research performance] lecture by director of the Interuniversity Research School for Resource Studies for Development (CERES), on 19 March 2007, Maagdenhuis Amsterdam, *Folia* 25 13.
- Disciplineoverleg Orgaan Rechtsgeleerdheid (DRG), Naar prestatie-indicatoren voor rechtswetenschappelijk onderzoek, March 2007. [Disciplinary Committee for Law, *Towards Indicators for Research Performance in Law*]
- ESF conference report, *Peer review – its present and future state*, Prague, 12-13 October 2006, released in Brussels April 2007
- Glaeser, Jochen, et al., Intraorganisational evaluation: are there “least evaluable units”? *Research Evaluation* 13 (2004) 19-32.
- Godin, Benoit, *Measurement of Science and Technology: 1920 to the Present*, London: Routledge, 2005.
- Gubba, E.G. and Y.S. Lincoln, *Fourth Generation Evaluation*, Newbury Park, CA: Sage Publications, 1989.
- KNAW, VSNU, NWO, *Kwaliteit verplicht. Naar een nieuw stelsel van kwaliteitszorg voor het wetenschappelijk onderzoek*, rapport Commissie Kwaliteitszorg, Amsterdam, 2001 [Quality obliges. Towards a new system for quality assurance of scientific research, report Committee Quality Assurance]
- Langfeldt, Liv. Expert panels evaluating research: decision-making and sources of bias. *Research Evaluation* 13 (2004) 51-62.
- Laredo, Ph. & P. Mustar, Laboratory activity profiles: An exploratory approach, *Scientometrics* 47/3, 515-539 (2000).
- Merkx, Femke, et al., *Evaluation of Research in Context; a quick scan of an emerging field*. The Hague: Rathenau Institute / ERiC, 2007.

- Merkx, Femke, and Peter van den Besselaar, Positioning Indicators for cross-disciplinary challenges: the Dutch coastal defense research case, *Research Evaluation* 20 (2008) (forthcoming).
- Meta Evaluatie Commissie, [Meta Evaluation Committee] *Trust, but verify*. Amsterdam, 2007.
- Michelson, Approaches to research and development performance assessment in the US: an analysis of recent evaluation trends. *Science and Public Policy* 33 (2006) 546-560.
- Moed, H.F., *Citation Analysis in Research Evaluation*, Springer, 2005.
- Oostveen, Anne-Marie et al., *Research evaluation – an overview*. Science System Assessment Rapport 0707. The Hague: Rathenau Institute (in preparation).
- Raad voor de Medische Wetenschappen, *The societal impact of applied health research*. Amsterdam: KNAW, 2002. [Council for the Medical Sciences]
- Raad voor de Medische Wetenschappen, *Gezondheidsonderzoek: het investeren waard*, KNAW, 2007. [Council for the Medical Sciences, Health Research : Worth the Investment]
- Spaapen, Jack, Huub Dijstelbloem and Frank Wamelink, *Evaluating research in context. A method for comprehensive assessment*, 2nd edition, The Hague: COS 2007.
- Sociaal Wetenschappelijke Raad (SWR) and Raad voor de Geesteswetenschappen (RGW), *Judging research on its merits*, Amsterdam: KNAW, 2005. [Council for the Social Sciences (SWR) and Council for the Humanities (RGW)]
- Van den Besselaar P., & L. Leydesdorff, *Research budget allocation and bibliometric indicators – an assessment*. Science System Rapport 0707. The Hague: Rathenau Institute (forthcoming).
- Versleijen, A. (red.), Dertig jaar publieke onderzoeksfinanciering in Nederland (1975-2005); historische trends, actuele discussies. [Thirty years of public research finance in the Netherlands (1975-2005): historic trends, current discussions], Science System Rapport 0703. The Hague: Rathenau Institute, 2007.
- Wenneras & Wold, Nepotism and sexism in peer review, *Nature* 387 (1997) 341-343.
- Working Party on Innovation and Technology Policy, Peer review: its uses, demands and issues, OECD workshop on rethinking evaluation in science and technology, 29-30 October 2007, Paris, DSTI/STP/TIP(2007)13.

Appendix I Survey of assessment practices

This survey does not attempt to be complete but it does give a good idea of the range of assessment practices with which researchers are confronted. It shows how complex those practices are, or at least how complex they have become. The crucial question at every point when assessment takes place is how much effort it costs to provide the information requested and what the actual benefits are of that effort.

In general, one can distinguish three different levels of assessment:

- A. Assessment focusing on individual researchers:
 - 1. During selection procedures, by the future employer.
 - 2. During (annual) performance appraisal interviews, by the researcher's superior.
 - 3. During appraisals, ditto.
 - 4. When applications are made for projects, by peers and research councils, by potential commissioners, including internationally.
 - 5. When project results are reviewed, by financiers.
 - 6. When articles are submitted, by peers and editors.
 - 7. When a PhD is completed, by peers.

- B. Assessment focusing on groups of researchers (institutes, programmes):
 - 8. Self-evaluation
 - 9. Mid-term assessment
 - 10. Annual monitoring, METIS (difference between faculty-university-discipline)
 - 11. SEP: research institute, faculty
 - 12. ECOS: accreditation/reaccreditation of research school

C. Assessment focusing on national/supranational scientific/scholarly system

Such assessment increasingly takes place in an international (European) context, in the form of occasional studies, or structurally, for example via the European Commission, the OECD, or nationally via the Ministry of Education, Culture and Science, or via the Rathenau Institute (Science Systems Assessment).

In all assessments, and at all the various aggregation levels, decisions need to be taken regarding the primary objectives, criteria, indicators, and methods. At each level, there are questions and problems specific to that level which influence the way assessment is organised. They are:

System level

Questions/problems include (this is naturally not an exhaustive list):

- Is project selection effective? Does the money reach the best researchers?
- Is the career system effectively organised so that promising young researchers are given full scope to develop and explore new avenues of research?
- Is the organisation of research at faculty research institutes and national research schools with programmes functional in all disciplines?
- Is the system of academic ranks sufficiently challenging? Or do we need to abandon the idea that “university lecturer” and “senior university lecturer” can be final positions? Is the hierarchy of the job structure actually functional? What is the role of the *Principal Investigator* within that structure?
- Can research sponsors articulate their demands effectively?

Institute level

Institutes often do more than just scientific research. They teach, they work on research collections, they provide services to science and society. An assessment considers all these aspects and also the way the institute is run. This means that one needs to decide whether all the various aspects should be assessed together – which will affect the method and, for example, the membership of the external committee – or whether separate assessments should be organised for the various different aspects (for example SEP for the research and ECOS for the PhD programmes) – which may increase the feeling that there is an accumulation of assessments.

Group level

In the case of research groups, the question is what exactly should be assessed – the group as a whole, or the individual research coordinators? Is a good research group a collection of top-flight researchers, or a group with a good coordinator and an effective and productive programme? What does this say about the form and objective of self-assessment by groups? Is it useful for assessment at this level to be external or is it specifically the management of the institute that should be responsible?

Individual level

Where individuals are concerned, there are usually three points when assessment takes place: when they apply for a position, during regular internal procedures (performance appraisal interviews etc.), and when external applications are made. In the latter case, the accent is generally on past performance, which says a lot about the quality of the researcher and indicates what can be expected of him or her in future. This also plays a role in the first two cases, but there other aspects are also considered, for example leadership qualities, the ability to work with others, innovativeness, and effectiveness at acquiring funds and recruiting good co-workers.

Appendix 2 Peer review and bibliometrics: advantages and disadvantages

Peer review

Peer review is one of the most common methods of assessing the quality of scientific/scholarly work. It is used at all the various levels of assessment (see Appendix 1). The method has many advantages: it is relatively fast and easy to use, inexpensive, widely accepted, and versatile, i.e. it can deal with a large variety of questions. There are disadvantages, however, partly because the peers involved are of course human – allowing subjective factors to perhaps gain the upper hand – and partly because so many demands are made on a restricted number of people that scrupulousness may be at risk (and “referee fatigue” may occur). It is also increasingly questionable whether peers can in fact deal with the questions that are currently posed in assessments, questions with a much broader impact than those raised in traditional discipline-oriented assessments.

Because it is such an important mechanism, the way peer review operates has always been the object of a great deal of research. A number of important conferences have been organised in the last couple of years by the OECD and the ESF with a view to subjecting peer review to broad analysis and seeking solutions. It was noted that the peer review process is becoming increasingly international, a trend that brings with it new problems. By way of illustration, these are a few of the questions that arise in the course of discussion of peer review.

- The enormous pressure to publish a great deal means that reviewers are reading less and less of the output. It is becoming increasingly normal to read only a few ‘key publications’.
- The ‘halo effect’: if something already has a good reputation, the likelihood of a good review increases.
- Bias due to disciplinary differences (Van den Besselaar & Leydesdorff 2007), or paradigm differences.
- Multidisciplinary, interdisciplinary, and transdisciplinary research is difficult to assess by means of traditional peer review: Disciplines discipline! ‘Peer’ review that attempts to deal with this criticism is sometimes known as ‘expert’ review. This is not always without its problems, however. When the members of a committee have differing competencies, or in any case differing disciplinary backgrounds, each of them will often be a ‘peer’ only as regards part of the subject of the review. This often leads to tacit or vague compromises (Langfeld 2004). This is also a danger affecting foresight studies of a wide area.
- Peers may have a particular interest. More specifically, studies have shown that there is ‘Nepotism and sexism in peer review’ (Wenneras & Wold 1997). One study found that women had to publish significantly more top publications to

achieve the same assessment. A good relationship with committee members also helped a great deal.

- Low reliability of the ratings. This also occurs when papers and proposals are assessed, and definitely not just in the humanities and social sciences. (Cicchetti 1991 and Rothwell & Martyn 2000 on the life sciences, for example).

Bibliometrics

In the twentieth century, criticism of peer review led to more objective methods being developed, for example bibliometrics and scientometrics. There was growing interest in these methods, particularly in the area of science policy, because they appeared to offer a relatively simple way of arriving at decisions on the allocation of research funding. Here too, it quickly turned out that bibliometric methods have not only advantages but also major disadvantages.

- The smaller the unit to be assessed, the less valid the methods are (Glaeser et al., 2004).
- The quality of the indicators is unclear: what measures what? This has been the object of a great deal of discussion and there is a need for more effective studies to determine the proper indicators that should be applied. One example is the increasingly popular “H index”, which has also led to much discussion.
- Bibliometrics is blind to new developments because they have not yet found their way into the established ISI-indexed journals: one needs to look back over at least a number of years in order to produce a successful citation analysis – past performance says little about how up to date research actually is.
- Quantitative criteria invite abnormal effects: if one requests more articles, one gets more articles; if one asks for more patents, one gets more patents; low-indexed journals are omitted from assessments. This then results in strategic behaviour rather than quality being measured. Using a number of indicators side by side makes all this even more difficult.
- Although the ISI database is the most frequently used when deriving bibliometric indicators, other databases are being developed, generally focussing on specific disciplines, which can lead to significantly different results, for example because they (also) use other media or include other output.

Appendix 3 Membership and task of the Committee for Quality Assurance

Task

The Committee for Quality Assurance of the Royal Netherlands Academy of Arts and Sciences (KNAW) was set up in 2006 to advise the Board of the Academy – both on request and at its own initiative – on the broad field of quality assurance in research. The Committee is chaired by Prof. Peter van der Vliet and has the following members:

Membership

Prof. Peter van der Vliet, chair, Emeritus Professor of Physiological Chemistry, Utrecht University (Chair)

Prof. Peter van den Besselaar, head of Science Systems Assessment at the Rathenau Institute, Professor of Social Sciences Informatics, University of Amsterdam

Prof. Trudy Dehue, Professor of Theory and History of Psychology, University of Groningen

Prof. Els Goulmy, Professor of Transplantation Biology (Histocompatibility Antigens), Leiden University

Prof. Bas ter Haar Romeny, Professor of Old Testament and Eastern Christian Traditions, Leiden University

Prof. Richard Grol, Professor of Quality of Health Care, Radboud University Nijmegen and Maastricht University

Prof. Ad Lagendijk, Distinguished University Professor, University of Amsterdam, researcher at FOM Institute for Atomic and Molecular Physics (AMOLF)

Prof. Emmo Meijer, Board of Management, Unilever R&D and Professor of Macromolecular and Organic Chemistry, Eindhoven University of Technology

Prof. Theo Mulder, Director of Research, Royal Netherlands Academy of Arts and Sciences

Prof. Frits van Oostrom, Distinguished University Professor, University of Utrecht, former president of the Royal Netherlands Academy of Arts and Sciences

Dr Jack Spaapen, Co-ordinator Quality Assurance and Research Evaluation, Royal Netherlands Academy of Arts and Sciences

Prof. Wiecher Zwanenburg, Emeritus Professor of French Language and Literature, Utrecht University

Jacco van den Heuvel, policy officer, Royal Netherlands Academy of Arts and Sciences, Secretary to Quality Assurance Committee

