

**Title**

Two decades of national research evaluation in The Netherlands  
An analysis of the development of the system and its outcomes

**Authors**

Stefan de Jong, Science System Assessment, Rathenau Instituut, Anna van Saksenlaan 51, The Hague, 2593HW, The Netherland.  
E-mail: s.dejong@rathenau.nl

Leonie van Drooge, Science System Assessment, Rathenau Instituut, Anna van Saksenlaan 51, The Hague, 2593HW, The Netherland.  
E-mail: l.vandrooge@rathenau.nl

**Keywords**

Academic research evaluation policy.

## Intro

The demand for accountability is widespread in society. Research is not excluded from this quest for accountability. In the past two decades, a number of countries have introduced national systems for ex-post research evaluations. The first being the UK and The Netherlands. The Dutch system builds upon a tradition of formalized quality care for academic research going back to national policies developed in the 1980s. Currently, a third version of a standard protocol for evaluations is being used.

This paper aims to provide an overview of the development of the Dutch system of research evaluations from its start in 1994 to present date. We take the three most commonly discussed policy issues and complaints about the system as our focus.

## Policy issues

1. *'The system is biased towards natural and biomedical sciences.'*

In general, leading theories and research methods and the definition of quality are much more agreed upon in the natural and biomedical sciences than in the social sciences and humanities. Also, in general the national context in case of the first is of less importance than in the case of the latter. The Dutch evaluation system with set criteria and committees of international peers is believed to be biased towards the natural and biomedical sciences, resulting in higher scores for these disciplines;

2. *'Single organization evaluations have an advantage.'*

The first protocol involved discipline wide evaluations; all research groups in a discipline were evaluated by the same committee in a single evaluation. The succeeding protocols allow evaluations in disciplinary sub groups. The rationale is some disciplines are too diverse to be covered by one committee. In practice, this results in the evaluation of research of a single research organization or even a single research group.;

3. *'Ranking leads to score inflation.'*

There are two main types of evaluation: summative evaluation and formative evaluation. The aim of the first is to assess whether certain quality standards are met, while the aim of the latter is to improve.

In summative research evaluations, which allows ranking, the goal of evaluands will be to score as high as possible to avoid negative consequences. In this scenario research groups might only highlight their best results. In contrast, formative evaluations require a realistic view to identify potential improvements. Currently, results are increasingly discussed in terms of scores and it is believed the summative goal of the system overshadows the formative goal. As a result, scores would be on the rise with a risk of the system losing its distinctive capacity.

## Method

We combined document research with statistical analyses. Our first main data source are the four succeeding evaluation protocols and intermediary updated versions (five documents in total). For each protocol the following data were collected and compared: goal of evaluation; organizations responsible for the evaluation; planning; frequency and organization of evaluation; composition and appointment of committee; evaluands; criteria; scoring; dissemination of results; and meta evaluation of the system.

Our second main data source are the reports in which results are published. A total of 223 reports were collected and this is believed to include all reports published from 1994 to 2013. For each report the following data is collected: version of the protocol; year of publication; discipline (based on classification by Ministry of Science) number of involved universities; and number of involved groups. These data were linked to the research groups. Furthermore, for each research group the scores on the four criteria as well as the university it belongs to was determined. This resulted in a database with all the scores attributed to 4799 evaluated units<sup>12</sup>. Mean, median and mode per year and protocol were determined per criterion and mean of the criteria.

## Results

### Protocols

When comparing the succeeding protocols, similarity and continuity stand out:

- Quality care is mandatory by law;
- National system for the evaluation of academic research, serving two goals: accountability (summative) and improvement (formative);
- Evaluation on the level of research groups and periodically organized;
- Four criteria: quality, productivity, relevance and viability;
- The protocol describes what information should be provided to enable a committee of scientific peers to evaluate the research;
- Results are a score (1-5) on each of the four criteria accompanied by a textual explanation;
- The board of the institute is the official recipient of the report and has to provide a formal reaction. The board is free to decide what consequences of the report are.

Nevertheless, there are two major changes. Firstly, initially evaluations were organized on a central level. Only university based research was evaluated per discipline in a national system, which was organized by the Association of Dutch Universities (VSNU). From 2003 onwards, also the research institutes part of the Dutch Royal Society of Arts and Sciences (KNAW) and Netherlands Organisation for Scientific Research (NWO) are evaluated in the national system. The organization of evaluations is a responsibility of the boards of the institutes. It is them who decide what the task of the committee is and which groups are evaluated when and by whom. As a cons performance they have to publish the results as well as their formal reaction to the report.

Secondly, the definition of the scores changed. In the first protocol scores 1 and 2 indicate an insufficient level, in protocol two and three only score 1 is insufficient. Furthermore, in the second protocol scores reflect a *potential* whether in the third protocol they reflect a *reality* (for instance score 5 in the second protocol means *international leader; most likely important and substantial impact* while in the third protocol it *means world leading; has important and substantial impact*.)

---

<sup>1</sup> PER Base, constructed by CHEPS, University of Twente, The Netherlands.

<sup>2</sup> Not every research group has been evaluated on each of the four criteria; a single research group is likely to have been evaluated multiple times from 1994-2008.

Scores

Issue 1: The system is biased towards natural and biomedical sciences.

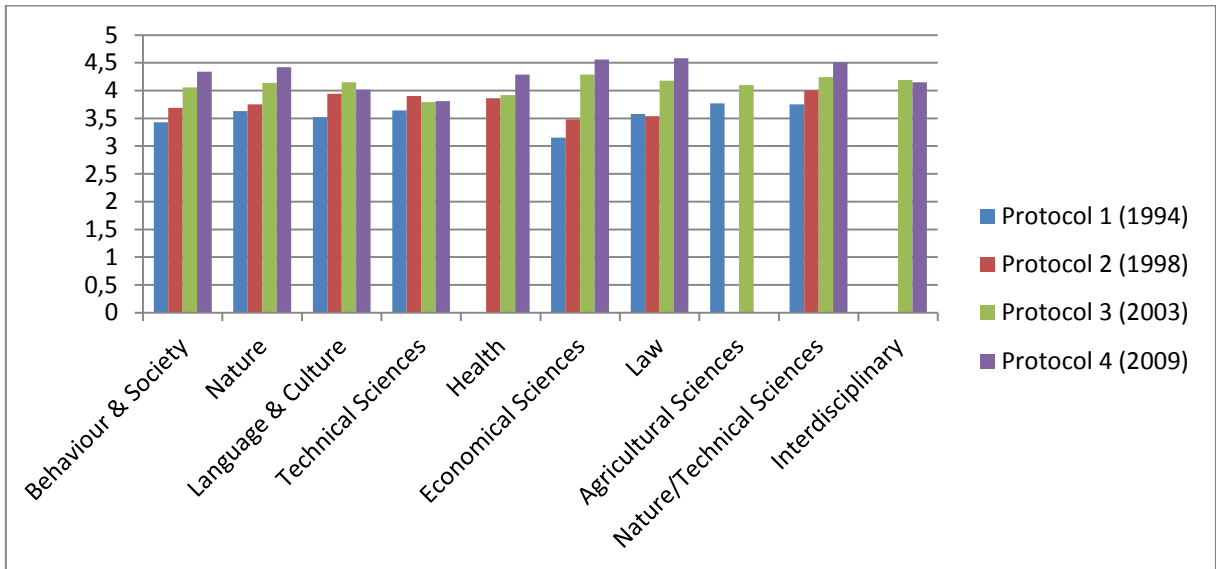


Figure 1: Mean score per discipline per protocol

Figure 1 shows the mean scores per discipline. As becomes clear from this graph, scores increased in all disciplines. Although there are difference between disciplines, there is no clear distinction between natural and biomedical sciences and social sciences and humanities.

Issue 2: Single organization evaluations have an advantage

Figure 2 shows the number of single research organization evaluations increased with the introduction of the second protocol. However, as table 1 shows, differences between mean scores of single organization evaluations and multiple organization evaluations are small and not per se in favour of single organization evaluations.

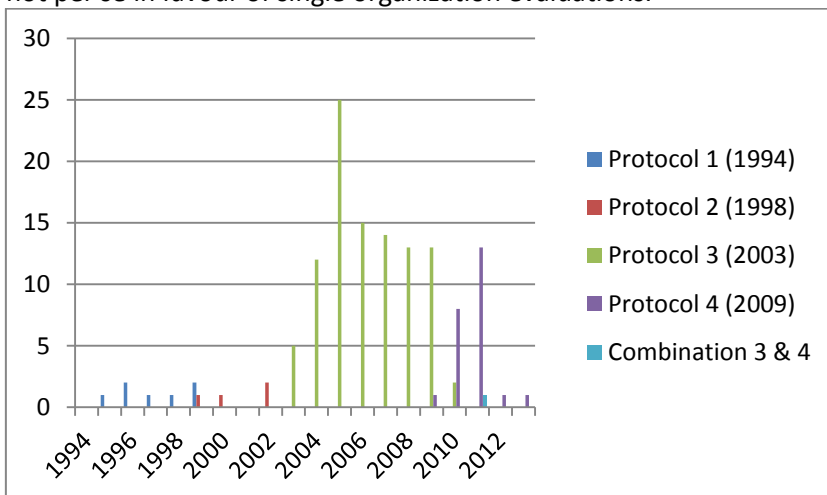


Figure 2: Single university evaluations per protocol and year

**Table 1: Mean scores of single organization evaluations and multiple organization evaluations per protocol**

	Single organization evaluation	Multiple organization evaluation
<b>1<sup>st</sup> Protocol (1994)</b>	3.57 (n=58)	3.57 (n=1118)
<b>2<sup>nd</sup> Protocol (1998)</b>	3.70 (n=30)	3.79 (n=953)
<b>3<sup>rd</sup> Protocol (2003)</b>	4.13 (n=585)	4.07 (n=680)
<b>4<sup>th</sup> Protocol (2009)</b>	4.34 (n=133)	4.37 (n=256)

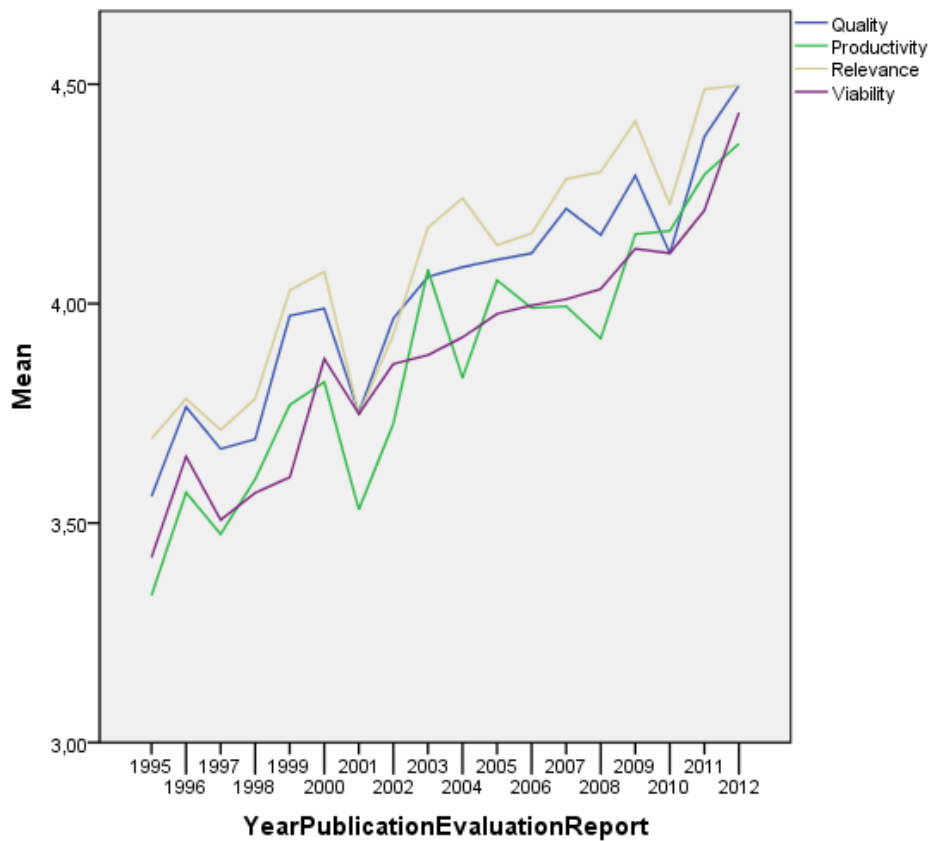
### Issue 3: Summative evaluations leads to score inflation

Table 2 shows mean scores for all criteria have increased by at least 0.75 point since 1994. This could indicate the overall quality of research in the Netherlands increased. Though this may be the case, table 2 also shows the score of 5 currently is the most obtained score for all four criteria<sup>3</sup>. On top of that, 50% of the scores are 4 or higher. This indicates the distinctive capacity of the system indeed is quite limited. Notably, the meaning of score 5, ‘world leading’ is the most given assessment at the moment.

**Table 2: Statistical values per evaluation criterion per protocol**

	Value	Quality	Productivity	Relevance	Viability
<b>1<sup>st</sup> Protocol (1994)</b>	Mean	3.63 (n=1168)	3.45 (n=977)	3.69 (n=1056)	3.52 (n=1000)
	Median	4.00 (n=1168)	3.00 (n=977)	4.00 (n=1056)	4.00 (n=1000)
	Mode	4.00 (n=1168)	3.00 (n=977)	4.00 (n=1056)	4.00 (n=1000)
<b>2<sup>nd</sup> Protocol (1998)</b>	Mean	3.87 (n=978)	3.70 (n=821)	3.92 (n=978)	3.72 (n=942)
	Median	4.00 (n=978)	4.00 (n=821)	4.00 (n=978)	4.00 (n=942)
	Mode	4.00 (n=978)	4.00 (n=821)	4.00 (n=978)	4.00 (n=942)
<b>3<sup>rd</sup> Protocol (2003)</b>	Mean	4.15 (n=1231)	4.03 (n=1226)	4.23 (n=1246)	4.03 (n=1200)
	Median	4.00 (n=1231)	4.00 (n=1226)	4.00 (n=1246)	4.00 (n=1200)
	Mode	4.00 (n=1231)	4.00 (n=1226)	4.00 (n=1246)	4.00 (n=1200)
<b>4<sup>th</sup> Protocol (2009)</b>	Mean	4.39 (n=385)	4.31 (n=384)	4.48 (n=387)	4.28 (n=379)
	Median	4.50 (n=385)	4.50 (n=384)	4.50 (n=384)	4.00 (n=384)
	Mode	5.00 (n=385)	5.00 (n=384)	5.00 (n=384)	5.00 (n=384)

<sup>3</sup> in the case of the criterion productivity scores 5 and 4 are obtained in equal numbers.



Figuur 3: Mean score per criterion per year

### Conclusion

The Netherlands has a long lasting and stable system of quality control. The system does not seem to be biased towards some disciplines or size of the evaluand population. The goals of the system have always been and still are quality care and accountability. However, the emphasis on the scores suggests ranking is the goal. At the moment the scores suggest almost all Dutch research is at least internationally competitive. However, inflation of the scores makes it difficult to show differences between groups. The inflation is also known for the UK's RAE. In contrast to the UK, in The Netherlands funding is not directly related to evaluation outcomes. Therefore it is an option to revise the system of scoring.