Facts&Figures

WWW.RATHENAU.NL • 2013

Rathenau Instituut

Twenty years of research evaluation

Leonie van Drooge, Stefan de Jong,[1] Marike Faber and Don Westerheijden[2]

The Netherlands has a long tradition of quality assurance in academic research. This edition of Facts & Figures looks at the development of the quality assurance system and at the results of research evaluation.

Contents

1. Introduction

What has twenty years of quality assurance in academic research achieved? This edition of Facts & Figures considers this question. For the first time since 1993, it provides a review both of the development of the system and of the results of evaluation, providing an insight into how universities and research institutes assure the quality of their research.

A description of the current system and a brief outline of the developments that gave rise to this system are presented. Figures on the number of evaluations performed and scores awarded provide an insight into the results; a summary of surveys shows how users perceive the protocols. A characterisation of a number of systems in other countries then places the Dutch system in an international perspective.

The Rathenau Instituut promotes the formation of political and public opinion on science and technology. To this end, the Institute studies the organisation and development of science systems, publishes about social impact of new technologies, and organises debates on issues and dilemmas in science and technology.

1 Leonie van Drooge and Stefan de Jong are researchers working at the Rathenau Instituut. This publication was produced with the assistance of Jasper Deuten, Catherine Chiong Meza and Barend van der Meulen of the Rathenau Instituut.

2 Marike Faber and Don Westerheijden are researchers working at CHEPS (Center for Higher Education Policy Studies), University of Twente.

The main conclusions:
- The Netherlands has a long-standing and stable system of quality assurance in academic research, including in comparison with other countries. Thus far, however, we have no overview of the evaluations performed and positions adopted by the boards of research organisations in response.
- There is great variety in terms of the scope of evaluations, ranging from entire disciplines and entire disciplines minus a few organisations, or a combination of disciplines within a research organisation, to a single centre or research group. The quality of the research covered by these evaluations cannot therefore be systematically compared.
- Scores for the quality of research have risen over the past twenty years. Currently, virtually all aspects of all research rate as at least internationally competitive. As a result, there are barely any observable differences between the scores of different research units.
- The Dutch system differs considerably from other national systems. The Netherlands has no national goal, predefined consequences or central organisation that is responsible for the system. Goals are defined at research organisation level and organisations themselves are responsible for the evaluations, and for deciding what consequences should apply.

2. The current system of institutional research evaluation: SEP 2009-2015

The Standard Evaluation Protocol (SEP) 2009-2015[3] describes the current system of research evaluation. It concerns institutional evaluation, i.e. evaluation of research units at universities, including university medical centres, and at the institutes affiliated to the Netherlands Organisation for Scientific Research (NWO) and the Royal Netherlands Academy of Arts and Sciences (KNAW).

There is a fixed protocol for evaluation:
- All research should be evaluated externally once every six years.
- An internal mid-term evaluation three years later serves to monitor measures taken in response to the external evaluation.
- Evaluations take place at two levels: that of the individual research unit (group, programme) and that of the coordinating research institute as a whole.
- The criteria are:
  - academic quality;
  - academic productivity;
  - societal relevance;
  - vitality and feasibility.
- The goals are:
  - to improve the quality of research;
  - to provide accountability for the use of public money towards the research organisation's board, funding bodies, the government and society at large.

The board of the research organisation commissions the evaluation.
- The board decides which institute is to be evaluated when.
- It approves the terms of references (TOR) for each evaluation.
- It establishes the peer review committee (PRC) and appoints its members.

The institute presents the research in a self-evaluation report.
- The unit evaluated/the institute describes the mission, goals and context of the research in a self-evaluation.
- The self-evaluation includes a description of academic quality and relevance, societal relevance and prospects for the future.
- It includes a strategy for the future based on an analysis of strengths and weaknesses, or SWOT (Strengths, Weaknesses, Opportunities, Threats).
- It provides an overview of input data (resources, staff) and relevant output data.

3 VSNU, NWO, KNAW (2009). *Standard Evaluation Protocol 2009-2015*. Amsterdam: KNAW.

The PRC makes an assessment.
- The PRC's assessment is based on the self-evaluation report and a meeting with representatives of the institute and the units, generally during a site visit.
- The PRC describes its assessment in an evaluation report.
- The assessment of a research unit should encompass all four criteria and relate both to performance over the period under review and to plans for the future.
- The assessment should include a qualitative summary of the main findings and a score on a specially developed five-point scale.
- The assessment of an institute should cover policy and management, be geared to the future and should at least include an account of quality.

The board of the research organisation rounds off the evaluation.
- The board receives the evaluation report.
- After consultations with the institute, the board arrives at a position in response to the assessment and recommendations of the PRC.
- The evaluation report and the board's position must be made public.

The SEP 2009-2015 also describes the meta-evaluation of the protocol itself. In other words, how the participating organisations must account for the proper use of the protocol.
- Each organisation must publish a list of forthcoming evaluations.
- Each organisation should list completed evaluations and the board's positions in its annual report.
- KNAW, NWO and the Association of Universities in the Netherlands (VSNU) will together ensure that an independent expert committee evaluates the SEP 2009-2015 in 2013; the evaluation should concern not only use of the protocol, but also the impact of the evaluations on the policies of the organisations. The results will be made public.

The SEP 2009-2015 is just part of a thirty-year tradition of institutional research evaluation in the Netherlands.

3. History and statutory framework

Until 1982, responsibility for quality assurance in university research lay with the faculties themselves. Faculty committees assessed research on its academic merits, its feasibility and its composition. The system was criticised for its failure to provide insight into research efforts and the absence of accountability to the funding body, the Ministry of Education, Culture and Science. This all changed in 1982, with the introduction of Conditional Financing, a procedure whereby assessment committees made up of external experts would assess research programmes.

In 1985 the science minister introduced a new administrative philosophy for universities,[4] based on autonomy and self-regulation. In exchange for greater autonomy, universities would have to show they could deliver quality. The new approach was placed on a statutory footing in 1992 and universities were also given the responsibility of ensuring that the quality of their work was regularly assessed. The relevant provision of the Higher Education and Research Act is shown in Box 1.

4 House of Representatives of the States-General (1985-1986 session). *Beleidsnota Hoger Onderwijs Autonomie en Kwaliteit*, 19 253, nos. 1-2.

Box 1: Higher Education and Research Act
The board of the institution (…) shall ensure that regular assessment of the quality of the activities of the institution takes place, with the involvement of independent experts, and in collaboration as far as possible with other institutions. (…) Insofar as the assessment involves independent experts, the outcomes shall be made public. (Higher Education and Research Act, section 1.18).

Our Minister may subject the funding of research at universities to certain conditions relating to quality assurance. (Higher Education and Research Act, section 2.5, subsection 2).

4. VSNU protocols 1993, 1994 and 1998[5]

In response to the new legislation, in the early 1990s universities association VSNU developed a national system of research evaluation, in consultation with NWO and KNAW. At the core of the system lay regular assessment of all university research performed in a particular discipline by international assessment committees.

The first general protocol for quality assessment in academic research was adopted in February 1993. Trial evaluations were conducted in mechanical engineering, biology, psychology and historical sciences. In response to these trials, VSNU made some adjustments to the protocol. Over the following four years, all other university research was assessed using the VSNU protocol 1994. After an evaluation of the protocol, VSNU decided to organise a new round of research assessment. The goals and criteria in the VSNU protocol remained largely unchanged in the new 1998 version.

5. Standard Evaluation Protocols 2003-2009[6] and 2009-2015

In 1999 VSNU, NWO and KNAW established the Quality Assurance in Academic Research working group, which in 2000 published a report outlining a new national system of quality assurance.[7] This report provided the basis for the Standard Evaluation Protocol (SEP).

The SEP 2003-2009 introduced some important changes relative to the VSNU protocols. VSNU was no longer responsible for organising evaluation; this would henceforth be up to the research organisations themselves. The system of comparative assessment at national level was also abandoned, resolving a major problem. Too often, the discipline-wide assessments had been a matter of comparing apples and oranges, undermining the broader value of the assessment.[8]

The Quality Assurance in Academic Research Meta Evaluation Committee (MEC) evaluated the SEP 2003-2009 in 2007[9] and 2009.[10] Although the outcomes were overwhelmingly positive, there were two points of criticism. It was not clear what research organisations were doing with the results of the evaluations, and the scores were being inflated. The SEP 2009-2015 is largely a continuation of the SEP 2003-2009, with a few minor changes, some of them in response to the MEC's criticisms.

5 VSNU (1993). *Quality Assessment of Research – protocol 1993*. Utrecht: VSNU; VSNU (1994). *Quality Assessment of Research – protocol 1994*. Utrecht: VSNU; VSNU (1998). Protocol 1998. In: *Series Assessment of Research Quality*. Utrecht: VSNU.

6 VSNU, NWO, KNAW (2003). *Standard Evaluation Protocol 2003-2009*. Utrecht: VSNU.

7 Werkgroep Kwaliteitszorg Wetenschappelijk Onderzoek (2001), *Kwaliteit verplicht. Naar een nieuw stelsel van kwaliteitszorg voor het wetenschappelijk onderzoek*. Amsterdam: KNAW.

8 See note 7, p. 41.

9 Meta Evaluatie Commissie Kwaliteitszorg Wetenschappelijk Onderzoek (2007). *Trust but Verify*. Amsterdam: KNAW.

10 Meta Evaluatie Commissie Kwaliteitszorg Wetenschappelijk Onderzoek (2009). *E-VA-LU-ER-EN. Het beoordelen van wetenschappelijk onderzoek in de praktijk*. Amsterdam: KNAW.

6. Twenty years of evaluation in the Netherlands: large similarities, slight differences

The protocols used in the Netherlands over the past twenty years are all different. However, there are great similarities between them, and there has been continuity. See Box 2 for a summary of the similarities between the evaluation protocols.

Box 2: similarities between VSNU and SEP protocols

The VSNU and SEP evaluation protocols have the following things in common:

- They are based on a national system for the regular assessment of all academic research.
- The goals are quality enhancement and accountability.
- Assessment takes place at research unit level.
- The four assessment criteria are:
  - academic quality
  - academic productivity
  - relevance
  - feasibility
- Academic peers form a judgment based partly on information provided by the unit.
- The board of the organisation receives the report and is expected to give its response.

If we examine the protocols more closely, however, we are struck by a number of differences between the VSNU and SEP protocols. The main differences are listed in Box 3.

Box 3: differences between SEP and VSNU protocols

The SEP protocols differ from the VSNU protocols in a number of respects:

- Discipline-wide assessment has been abandoned; research organisations are no longer obliged to organise joint evaluations with all other organisations performing research in the same discipline.
- The secondary goal of enabling the government to use the evaluation to survey the discipline has been abandoned.
- The protocol also applies to research conducted at KNAW and NWO institutes.
- Responsibility for the system lies entirely with the research organisations, so each organisation is free to make its own decisions concerning:
  - the scheduling of evaluations;
  - drafting the terms of reference for the evaluation;
  - setting up and appointing the PRC;
  - responding to the PRC's report and deciding what consequences should apply.
- Research organisations are no longer obliged to submit the report to the Ministry of Education, Culture and Science.[11]

11 SEP 2003-2009 did still oblige research organisations to report results to the Ministry; SEP 2009-2015 merely stipulates that the results must be made public, preferably via the research organisation's website.

Table 1 Meaning of scores in the different protocols

Score	VSNU 1994 en 1998	SEP 2003-2009	SEP 2009-2015
5	Excellent	Excellent international leader; most likely important and substantial impact	Excellent world leading; has important and substantial impact
4	Good	Very good internationally competitive and national leader; expected to make significant contribution	Very good internationally competitive and nationally leading; makes a significant contribution
3	Satisfactory/average	Good internationally visible and national player; will probably make valuable contribution	Good internationally visible and nationally competitive; makes a valuable contribution
2	Unsatisfactory	Satisfactory nationally visible; solid, not exciting, will add to understanding	Satisfactory nationally visible; solid, not exciting, adds to understanding
1	Poor	Unsatisfactory neither solid, nor exciting; not worth pursuing	Unsatisfactory neither solid, nor exciting

Source: VSNU (1994), VSNU (1998), VSNU, NWO, KNAW (2003), VSNU, NWO, KNAW (2009)

Rathenau Instituut

The meaning of the scores on the five-point scale has changed several times (Table 1). While, in the VSNU protocols of 1994 and 1998, a score of 1 or 2 was unsatisfactory, under the SEP only a score of 1 is unsatisfactory. Another new feature of the SEP is the score 'very good', between 'good' and 'excellent'. And, while the VSNU protocols gave only a one-word explanation of the score, the SEP protocols have given increasingly detailed descriptions. The SEP 2003-2009 expressed expectations, in terms of 'most likely' and 'will probably', while the 2009-2015 includes more precisely worded observations ('has', 'makes'). These changes were made in response to criticism from the MEC that scores were being inflated.

7. Twenty years of evaluation in the Netherlands: gathering and providing access to data

The development of the PER base by the Center for Higher Education Policy Studies (CHEPS) provided access to the results of twenty years of research evaluation.[12] The PER Base is a database containing data on all known evaluations since 1993.[13] It covers institutional research evaluation conducted in accordance with the VSNU and SEP protocols. A list of reports is available online at www.rathenau.nl.

It proved difficult to obtain a good overview of all research evaluations conducted in the past twenty years. A number of things remain uncertain. It is not for instance clear what evaluations were conducted. The evaluation reports published up to 2000 contain a list of evaluations already published. The reports mentioned in these lists have been entered in the PER Base. The database is therefore complete in terms of evaluations conducted under the VSNU 1993 and 1994 protocols. This cannot be established with certainty in the case of the VSNU protocol of 1998.

Since the introduction of the SEP, no central record of planned and completed evaluations has been kept. Some research organisations fail to comply with the obligation to make planned evaluations and results public. Nor are they required to submit reports to a particular body.

It is also unclear precisely what research has been evaluated. Not all research has been evaluated once under each protocol. Physics, astronomy and some areas of agricultural science were not for example evaluated under the 1998 VSNU protocol. To what extent all units have been evaluated under the SEP protocols cannot be established because of the sheer diversity of evaluations.

12 The PER Base was developed by the University of Twente's Center for Higher Education Policy Studies, with funding from the Ministry of Education, Culture and Science as part of the CHERPA project. CHEPS is responsible for the data on evaluations up to and including 2009; the Rathenau Instituut is responsible for the data since 2010 Persistent Identifier: urn:nbn:nl:ui:13-c8tn-eh.

13 It contains only data from research organizations that comply with the definition of the protocol applicable at the time (VSNU protocols: universities; SEP: universities, KNAW and NWO institutes).

Furthermore, we have no idea of the scale of the research evaluated in terms of FTEs. Some reports contain no data on this, while others report overall figures (the total over the years), annual figures for individual categories of staff, or figures for a particular year. As such, it is not possible to tell whether all the research performed by an evaluated unit has actually been assessed.

Finally, the scope of the evaluation reports varies widely, particularly since the introduction of the SEP 2003-2009. This restricts the view of which units have been evaluated. Under the SEP, for example, units across an entire discipline are no longer obliged to act in collaboration. Evaluations involving all research organisations together still take place; in some disciplines virtually every university organises an independent evaluation, as in the case of physics (seven reports on university research groups and three on NWO institutes over the period 2003-2009). There are also hybrid situations, whereby a number of universities work together and just one or a few organise an independent evaluation, as has happened in philosophy (2003-2009), psychology (2009-2015) and law (2003-2009).

The disciplinary composition of the research evaluated also changed with the introduction of the SEP. Evaluations of several disciplines are sometimes conducted simultaneously, as in the evaluation of science and technology (University of Groningen, 2005) or of social sciences (Vrije Universiteit Amsterdam, 2008). Mainly, however, subdisciplines are evaluated, as in the evaluation of a centre for gender and diversity (Maastricht, 2005), a single chair in meteorology and air quality (Wageningen UR, 2004) and an institute for sociocultural research (Radboud University Nijmegen, 2006).

The many changes over the past twenty years in the names, size and composition of research groups make it impossible to analyse the development of individual research units. In only a few cases is it possible to track the performance of a unit over the years. The Information Processing and Task Performance research group (University of Groningen, psychology) is mentioned in all evaluations over the past twenty years. This is an exception, however.

8. Numbers of evaluations

Twenty years of institutional research evaluation has produced 222 evaluation reports presenting assessments of 4765 units.[14] Figure 1 shows the numbers of evaluation reports by year; Figure 2 shows the numbers of units evaluated (research groups, programmes, departments) each year.

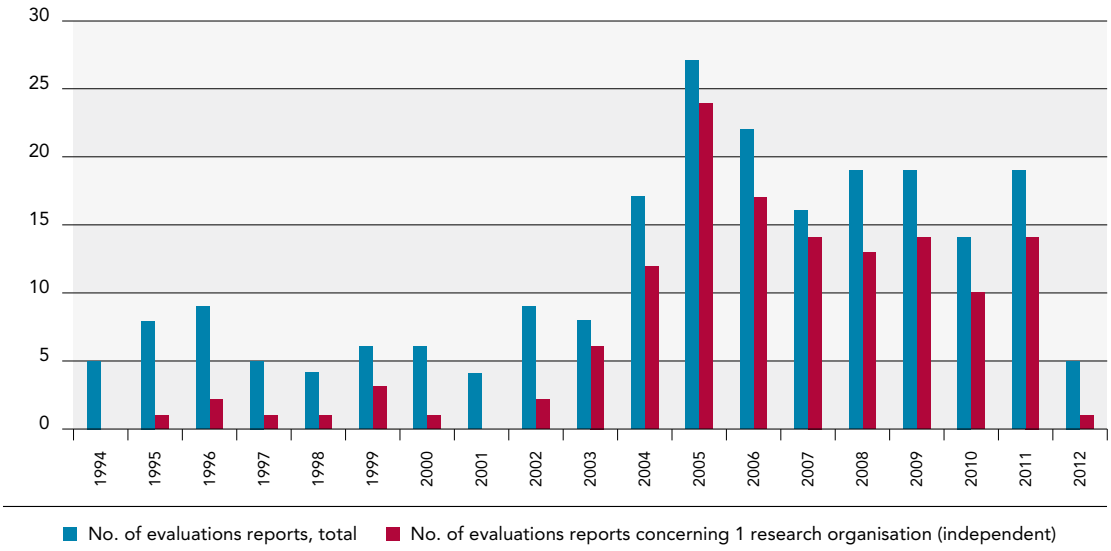
With the introduction of the SEP in 2003 the number of evaluation reports published each year increased sharply. A large proportion of this increase can be attributed to the increase in the number of independent evaluations. An independent evaluation is conducted at a single research organisation (university or institute), and may involve one or more research units at that organisation. This rise in independent evaluations is in accordance with the changes to the protocol, whereby discipline-wide assessment was abandoned, the boards of the research organisations became responsible and the KNAW and NWO institutes also began to take part.

Besides the total number of evaluation reports, Figure 1 also shows the number of independent evaluation reports by year. They account for 136 of the 222 reports published between 1994 and 2012. The figure also shows that a few independent evaluations took place prior to 2003, because some disciplines exist at only one research organisation, such as aerospace engineering and marine technology.

An evaluation report assesses an average of 21 units. There are 25 known reports in which a single unit was assessed. By far the longest report is that on medicine from 1994, which assessed 572 units. This is followed by a report on chemistry from 1996, presenting assessments of 162 units.

14 Only those units that were covered by the protocol at the time of evaluation are included in the database. This means that figures on TNO institutes evaluated under the VSNU protocol, and the NGI's BSIK programme evaluated under the SEP have been disregarded.

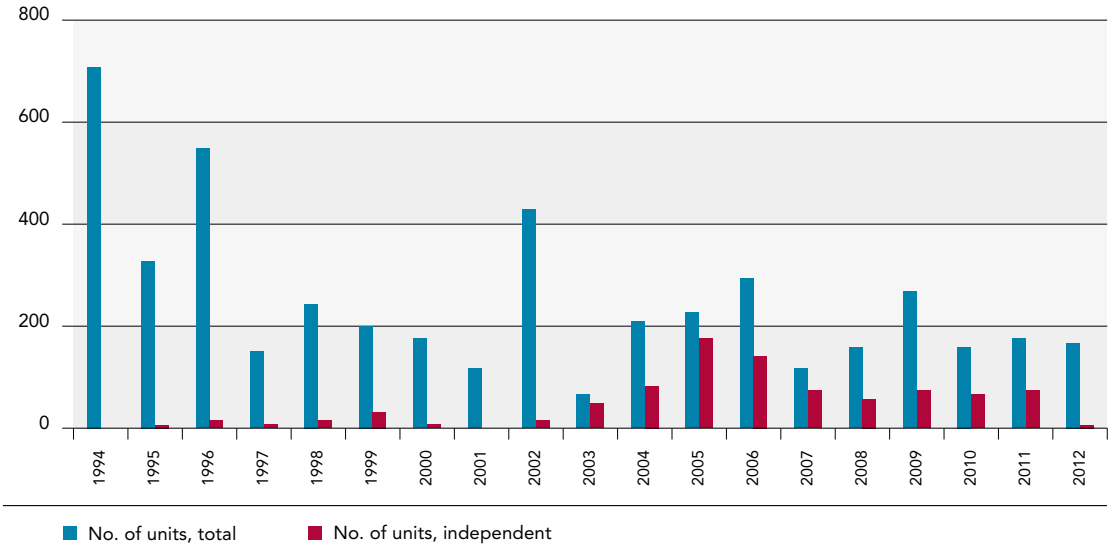
Figure 1 Number of evaluation reports by year



Source: PER Base (CHEPS and Rathenau Instituut)

Rathenau Instituut

Figure 2 Number of units evaluated by year



Source: PER Base (CHEPS and Rathenau Instituut)

Rathenau Instituut

9. Scores

The PER Base contains each unit's scores on the five-point scale for the four criteria. The analysis below is based on these scores. The PRCs also give a qualitative assessment in the report. Those assessments, of 4765 units times four criteria, have not been included in the analysis, which focuses on the scores on the five-point scale. In practice, too, these scores are regarded as very important.

One point to consider when interpreting the scores is how they are actually awarded. The protocols describe the scores as a code for a particular assessment. In practice, a considerable number of scores awarded are not in the form of a whole number. The use of fractional scores (such as 4.5 or 4.13) suggests that the PRC does not regard the score as a code, but as a sliding scale. Our analysis of the scores is based on this practice.

Some remarks about the analysis of the scores. The first four evaluations – mechanical engineering, biology, psychology and historical sciences – have been disregarded, as the VSNU protocol 1993, which was used for these evaluations, used a three-point scale.

The analysis concerns only scores that fit the system used in the protocol applicable at the time.[15] This means that, though the report on medicine from 1994 has been included in the database, its scores were not analysed. This evaluation awarded only two scores per research group, and they could not be clearly correlated with the prescribed criteria.

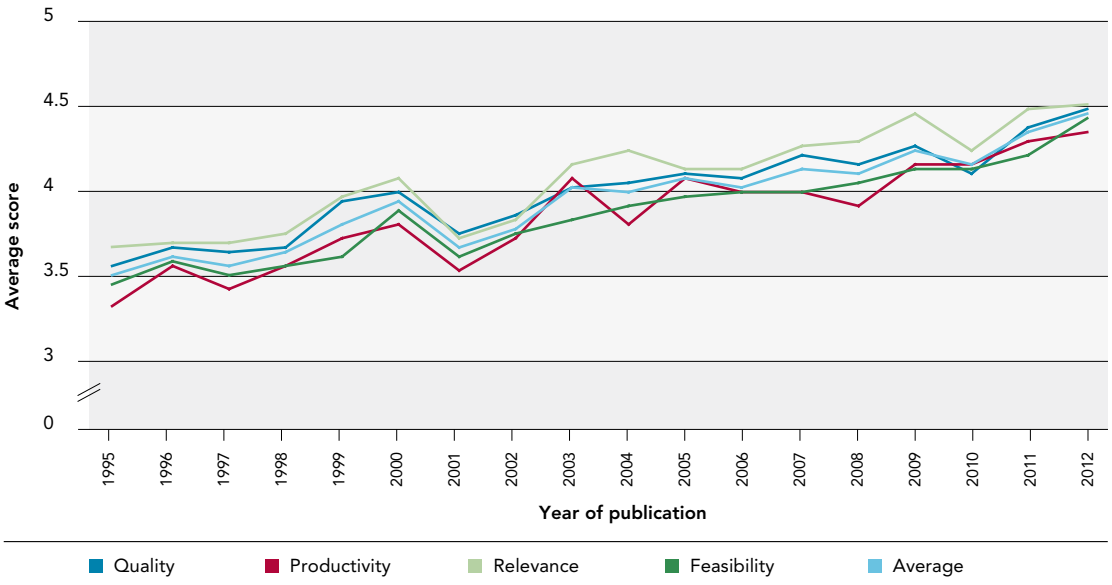
The scores 'not applicable' or 'not assessable' were not considered in the analysis. They apply to 13% of cases, ranging within individual criteria from 10% (academic quality) to 16% (academic productivity). The implication is that a different number of scores has been registered for each criterion.

The widespread impression that scores are being inflated is confirmed by the data in Figure 3 and Table 2, which show the average scores for each of the criteria and the average for the four criteria by year (Figure 3) and by protocol (Table 2). The scores for each of the four criteria show an upward trend. Further analysis shows that the distribution of the scores is small and has narrowed under each successive protocol. Since the introduction of SEP 2009-2015 the most common score for all criteria is a 5, or 'world leading'. In our description of the protocols we pointed out that the meaning of the scores has changed on a number of occasions (Table 1). As a result, scores should in fact have fallen, assuming quality has remained the same. If we take the description of the scores in Table 1 seriously, the trend means that the rise is even greater than is suggested by the table and figure.

Table 3 undermines the expectation that the rise in the scores can be attributed to the increase in independent evaluations (involving just a single research organisation). The differences between the two types of evaluation are small. The suspicion had been that independent evaluations might be more tailored to the unit or institution, and would therefore produce higher scores.

15 In a few cases the reports give only an assessment of each aspect in words. These were included only if this complied with the scoring system, and could therefore be translated into a score. One example is the report of the evaluation of LUMC in 2006.

Figure 3 Average score by criterion and year



Source: PER Base (CHEPS and Rathenau Instituut)

Rathenau Instituut

Table 2 Average scores by unit, criterion and protocol[16]

	Quality	Productivity	Relevance	Feasibility
VSNU 1994	3.65 (n=1179)	3.47 (n=988)	3.70 (n=1066)	3.53 (n=1014)
VSNU 1998	3.88 (n=1009)	3.70 (n=852)	3.93 (n=1009)	3.72 (n=970)
SEP 2003-2009	4.14 (n=1205)	4.03 (n=1201)	4.23 (n=1217)	4.03 (n=1166)
SEP 2009-2015	4.39 (n=385)	4.31 (n=384)	4.48 (n=387)	4.28 (n=379)

Source: PER Base (CHEPS and Rathenau Instituut)

Rathenau Instituut

Table 3 Average score by unit and protocol, divided into independent and joint evaluations

	Independent evaluation	Joint evaluation
VSNU 1994	3.46 (n=49)	3.58 (n=1138)
VSNU 1998	3.80 (n=39)	3.80 (n=975)
SEP 2003-2009	4.16 (n=565)	4.05 (n=663)
SEP 2009-2015	4.37 (n=132)	4.37 (n=256)

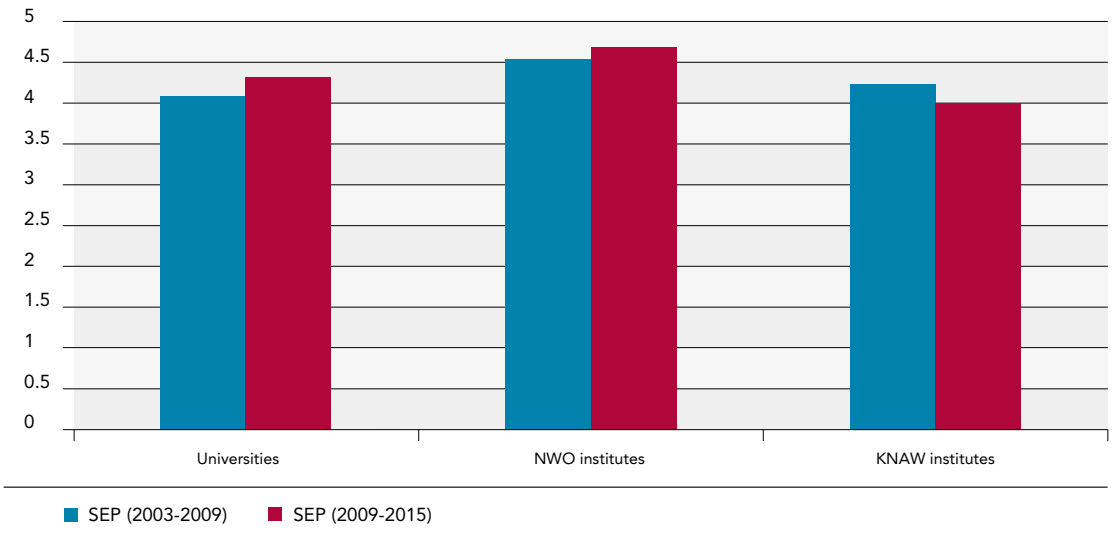
Source: PER Base (CHEPS and Rathenau Instituut)

Rathenau Instituut

16 Evaluations where it was not clear what protocol had been used were excluded.

Since the introduction of the SEP, both NWO and KNAW institutes have been covered by the protocols as well. These research institutes are almost always evaluated independently. Figure 4 shows the average scores for NWO and KNAW institutes as compared with universities. The NWO institutes, which were all evaluated under both protocols, score very highly. The average score for the KNAW institutes seems to be falling. It should however be noted that only a small proportion of KNAW institutes, mainly the smaller ones, have been assessed under SEP 2009-2015.

Figure 4 Average score of universities, and NWO and KNAW institutes

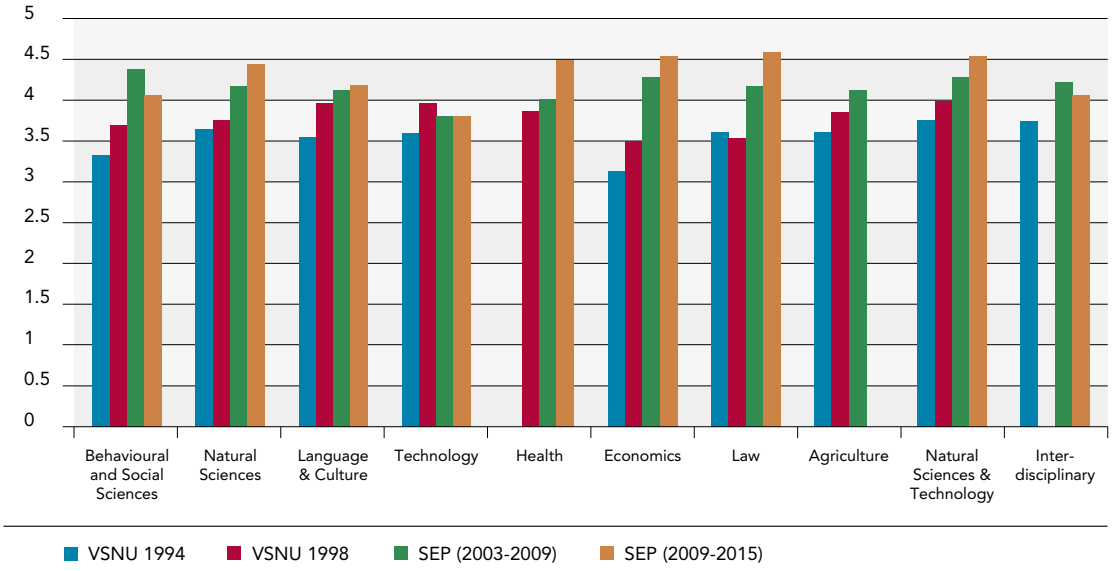


Source: PER Base (CHEPS and Rathenau Instituut)

Figure 5 lists the average score by HOOP field;[17] Figure 6 shows the number of units in each HOOP field that received a score for at least one criterion. The average score has risen in all HOOP fields since the first evaluations, in some by more than a whole point. The only field where the rise has been smaller is Technology. In some fields, a slight fall can be seen over a certain period. In the Behavioural and Social Sciences field, where almost all units have been evaluated under the SEP 2009-2015, the average score has fallen relative to evaluations under the previous protocol. The score for interdisciplinary evaluations has also shown a slight fall, though it should be noted here that this concerns only a small number of units. In general, we can conclude that the rise in scores cannot be attributed to a particular field.

17 HOOP fields are fields of academic study as identified in the Ministry of Education, Culture and Science's Higher Education and Research Plan (known by the Dutch acronym HOOP). For HOOP fields in each report: see online overview (www.rathenau.nl). The HOOP field Natural Sciences/Technology covers disciplines that fall under both Natural Sciences and Technology: physics, astronomy, chemistry, mathematics, computer science.

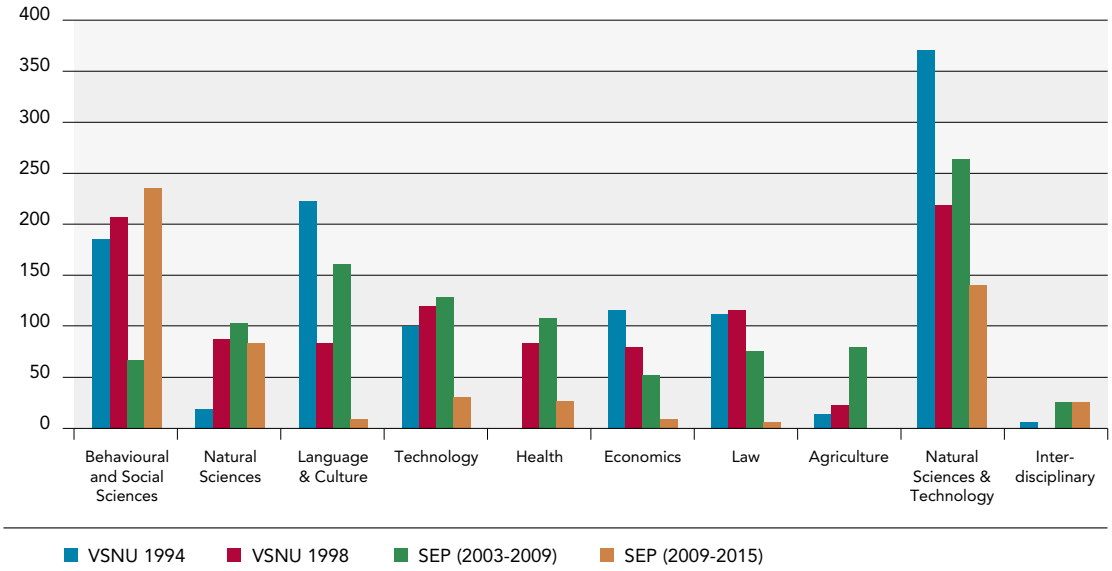
Figure 5 Average score by HOOP field and protocol



Source: PER Base (CHEPS and Rathenau Instituut)

Rathenau Instituut

Figure 6 Number of units evaluated by HOOP field and protocol



Source: PER Base (CHEPS and Rathenau Instituut)

Rathenau Instituut

10. International

To place experiences in the Netherlands in a broader perspective, we have compared the Dutch system with that in a number of other countries that have a top-rate science system. Table 4 shows national systems of institutional evaluation. The comparison is not concerned with funding instruments or programmes, but with the evaluation of research units.

The comparison reveals that other countries have made different choices. Germany and the United States have no national system, for example. Various German *Länder* do have their own system of institutional evaluation, but they are not very stable. There is no system at federal level. The United States recently introduced STAR metrics, a system of accountability for government-funded research, the main goal of which is to identify the impact of the funding system on employment.

Countries that do have a national system also take an approach different to that taken in the Netherlands.[18] They have an organisation (e.g. a national agency or ministry) that is responsible for evaluation. The Netherlands is the only country where the research organisations themselves bear full responsibility.

Those countries also have national goals: to strengthen their international position, to raise quality to 'world leading'. The Netherlands is the only country that has defined its goal as improving quality at research unit level.

Box 4: Goals of evaluation

Evaluation serves various purposes: accountability, reward and improvement. Different questions and outcomes are associated with each of them.

Accountability

The key question in this type of evaluation is whether resources have been used correctly and whether procedure has been followed. It is important that the requirements are clear. A significant proportion of evidence can be gathered by parties other than the institution concerned, in the form of overviews of resources and results, and descriptions of procedure. In essence, the assessment comes down to satisfactory/unsatisfactory.

Ranking and reward

The key question in this type of evaluation is who is the best, and therefore qualifies for a (greater) reward. It is important that agreement is reached on criteria and their weight. The assessment can produce an absolute ranking – from best to worst – or a relative ranking, whereby several evaluated units may be awarded the same ranking.

Improvement

The key question in this type of evaluation is whether performance is making the best possible contribution to the intended mission or goals, and to identify opportunities for improvement. The evaluated unit must have a capacity for self-reflection. The assessment includes a diagnosis of the current situation and recommendations for the future. It will be partly descriptive and partly prescriptive.

18 This concerns England, Spain, Italy, France, Norway, Sweden and Denmark.

Table 4 International comparison of national systems of institutional evaluation

Country	Since	Organisation responsible	Method of evaluation
England[19 20]	1986	Higher Education Funding Council for England (HEFCE)	Peer score for nominated performances of individual researchers (ranking)
Spain	1989	ANEP agency	Assessment of individual researchers' output by expert panel
Netherlands	1993	VSNU/organisations	Peer score for nominated research performances of groups and institutes
Norway	2005	Ministry of Education and Research	Measuring of registered research output using performance indicators (partly determined by discipline)
Denmark[21]	2006	Danish Agency for Science, Technology and Innovation	National evaluation of parts of science system (funding instruments, disciplines, programmes, system components) by peers; based partly on self-evaluation; within framework defined by minister
France[22]	2006	AERES Agency	Self-evaluation and site visits by peers
Australia	2008	Australian Research Council (ARC)	Disciplinary national reviews based on indicators and peer review
Sweden	2008	Ministry of Education and Research, supported by Vetenskapsrådet (research council)	Performance indicators: output and external funding
Italy	2009	ANVUR agency	Peer score based on performance indicators
Germany	No national system		
USA[23]	No national system		

19 HEFCE (2010). Guide to Funding. http://www.hefce.ac.uk/media/hefce1/pubs/hefce/2010/1024/10_24.pdf (accessed 21-11-2012).

20 For England; Scotland, Wales and Northern Ireland have sister organisations. The system is the same.

21 Danish Agency for Science, Technology and Innovation. Research Evaluation Guidelines. http:/en.fi.dk/research-evaluation/framework-and-methods/action-plan-for-research-evaluation/Research%20Evaluation%20Guidelines.pdf (accessed 21-11-2012).

22 LERU (2012). Research Universities and Research Assessment. Louvain: LERU.

23 https://www.starmetrics.nih.gov/ (accessed 26-11-2012).

Country	National goal of system	Goal and implications of evaluations within system
England	To strengthen international status of research by fostering top-quality research	Ranking and reward: performance funding – distribution of 70% of direct govt funding (lump sum portion)
Spain	To raise quality of research to world leading by increasing research effort/output	Ranking and reward: allocation (or refusal) of 6-year research appointments
Netherlands	None	Quality improvement and accountability: no predefined consequences; up to research organisation boards to decide
Norway	To increase research activity and foster excellence	Ranking and reward: performance funding – redistribution of part of direct govt funding (less than 15%; lump sum portion)
Denmark	To provide accountability for national investments in research and to improve the system so that investments lead to excellence	Ranking and reward: performance funding – redistribution of growing proportion of funding (extent to which this happens unknown)
France	To identify excellence and for ranking; also to provide insight into quality for research groups and institutes (improvement) and government (strategic decisions); to inform students, companies and society (accountability)	Ranking and reward: performance funding – redistribution of funding by ministry and within research organisation
Australia	To identify excellence and new areas; international benchmarking; to create incentives to improve quality of research	Ranking and reward: performance funding (planned) – redistribution of lump sum funding, as well as increase in national budget for research
Sweden	Strategic university management designed to encourage improvements in research quality	Ranking and reward: performance funding – distribution of 25% of lump sum funding
Italy	To identify and foster quality	Ranking and reward: performance funding – distribution of 5% of lump sum funding, and disincentives for poor individual performance
Germany		
USA		

In addition, other countries have certain rules concerning the consequences associated with the evaluation in order to help them achieve their goal. In most cases, it is a matter of fostering quality or excellence by some form of performance-related funding: redistribution of part of the lump sum allocation or awarding individual appointments. The Netherlands has no such rules concerning consequences. Box 4 explains the various goals of evaluation.

The implication is that assessment plays a very important role in these countries, as it can have far-reaching consequences. Evaluation must make clear who or what is good enough to receive funding, and who or what is better and therefore deserves more. This explains the commonly used method of precisely defined performance measures.

The Netherlands uses scores too. This suggests that ranking is important. However, this is not the case; there is no need to establish a ranking to achieve the general goals of quality improvement and accountability.

11. Evaluation in practice: perception, utilisation and follow-up

A number of studies of the practice of institutional evaluations have been conducted in the Netherlands over the past few years.[24] They focused on use of the protocols and the consequences of evaluation. The studies revealed the following:

- the system of regular evaluation, of producing a self-evaluation report, of peer review and a management interview involving unit and research organisation board is generally well-regarded as a management instrument by administrators, deans and research group leaders;
- the results of evaluations can play a role in administrative decisions (such as whether to disband a unit or cut funding), but are never the only grounds for such decisions;
- low scores often have direct consequences, perhaps in the form of a binding order to improve, and resources may be made available for the purpose;
- high scores rarely lead to direct reward, as no financial resources are available for this;
- high-scoring units are however indirectly rewarded, as their ability to recruit staff or attract funding is enhanced;
- the PRC's final assessment and the numerical score awarded cannot always be clearly deduced from the arguments presented;
- those concerned perceive evaluation as a major administrative burden;
- by no means all research organisations publish the results in full.

The various studies conducted in the Netherlands have thus shown that evaluation is appreciated as an instrument of management. There is also appreciation of the fact that administrators are free to take autonomous decisions in response to the outcomes. In terms of the goals of the SEP, the focus is on improving quality and providing accountability to the research organisation board. Those concerned seem to overlook the goal of providing accountability to funding bodies, the government and society. The lack of openness reinforces this impression.

Compared with other countries, there is little criticism of assessments, either of the indicators used or of the results. However, the administrative burden is felt to be excessive.

24 Meta Evaluatie Commissie Kwaliteitszorg Wetenschappelijk Onderzoek (MEC) (2007). *Trust but Verify. Amsterdam*: Meta Evaluatie Commissie Kwaliteitszorg Wetenschappelijk Onderzoek (MEC) (2009). *E-VA-LU-E-REN. Het beoordelen van wetenschappelijk onderzoek in de praktijk.* Amsterdam: KNAW; Ben Jongbloed & Barend van der Meulen (2006). *De follow-up van onderzoeksvisitaties. Onderzoek in opdracht van de Commissie Dynamisering*. Enschede: CHEPS; Barend van der Meulen (2007). 'Interfering Governance and Emerging Centres of Control'. In: Whitley and Glaser (eds), (2007). *The Changing Governance of the Sciences, Sociology of the Sciences Yearbook 26*. Springer. pp. 191-2004; Leonie van Drooge, Stefan de Jong, Jos de Jonge (2012). *Focusgroepen SEP. Verslag van twee bijeenkomsten op 26 september 2012.* The Hague: Rathenau Instituut.

12. Summary

Twenty years of research evaluation have produced the following picture:

The Netherlands has a longstanding and stable system of quality assurance in academic research, particularly in comparison with a number of other countries, which either have no such system or have only introduced one over the past decade.

The Dutch system differs significantly from that in other countries. In contrast to those systems, the Netherlands has no national goal, no central organisation that bears responsibility and no rules concerning the consequences associated with the outcome of an evaluation. The Dutch situation is unique, as responsibility lies entirely with the research organisations themselves. Users appreciate the system as a management instrument and are in favour of the autonomy that administrators enjoy in responding to the results of evaluations.

The research organisations have taken advantage of the opportunity to decide for themselves which units should be evaluated. There is great variety in terms of the scope of evaluations, ranging from national evaluations of entire disciplines and evaluations of entire disciplines minus a single organisation, or a combination of disciplines within an organisation, to evaluations of a single centre or research group. No complete overview of evaluations has ever been produced before. There is still no overview of the positions the research organisation boards have adopted in response to evaluations, and of the consequences attaching to the results. Comparability between evaluations is poor.

Scores for all criteria – academic quality and productivity, relevance and feasibility – have risen over the past twenty years. Virtually all research currently qualifies as internationally competitive. As a result, it is virtually impossible to identify any differences in quality between research units.

Box 5: Abbreviations

HOOP	Academic fields identified in the education ministry's Higher Education & Research Plan
KNAW	Royal Netherlands Academy of Arts and Sciences
MEC	Quality Assurance in Academic Research Meta Evaluation Committee
NWO	Netherlands Organisation for Scientific Research
PRC	Peer Review Committee
SEP	Standard Evaluation Protocol
SWOT	Strengths, Weaknesses, Opportunities, Threats
TOR	Terms of Reference
VSNU	Association of Universities in the Netherlands

13. Sources and references

CHEPS and Rathenau Instituut: PER Base. Persistent Identifier: urn:nbn:nl:ui:13-c8tn-eh.

Danish Agency for Science, Technology and Innovation. Research Evaluation Guidelines. http://en.fi.dk/research/research-evaluation/framework-and-methods/action-plan-for-researchevaluation/Research%20Evaluation%20Guidelines.pdf (accessed on 21-11-2012).

Leonie van Drooge, Stefan de Jong, Jos de Jonge (2012). *Focusgroepen SEP. Verslag van twee bijeenkomsten op 26 september 2012*. The Hague: Rathenau Instituut.

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

HEFCE (2010). Guide to Funding. http://www.hefce.ac.uk/media/hefce1/pubs/hefce/2010/1024/10_24.pdf (accessed on 21-11-2012).

Ben Jongbloed & Barend van der Meulen (2006). *De follow-up van onderzoeksvisitaties. Onderzoek in opdracht van de Commissie Dynamisering*. Enschede: CHEPS.

LERU (2012). *Research universities and research assessment*. Louvain: LERU

Meta Evaluatie Commissie Kwaliteitszorg Wetenschappelijk Onderzoek *Trust but Verify*. Amsterdam: KNAW.

Meta Evaluatie Commissie Kwaliteitszorg Wetenschappelijk Onderzoek *E-VA-LU-E-REN. Het beoordelen van wetenschappelijk onderzoek in de praktijk*. Amsterdam: KNAW.

Barend van der Meulen (2007). 'Interfering Governance and Emerging Centres of Control'. In: Whitley and Glaser (eds), (2007). *The Changing Governance of the Sciences, Sociology of the Sciences Yearbook 26*. Springer. pp. 191-2004.

https://www.starmetrics.nih.gov/ (accessed on 26-11-2012).

House of Representatives of the States-General (1985-1986 session). *Beleidsnota Hoger Onderwijs Autonomie en Kwaliteit*, 19 253, nos. 1-2.

VSNU (1993). *Quality Assessment of Research – protocol 1993*. Utrecht: VSNU.

VSNU (1994). *Quality Assessment of Research – protocol 1994*. Utrecht: VSNU.

VSNU (1998). Protocol 1998. In: *Series Assessment of Research Quality*. Utrecht: VSNU.

VSNU, NWO, KNAW (2003). *Standard Evaluation Protocol 2003-2009*. Utrecht: VSNU.

VSNU, NWO, KNAW (2009). *Standard Evaluation Protocol 2009-2015*. Amsterdam: KNAW.

Werkgroep Kwaliteitszorg Wetenschappelijk Onderzoek (2001). *Kwaliteit verplicht. Naar een nieuw stelsel van kwaliteitszorg voor het wetenschappelijk onderzoek*. Amsterdam: KNAW.

Wet Hoger Onderwijs en Wetenschappelijk Onderzoek ('Higher Education and Research Act').

About this publication

This is the eighth publication in the Science System Assessment Facts and Figures series. This edition surveys the development of the system of quality assurance in academic research, using data gathered from a range of sources.

For further information on this publication, please contact the authors, *drs.* Leonie van Drooge (l.vandrooge@rathenau.nl) or Stefan de Jong MSc (s.dejong@rathenau.nl), or the head of the Science System Assessment Department, Dr Barend van der Meulen (b.vandermeulen@rathenau.nl).

Colofon:

May 2013

Rathenau Instituut
Postbus 93566
2509 CJ Den Haag
Tel.: +31 (0)70-3421542
Website: www.rathenau.nl

This publication may be cited as follows:
Leonie van Drooge, Stefan de Jong et al. (2013):
Twenty years of research evaluation,
Facts & Figures 8. The Hague: Rathenau Instituut.

Rathenau Instituut