Barend J.R. van der Meulen

Understanding evaluation processes in research systems in transition

Traditional evaluation practices in science appear to function smoothly. True, there are criticisms. But as things go, there is a tendency to see the conduct of evaluations and their function in the system of science, as self-evident, and somehow made possible by a natural competence of scientists to evaluate. This assumption is enshrined in the various peer review practices, and in the way new quality measures like bibliometric ones are validated against peer judgments. As long as one limits oneself to the study of stabilized, institutionalized evaluation practices, the assumption of smoothly functioning practices and natural competence is readily adopted. This is exactly what the classical studies did, often in terms of the reward system of science (Merton, 1973; Hagstrom, 1965; Crane, 1972; Zuckerman, 1967; Gaston, 1968; Blume and Sinclair, 1973).

For new evaluation demands, and for newly emerging practices, smooth functioning and competence to evaluate must still be constructed. This is indeed what discussions are about with new and more demanding sponsors of science. Because of the dependence on resources scientists cannot avoid the evaluations. They only can try to influence the evaluation practices by criticizing, discussing or joining them – and so they do. Several questions can then be raised: Who should actually conduct the evaluation: who is competent? What should be evaluated: scientists, programs, research organisation, articles? Which criteria should be used? What are the implications of the evaluation? What sort of conclusions can be drawn from it?

Most of the time these questions are discussed ad hoc, i.e. with reference to a specific evaluation practice, a single evaluation or to a study that is thought to bring abuses in an evaluation practice to light. These discussions are valuable as far as they are bring into question the seemingly obviousness of evaluations, and afford a glance behind the scenes of evaluations. They are less valuable if one is interested in the understanding of such things as the functioning of the evaluation process, the position and competence of the evaluator or possibilities to improve evaluation practices. How-

Science Studies, Vol. 8 (1995), No. 1, 24—35.

ever, such understanding is a prerequisite for the organisation and improvement of evaluation processes.

And for the discussion on evaluations in science one can add. If one thing is obvious in the study of evaluation processes it is that almost anybody involved in science has an opinion. The increasing competition over resources, evaluation processes are debated more and more.[1] But the high interests of scientists in peer review, the opportunities to publish almost any opinion about it and the difficulties to study it, have not improved the level of the debate. Moreover, with the high number of participants in the debate one can not see the forest for the trees.

To get the forest into sight again, it is useful to underscore the different levels from which one can look at evaluation processes. In the next sections peer review will be analyzed at the level of the peer, at the level of the decision processes peer review is part of, and at the level of the research system. In the last part the multi-perspective developed is used to discuss the new evaluation processes and their role in the research system in transition. The main topics addressed are evaluation competence and stabilization of evaluation processes.

How peers evaluate

At the micro level one finds the peer evaluating a research product or proposal. The most simple question one can ask at this level is: what does the peer do when (s)he reviews? The question can be asked empirically, e.g. by just studying what kind of criteria peers use to evaluate. Only a few studies limit themselves to that question. Those who did found a variety of criteria, differences in peer review processes, and changing roles of peers during the peer review process itself: from judging 'quality' in the first instance to advising about possible improvements.

Most of the studies combine the empirical question with a normative interest in the reliability, validity, fairness etc. of the peer review system. In the wake of some famous studies of the Coles (1977, 1981) and of Peters & Cecci (1982) there are numerous studies trying to find out whether the peers made good judgments of manuscripts and research proposals. Consequently the findings of these studies are often not reported in terms of how peers evaluate, but whether their evaluations are unbiased, valid, impartial or not.

Definitely, within these studies the competence of evaluators to evaluate is considered to be the cornerstone of the peer review system. But one should be cautious to lay such a burden on the peers shoulders. Quality is a relational concept and evaluators cannot but focus on certain qualities of an object. In order to make an evaluation the evaluator needs a frame or reference with which he can identify himself and to which the work to be judged can be related. Manuscripts submitted to a journal are evaluated on their contribution to a certain field or discipline. Proposals submitted in the context of an R&D stimulation programme are evaluated on their expected contribution to (some of the) the objectives of the programme. The quality that is ascribed to the objects evaluated is not a property of the objects but a relation between the object and the frame of reference, ascribed by an evaluator.

If within actual evaluations frames of references and relations had to be explicated fully, the evaluator would be burdened with an unbearable load. Reviewers need heuristics to form an acceptable judgment and make reductions of complexity.[2] The selection of a frame of reference itself is such a reduction of complexity. Manuscripts, proposals, research programs etc. are seldom linked to one research area only. Mostly they are embedded in a more heterogeneous context and, hence, one can think of various frames of reference to relate the object to. Quality can ascribed to the object with respect to each of these frames of reference and these ascriptions are likely to have different outcomes. Generally, evaluators will limit themselves in the number of frames of

reference as a first reduction of the complexity.

Reductions are also made on the level of the articulation of the relation between object and frame of reference. A well-known reduction on this level is the reputation of scientist(s) and/or institutions (Luhmann, 1990, esp. pp. 245–251). Within some evaluative practices this reduction is made impossible by sending anonymized texts (manuscripts, proposals) to the peers. So, useful pieces of information are given up. In other evaluations, like those of the Netherlands Research Council, the reputation of the applicants is explicitly taken into account. In these evaluations reputation is regarded as an indicator of the capacity to conduct the research proposed, and, thus, of the chance on a successful investment (Rip, 1985). One can think of other reductions, such as central research topics – does the work to be evaluated contribute to these? –, outcomes of previous discussions in the field – are these outcomes used as "shoulders to stand on" – and bibliometric indicators.

Conceptualizing an evaluation process as ascribing quality by relating the evaluated object to a frame of reference, brings up the question what frame of reference should be used. A first acceptable answer is the field of research to which the evaluated object is implied to contribute. Within different fields of research different evaluation criteria can be dominant. And in some fields of research it might even not clear what is good for the field: when the ways to handle research are well articulated and shared it is easier for reviewers to form a judgment than when there is uncertainty about tasks and methods (Whitley, 1984). Scientists (and thus, peers) might have different perceptions of their field and the quality of a contributions is inevitable uncertain. The implications for the evaluation multidisciplinary research are directly clear: it might be more worthwhile to get differences in evaluations then try to get convergence (Porter and Rossini (1985). But also the problems to transfer evaluation practices from one context to another: what is an appropriate procedure in one field of research, might not be appropriate in another.

The relation of the peer with the reference, the field is also relevant. Review can be done at different levels: a manuscript or research proposal in the own research area is treated differently from manuscripts and proposals outside it. In the latter case, the reviewer falls back on his/her general understanding of the speciality or discipline, and often focuses on craft and general methodology aspects, rather than on the substance of the research. When the reviewer is cognitively too close to the object to be reviewed the review will be different (another reference is used) from the case of a reviewer more at distance. The received opinion is that the review should be done by persons fully familiar with the domain. This assumes that you can only avoid errors by going *up* in level of detailness of expertise. Travis and Collins (1991) on the contrary give an example of a situation in which this policy led to problems. The proposal was on a topic of a very special area with researchers all knowing and backing each other. Reviews were made only by researchers in this small domain (other researchers found themselves not competent!) and hence were found to be "cognitively particular". So going *down* in expertise might be a good policy as well.

The context of peer review

Once we enlarge the focus only slightly, it becomes obvious that peer review, at first glance appearing a strictly insiders business -scientists judging scientists-, is anchored to the external world. In all the contexts, the evaluations are made for specific purposes. They are meant to be used as information in a decision process: decisions on publication of a manuscript, funding of a research proposal, — and if we look at new evaluation processes — on stimulation of a scientific field, on allocation of resources in a research field, on regulations of environmental, industrial and health risks.

The configuration in which these kind of

decisions are made have a recurrent structure. The basics of this configuration can be found already in the founding of the *Philosophical Transactions.* In 1665 Henry Oldenburg, the then secretary of the Royal Society, took the initiative to make public, and thus delocalize, what was already done bilaterally and in closed meetings by the new philosophers (scientists we call them now with hindsight): exchange results, claims, opinions, experiments, theories, etc. He convinced the Council of the Royal Society of the value of publishing these exchange. Oldenburg himself became the first editor, a role which was less articulated than it is nowadays. With this decision the Royal Society removed at least partly some of its central functions, reporting, experimenting, discussing and legitimating knowledge claims, from the meetings of the Royal Society to this new forum, but it still kept them under its authority. In order to keep control, the Council linked a condition to the delegation of these tasks, to the provision of journal space and, most of all, to the use of its reputation: manuscripts being published had to be first reviewed by some of the members of the Council (Zuckerman and Merton, 1971).

The peer review system of the *Philosophical Transactions* is important because it is an example of the configuration of actors in peer review.[3] With the decision to link a referee system to the *Philosophical Transactions* the Council set up a configuration between itself as publisher, its secretary Oldenburg as editor, its members as peers, and its members and other scientists as possible authors. The configuration returns in most of the present day peer review processes and the positions can in general be labeled as "patron", "intermediary", "peer" and "science".

Analytically, two relations are at the base of the configuration: exchange and delegation. The very basic relation is between the scientists producing scientific knowledge and an actor willing to provide resources: within manuscript peer review processes the publisher, with proposal peer review the state. They exchange goods and mutually profit from the relation. As these patrons of science could effectively function in this relation, peer review would not be necessary. However, for reasons of effectiveness or (perceived) competency patrons delegate part of the discretion on the resources to intermediate bodies like editors and research councils. They decide effectively on the allocation of the resources, being able to pursue own goals. What results is a typical control relation in which one actor transfers resources to another in the expectation that these actors will act his behalf. Within this relation peer review functions as a source of trust in and legitimation for the decisions of the intermediate bodies. The editor and the research council delegate one element of the decision process, the review, to (other) scientists, the peers. Most improvements of peer review processes can be understood as rebalancing some of the discretion over ultimate decisions and improve the trust in the outcomes.

The complexity of the configuration also results from the relations between the authors and the peers. One important element of journal peer review, which holds true for proposal peer review as well, is the latent "circulation" aspect: while I review X, and X reviews Y, Y will review Z, and she may well review me. Theoretically, this circularity might lead to a prisoner's dilemma. In an attempt to get a higher reputation himself one peer might give colleagues less reward than they deserve. If the others stick to the fair rules of the game he might profit from such a strategy, at least in the short run. However, if all peers adopt this strategy, the system will collapse and no one profits.

Despite this dilemma, most of the configurations happen to be stable. The alignment between the peers and the authors is due to the specific relations between scientists: they exchange goods as well. In their article on the emergence of the scientific journal with an institutionalized evaluation practice, Zuckerman and Merton (1971) have shown the importance of the triple role of the scientists involved: members of the Royal Society, contributors to the *Transactions* and read-

ers of it, i.e. consumers of the produced work. Especially the role of consumer is important for the stability of peer review processes.

Researchers, and hence peers, depend on the research products of each other (see also Bourdieu, 1975). Thus, they have an interest in spending time and effort on peer review; a strategy of rejecting good work is tantamount to cooking your own goose. The implication of this is that long-term self-interest can be combined with quality judgments at the disciplinary level.[4]

Stabilisation of peer review processes

If evaluation processes are repeated the different positions within the configuration become more aligned and routines about the frames of reference and the reductions of complexity that can be made are developing. In established evaluation practices evaluators are rarely asked to explicate which reductions they have made. Standards how to operate have sedimented in what can be called an *evaluative repertoire.* Evaluative repertoires facilitate the functioning of evaluation practices. They provide the means to which an evaluator can resort and the use of them legitimates the outcomes of the evaluation.

The concept of 'repertoire' is derived from sociolinguistics, in which *repertoire* is used as a concept for *ordered variability* in language uses of specific social groups or in specific social contexts. Within science studies it was introduced by Gilbert and Mulkay (1984) in their analysis of scientific discourses. They distinguished two interpretative repertoires, an empiricist one and a contingent one, in order to *"understand how scientists, as they reproduce different kinds of contexts within the social world of science through the use of different linguistic registers [or repertoires, bm]".*

Gilbert and Mulkay (1984) link the use of a certain repertoire to the competence that is ascribed to the user of the repertoire. Using empiricists and contingent repertoires at the right moment makes the actions of a scientist to be considered as professional, scientific and inevitable. So, repertoires provide evaluators with social resources. This competence is more than the ability to evaluate in a certain way: to form disciplinary judgments after reading a manuscript or to aggregate and interpret bibliometric data validly. Competence is also the ability to use a repertoire in the right context.[5]

There is construction of competence and peers at the level of each evaluation. Editors of scientific journals go to some length to find the "right" peers (if the field is large enough, algorithms for selection are sometimes developed). Research council staff pick and choose the peers that have to judge the proposals. Evaluators of institutes and the R&D programmes are invited on the basis of some relationship with the sponsor who commissions the evaluation. Selection of peers is an important instrument of program directors to manage the evaluation processes. In this sense already, peers are socially constructed: there is a process of search, selection and designation as peers.

Travis and Collins (1991) give some good examples of construction of peers and competence among researchers. In committee sessions of the SERC in which reviewed proposals were graded, reviews and peers came in for evaluation themselves (*meta* review). The qualifications of the peers were used to balance their reports and aggregate them into a final judgment: *"[D] is a hard referee," "[H] is the hardest referee I have ever come across."* In some discussions the choice of the peers on a certain proposal was questioned (and so were the reviews), as being limited to an in-crowd. The committee questioned also positive qualifications of a proposal, pointing to the fact that the three peers were from a same small specific field sharing with the proposal a controversial position. It was supposed that the good qualification was due to the choice of the reference by the evaluators.

Evaluation processes are mostly recurrent and, hence, the temporal constructions become stabilized and thus can be recon-

structed in other evaluations without much resistance. Once peers and competence are constructed and designated to one or more fields, the construction of peers in specific instances, i.e. the selection of peers, is not any more a matter of pure technical competence to evaluate, but of social authority and expertise. In other words the constructions and their actualization in concrete evaluations provide certain actors with a *social capital* (Bourdieu, 1975) that can be introduced and used in new evaluations.

The social construction of peers takes place against a backdrop of societal construction of peers. Peer review has come along with the professionalization of science with an emphasis on reputational control, rather than client or state control. Peer review can add to the autonomy of fields. Autonomy of a field implies then that scientists of this field are in advance labeled as peers, that their notion of good evaluation, and thus of good science, is dominant and that consequently their evaluations are regarded as good evaluations. The boundaries of peership and evaluation competence coincide with the boundaries of the field. Or, to be more precise, the boundaries of the field are reinforced now they not only designate members and non-members of the field but also peers and non-peers, competence and non-competence.

Differences in peer review processes

It is important to note that the struggles about who and what are competent peers can result in differences of roles and relations within the basic configuration. For instance, there are disciplinary differences in the way editors select peers. Gordon (1980) in a study on the evaluation of research papers quotes editors of medical journals who stress the authorial position of the peers. *"A referee must be a specialist in his field, recognized and accepted as an expert"*, and *"A referee must be someone who is regarded as an authority in the specialist subject of the paper"* (p.55). In mathematics editors try to combine evaluation competences of different peers. Young ones are asked for technical examination and older ones for the assessments of the importance of the piece of work.

Of interest are also the differences Gordon finds of strategies of peer selections by natural sciences journals and by those in philosophy and social sciences. For the former the peers are those working on the same subject or in the same field, no additional criteria are used.[6] For the latter peers also have to be able to make balanced judgments and to examine the interest of the paper for both the author's fellow researchers in the specialism and the broader journal readership.

> *"For the philosophy and social science journals] this role finds its rationale within a perspective viewing journal publication as a complement to monograph publication in the formal dissemination of research findings, and also as providing a reader-orientated current awareness function. This contrasts with the natural science and mathematical editors' depiction of the roles of their journals, which views them as considerably more author-orientated, with archival and priority recognition functions given greater weight." (Gordon, 1980, p. 66)*

The interesting point of this finding is that who is regarded as a peer is related to the specific communication role of the journals, hence, to the broader context of the evaluation.

Differences occur not only in the relations between 'intermediary' and 'peer', but also in their relation to the field of research. For instance, in some disciplines like psychology and behavioural science, the quality of the peer review processes has been subject of various experimental studies. The discussions seems not due to wrong use of peer review by psychological editors or to larger incompetence of psychologists to judge what is good science and what not. The studies seem to be part of a stabilized situation in which he editors and peers are considered

as representatives of the field and accountability is effected in a relatively strong monitoring of the evaluations and hence the editorial processes.

Differences between peer review procedures are larger when we look at peer review of research proposals. In the USA the National Advisory Cancer Council was the first council who asked individuals called peers to evaluate proposals. In 1937, it was *"was authorized to review applications for research funding and to certify approval to the Surgeon General for projects that had the potential of significantly contributing to knowledge about cancer."* (Quoted in Chubin and Hackett, 1990). Roy (1985) reports that in the late 1940s, at the Office of Naval Research, which was the first systematically organized government source of research funds for universities, peer review began as an informal "seeking of a second opinion" *among* the grant managers. With one of the ONR grant managers this practice moved to the National Science Foundation (NSF) where it was formalized. Nowadays funding agencies and research councils use a variety of mechanisms called peer review, asking scientists to judge individually or in committee sessions, to judge the proposals separately or ranking them, review in one or two rounds with possible lay juries, etc., etc.

Some funding agencies, like the Netherlands Research Councils and the former Science and Engineering Research Council (SERC) in the UK operate like publishers, i.e. leaving the discretion over the selection to representatives of the respective fields. Other funding agencies have clearly articulate own goals and peers are expected not only to identify with their own disciplinary field but also with the policy of the funding agency: not only the quality of the proposal with respect to the field is at stake but also the merit of the research proposed with respect to the goals set by the funding agencies.

It has been noticed that at funding agencies knowledge is built up about the judgmental behaviour of peers and officials can use this knowledge to steer the evaluation process. At the NSF program directors have explicitly the position as one of the evaluators and the final decision maker. (National Science Foundation, 1991) The procedure of the Dutch Technology Foundation (STW) with its monitoring of the judgment behaviour of the peers is another example of an active role of a funding agency (Van den Beemt and Le Pair, 1991).

The status of a peer within these selection processes has become ambivalent, and the space for funding agencies and their program directors, for discretion over the selection process has increased. Consequently, in these peer review configurations, boundaries between peers and non peers, competence and non-competence shift with reference to popular image of peer review and autonomy might be weakened.

New evaluation processes

So far, I have gone from the micro level of how peers evaluate up to meso analyses of individual and recurrent peer review processes. For going to the level of interactions between different evaluation processes and their position in the research system the focus on traditional peer review does not hold any more. We need to take into account new evaluation processes.

The introduction of new evaluation processes in the relations of science with especially the state, but also other sponsors and society, is one of the main features of the political transformation of the research system (Ziman, 1983; Rip, 1988; Cozzens *et al.*, 1990). In the eighties governments became eager to get 'value for money' and pressed scientists and research organisations to articulate their performances. Research budgets became more stringent and competition about resources increased. In addition direct linkages with industry and other *research users* were stimulated. Scientists, universities, research organisations have by and large adopted the quest for quality and initiated evaluation processes themselves. Evaluations have been initiated of research pro-

grammes, institutes, disciplines and of researchers. These evaluations take place for different reasons and in different policy contexts. The net result is a research system in which quality control is realized in an conglomeration of informal and formal evaluation processes.

We can analyze the development and stabilisation of these evaluation processes within the framework, developed above. The basic configuration of patron, intermediate body, peer, frame of reference and evaluated object is still visible. Often the new evaluation processes are labeled as peer review as well. Within the evaluation processes patrons and intermediate bodies call in scientists to use their *social capital* as peers and legitimate the evaluation processes.

There is a kind of naive expectation that by adopting the peers in the evaluation process, the stabilization of traditional peer review processes is adopted as well. The expectation is naive as those initiating the evaluation processes often appear to be much more active within the process itself as traditional patrons. Within these new evaluations more then ever patrons seek to have the pleasures of the expertise and authority of peers without the pain of losing the control over the decision process.

In the eighties, when new evaluations took place in the context of budget constraints, instead of stability, discussions arose on the validity of certain evaluation processes. The resulting struggles and negotiations have resulted in adaption of peer review processes to the new needs, as well as the emergence of new evaluative repertoires. The most profound changes in evaluation practices are the emergence of evaluation tools like bibliometric methods and the development of evaluation protocols to which peers have to conform.

The discussions on bibliometric methods both in the scientometric literature and among scientists and science policy makers, the successful and unsuccessful uses within evaluation, the changes in legitimations for bibliometric methods within evaluation reports are an excellent example of the stabilisation process of an evaluative repertoire. Presently bibliometric methods are acceptable in many evaluation processes. The basic rhetoric behind bibliometric indicators is that these evaluation tools help the peers to make the impartial and valid judgments needed for such decisions as on allocation of resources.

The intriguing paradox is then that although the peers are called in the evaluations because of their competence to evaluate, they are confronted with less trust in their individual evaluation competency then ever. The improvement of judgments might be a good reason to introduce bibliometric methods and protocols in the process, but at this point we should not forget the social aspect of judgments. Bibliometric tools are especially attractive for those in the evaluation process who do not evaluate. The bibliometric methods provide them with the tools to control the peers: to have them evaluating according to the aims of the evaluation. (Van der Meulen, 1992)

Another important change with regard to the traditional peer reviews is that more and more non-scientists enter the configurations and take part in the circulation. In manuscript peer review the position of editor (intermediator), peer and evaluation object belong to scientific actors. In proposal peer review research council intermediates might be staff members of research council – but if they are not (recently been) scientists themselves they mostly take up modest roles —, and sometimes *lay juries* are called in for balancing peer reviewed proposals. In new evaluation processes the intermediate position is often occupied by more active research policy makers, having opinions on the quality of science themselves. In some cases, like disciplinary evaluations, the ministries organize the evaluation themselves and seek peers directly. Note that in these evaluation processes society is often defined as a new patron for which the government mediates.

As to the position of the peer, non-scientists might be called in even in the first round. Especially, *research users* are considered to be legitimate evaluators in evaluation proc-

esses were social relevance of research is at stake. They are added to evaluation committees to balance against science-centred evaluations or might even dominate the evaluation process. A recent example is the evaluation of the German large research institutes. The strong focus in present UK science policy on research use will likely have that effect as well. Remark that 'research users' is a construction as much or even more as 'peers'. The competence of the non scientists is often linked to societal and industrial prestige rather then expertise with using research.

Also evaluation experts are coming up especially within the evaluation of research programmes. They actively go for articulating good evaluation processes and stress the importance and difficulties of balanced judgments —a task not easily to be done my amateur evaluators, that is the peers. Even the position of 'evaluation object' is not untouchable any more as management, again also of staff members, is an integral part of some evaluation processes.

In most evaluation processes scientists are still used as peers and actors involved in the evaluation tend to speak of peer review. But evaluation competence is not solely ascribed to members of the field any more. Others are called in as competent reviewer. Patrons or intermediaries come in with *meta*-review. They claim competence on how the evaluation should be conducted and try to impose their notions of good science on the evaluation. Decision makers build up the formal right and the social legitimation to deviate from the evaluative advices of the peers.

In conclusion: research evaluations in transition

Considering the conglomeration of evaluation processes it is hardly possible to give a definitive account of their position in the research system. Within studies on peer review one can still speak about *the* quality of the research product. Although quality is a relative concept routines have developed up to point that one can easily black boxing the social construction of quality. On the level of the research system however, evaluation processes are so diverse that one has to think in terms of different qualities. The question is then how good quality control can be maintained and what new stabilization takes place? The framework developed does not provide definitive answers for what the end result should be. But it identifies factors that are important in the development and stabilization of evaluation processes. These are:

- the exchange of resources and delegation of tasks within the configuration of the evaluation,
- the division of discretion within the configuration,
- the ascription of evaluation competence,
- the development of evaluative repertoires,
- the production of trust in reliability, fairness etc. of the actual outcomes.

With these in mind we can at least analyze which evaluations are still stable and will resist changes and which are sensitive to (further) changes. With regard to new evaluation practices, critical factors for the stabilization can be identified as well as driving forces for ongoing transitions.

Traditional peer review has a strong position in the research system and is likely to remain a reference for thinking about quality control. Especially the configuration of manuscript peer review is extremely stable. Criticism of its reliability and fairness can be heard, but have had little effects. Proposal peer review is more sensitive to changes but there are some stabilizing factors. Peer review processes are crucial for acquisition of resources, also those not directly linked to the review processes, like reputation and credibility. The resources directly obtained from peer review can be converted into scientific results and hence prestige and status. The *credibility cycle* of Latour and Woolgar (1979) still works.

There is another mechanism that makes peer review to have a strong position. Having peer reviewed publications and research funds is becoming an indicator of quality as

such and increases the reputation. The outcomes of traditional peer review processes are reproduced in other evaluations, both formalized and non-formalized. Bibliometric indicators are the evident example. Having projects within government funded research programmes might be a signal of quality as well, especially if the programme dominates the field of research.

But the transformation is still going on. Some new evaluative practices have stabilized, like the use of bibliometric methods and evaluation of research institutes by visitation of peer committees. These practices rely on traditional peer review and reproduce existing relations. Bibliometric indicators can in principle destabilize relations and induce new ones, but as long as the dominant ideology is that only peers can interpret the results, they are unlikely to do.

What seems to become a driving force for changes is the circulation of non-scientists, esp. research users, evaluation experts and research policy makers within the evaluation processes. Management of proposal peer review and the evaluation by research users are accepted practices. The question is to what extent these new actors are willing and able to put into the evaluation processes their own frames of reference and notions of quality. For some evaluation tasks research users can be called in, nowadays. But their is still little experience with handling their judgments. If stabilization occurs, it might affect existing peer review processes and induce further changes.

The crucial factor seems to be competence building and the development of evaluative repertoires. A decade ago an attempt by the Dutch government to have university research programmes evaluated on their social relevance failed (Blume and Spaapen 1988). But these programmes were mainly judged by scientists and these evaluation processes are replaced by other ones. The evaluation of pre competitive research programmes has give a new impetus to the issue and other methods have been developed, especially at the European level. Presently in several countries there are research and policy programmes to develop sound methods to develop methods for the evaluation of societal quality.

Another issue is the coupling of evaluations to foresight. One can presume that when foresight gets stabilized, patrons want to introduce the outcomes in the evaluations as part of the frame of reference. In that case evaluators might be expected to evaluate research on its contribution to the expected *future* of a field of research, rather then to the perceived state of the art. If users dominate the foresight process they might dominate these new evaluations as well.

Definitely, research evaluations are in transition as much as the research system. The most profound change occurring at the moment seems to be the introduction of criteria of user relevance and societal value of research within evaluation processes. The framework developed suggest we should not have too high expectations of possibilities to create such new evaluation processes without inaccuracies, disputes and resistance. Good evaluations are the result of evaluation processes in which roles and relations of those involved are well balanced with the evaluative repertoires.

Stabilization is a result of circulations of actors to positions and within processes, but also of an increase of accepted evaluations that can be used as reference and legitimation, and of growing trust between those involved in an evaluation. Theoretical and empirical understanding of what is going on within the evaluation process and its context, and understanding of the values and limits of evaluation tools contribute to stabilization as well.

NOTES

1. Especially in the USA, where science funding is seen as a matter of fair distribution among a group of equally eligible recipients, there is extensive pre-performance evaluation and extensive discussion about the fairness of peer review. The studies of Cole et al., which suggested bias of the peers towards institutions of high reputation and a high number of unreliable outcomes of the peer reviews, have played

a key role in this discussion (Cole, Rubin and Cole, 1978; Cole, Cole and Simon, 1981). Beside being praised for the insight they give in proposal peer review, these studies have been critized for the too easy conclusions on the fallibility of the peer review process (Harnad, 1985).
An informative overview of the NSF debate is given in Chubin and Hackett (1990).

2. The conceptualization of evaluation as linking the object to a frame of reference opens the possibility to a cognitive theory of evaluative performance. Stepping stones for such a theory are already given in cognitivist studies of science. In dialogue with sociological studies of 'science in the making', De Mey (1982) and Giere (1988) have made studies on how scientists adopt new results and relate them to their existing frame of reference. This has some evaluative component and it would be interesting to extend this kind of studies to the domain of evaluation practices. Tijssen (1992, especially chapter 12) has made an attempt to articulate mental maps or frames of references of scientists by using rating data of interviewees and multi-dimensional scaling techniques.
3. One should be cautious to consider the peer review system of the *Philosophical Transactions* as the first example of modern science's peer review, born fully armed from the head of Zeus, the Royal Society. The idea of science being a separate field of activities has came up only recently and it is significant that in most accounts of peer review with a historical review a big leap is made through history from the first days of the Royal Society to post WW II science. Taking such steps it is readily forgotten that most members of the Royal Society where certainly not only scientists, and that activities in politics, religion and economy were intrinsically intertwined with the new philosophy (see for instance: Shapin and Schaffer, 1985).
4. The emphasis on exchange relations between either a principal and an agent or a consumer and a producer, makes clear that collegial control in science is to a large extent a side product of delegated principal control (as peer) and of consumer/co-producer control. Compare this with for instance the medical professions. There is also control by colleagues, but these are just colleagues. The patients are the consumers. This is the main reason why collegial control in science is much stronger than in other professions. This explains why reputation is so important in science and is a strong proxy measure for quality: consumer control results in reputation differences. If consumer control lacks, professions can maintain to the outside the idea that all practitioners are of the same quality (Johnson, 1972). In science the colleagues are consumers as well and hence collegial control leads to reputation differences (and thus can better be viewed as consumer control).
5. This use of the concept of *repertoire* links up to the sociological critique of Bourdieu (1977) on linguistics. He wants to replace the notion of specific linguistic competence by the notion of symbolic capital connected to the use of certain repertoires (registers, languages, dialects).
6. However, numbers on the the status ranks of referees and authors of the Physical Review provided by Zuckerman and Merton (1971) make clear that there is a preference for peers with higher ranks.

REFERENCES

Blume, S.S. and Sinclair, R.
1973 "Chemists in British universities: a study of the reward system in science" American Sociological Review 38:126–138

Blume, S.S. and Spaapen, J.B.
1988 "External assessment and "conditional financing" of research in Dutch universities." Minerva 26: 1–30.

Bourdieu, P.
1975 "The specificity of the scientific field and the social conditions for the progress of reason", Social Science Information, 14, 6: 19–47.

Bourdieu, P.
1977 "The Economics of Linguistic Exchanges", Social Science Information, 16: 645–668.

Chubin, D.E. and Hackett, E.J.
1990 Peerless Science: peer review and U.S. science policy, Albany: State University of New York Press, 1990.

Cole, S., Cole, J.R. and Simon, G.A.,
1981 "Chance and consensus in peer review" Science, 214: 881–886.

Cole, S., Rubin L., and Cole, J.R.
1977 "Peer review and the support of science", Scientific American 237: 34–44.

Cozzens, S.E.
1990 "Autonomy and Power in Science." Pp 164–184 in S.E. Cozzens and T.F. Gieryn (eds) Theories of Science in Society, Bloomington: Indiana University Press.

Cozzens, S.E., Healey, P., Rip, A. and Ziman J. (eds.)
1990 "The Research System in Transition" Dordrecht: Kluwer Academic Publishers.

Crane, D.
1972 Invisible Colleges, Chicago: The University of Chicago Press.

Gaston, J.
1968 "The Reward System in British Science" American Sociological Review, 33: 718–732.

Gilbert G.N. and Mulkay, M.
1984 Opening Pandora's Box: a sociological analysis of scientists' discourse, Cambridge: Cambridge University Press.

Giere, R.
1988 Explaining Science, Chicago: The University of Chicago Press.

Gordon, M.D.
1980 "The role of referees in scientific communications" pp.263–375 in J. Hartley (ed.) The psychology of Written Communication: Selected Readings. London: Kogan Page.

Hagstrom, W.O.
1965 The Scientific Community, New York: Basic Books.
Harnad, S.
1982 Peer Commentary on Peer Review: a case study in scientific quality control, Cambridge: Cambridge University Press.

Johnson, T.J.
1972 Professions and Power, London and Basingstoke: The MacMillan Press Ltd.

Latour, B., and Woolgar S.
1979 Laboratory life. Beverly Hills: Sage.

Luhmann, N.
1990 Der Wissenschaft der Gesellschaft, Frankfurt a. M.: Suhrkamp Verlag.

Mahoney, M.J.
1977 "Publication Prejudices: An experimental study of confirmatory bias in the peer review system" Cognitive Therapy and Research, 1: 161–175.

Merton, R.K.
1973 The sociology of science, Chicago: The University of Chicago Press.

Mey, M. de,
1982 The cognitive paradigm, Dordrecht: Reidel.

National Science Foundation
1991 Grants Policy Manual, nsf8847a, electronic copy.

Peters, D.P. and Ceci S.J.
1982 "Peer review practices of psychological journals: The fate of published articles, submitted again" The Behavioural & Brain Sciences, 5: 187–255.

Porter, A.L. and Rossini, F.A.
1985 "Peer review of interdisciplinary research proposals" Science, Technology & Human Values, 10: 33–38.

Rip, A.
1985 "Commentary: Peer review is alive and well in the United States", Science, Technology & Human Values, 10: 82–86.

Rip, A.
1988 Contextual transformations in contemporary science, Pp 59–85 in: A. Jamison (ed.) Keeping science straight: a critical look at the assessment of science and technology, Department of Theory of Science, Göteborg, Report nr. 156.

Roy, R.
1985 "Funding science: The real defects of peer review and an alternative to it." Science, Technology & Human Values, 10: 73–81
Shapin, S. and Schaffer, S.
1985 Leviathan and the Air-pump: Hobbes, Boyle and the Experimental Life, Princeton University Press.

Shapiro, S.P.
1987 "The social control of impersonal trust, American Journal of Sociology, 93: 623–658.

Tijssen, R.J.W.
1992 Cartography of science: scientometric mapping with multidimensional scaling methods, Leiden: DSWO Press.

Travis, G.D.L. and Collins, H.M.
1991 "New light on old boys: cognitive and institutional particularism in the peer review system", Science, Technology & Human Values, 16 (3): 322–341.

Van den Beemt, F. and Le Pair, C.
1991 "Grading the grain: consistent evaluation of research proposals", Research Evaluation, 1: 3–10.

Van der Meulen, B.J.R.,
1992 "Indicators in a framework of judgment and control", Pp. 57–74 Weingart, P., Sehringer, R. & Winterhager, M. (eds.) Representation of Science & Technology, Proceedings of the Conference on Science and Technology Indicators, Bielefeld 1990, Leiden: DSWO Press.

Whitley, R.D.
1984 The intellectual and social organization of the sciences. Oxford: Clarendon.

Ziman, J.
1983 "The collectivization of science", Proceedings of the Royal Society B, 219: 1–19.

Zuckerman, H.
1967 "Nobel laureates in science, patterns of productivity, collaboration and authorship", American Sociological Review, 32: 391–403.

Zuckerman, H. and Merton, R.K.
1971 "Patterns of Evaluation in Science: Institutionalization, Structure and Function of the Referee System", Minerva, 9: 66–100.

Barend J.R. van der Meulen
Centre for Studies on Science, Technology and Society,

University of Twente,
P.O. Box 217
7500 AE Enschede
The Netherlands