Promises, Problems and Pitfalls of Peer Review

The Use of Peer Review in External Quality Assessment in Higher Education

D.F. Westerheijden

Abstract

In the discussions that raged since about 1981 on how to assess the quality of research and/or teaching in higher education, two broad strands of methods have been advocated: performance indicators and peer review. After thorough discussions of the merits of performance indicators, and after some years of experience with quality assessment procedures in several European countries, a consensus now seems to be growing that performance indicators should be the 'objective base' for assessment procedures, but that they should be used 'intelligently' in processes of assessment by human, knowledgeable individuals, for short: peer review. The merits of peer review have not received that much attention in the public discussions on quality assessment methods, even though a not insignificant body of research has been built up in the discipline of sociology of science.

The aim of this paper is to review some of the research into the actual operation of peer review, and to try and apply the research results to the 'new', external quality assessment procedures, where peers are in a quite different position from journal referees and the like, and where they face a different task.

§ 1 Introduction: Promises

British higher education first came into contact with external quality assessment procedures in a more dramatic way than anything experienced by other countries. Almost overnight, the University Grants Committee (UGC) in 1981 introduced substantial budget reductions to the universities by way of a 'highly selective approach', presumably based on teaching quality of the institutions (Sizer 1990: 156). Indeed 'presumably', for the criteria used remained unknown. This brought to the fore —as one of its more innocent consequences— the discussion on the question of how quality of higher education should be assessed: objectively or subjectively? By way of *performance indicators* or by way of *peer review*? In reaction to the apparently subjective approach used by the UGC, support grew for objective performance indicators.* Long lists of possible or recommended performance indicators have been published (see for an overview, e.g., Cave, Hanney, Kogan & Trevett 1988). In contrast to this development, the external quality assessment procedures initiated in the 1980s in the higher education system of the Netherlands relied to a large extent on subjective human judgement. The discussion in the Netherlands about performance indicators did lead to research (Segers, Dochy & Wijnen 1989, see also Dochy, Segers & Wijnen 1990), but not to policy changes. The contrast between the British and Dutch approaches to quality assessment in higher education should not be exaggerated, it is more a question of emphasis; the British do not rely solely on performance indicators, nor do the Dutch use peer review exclusively. In both countries quality assessment consists of human judgement based on —but more than a simple combination of— objective data.

The performance indicator side of this has been rather extensively discussed in the years since 1981. Peer review as a method is not often discussed; usually, it is taken for granted. For example, Banta & Fisher (1989: 6) call it '. . . a time-honored evaluative process that is almost

---

* Mixed up with this was the question on public, published *versus* 'secret', unpublished assessments. In principle, the publicity of statements is not related to the method of reaching these statements. Therefore, this question will not be addressed systematically in this paper.

universally accepted'. Awareness about the method of peer review with both decision makers and researchers does not seem to progress much further than that it is a subjective matter: quality assessors should use the objective base of performance indicators in an 'intelligent' way, supplementing the diverse, necessarily fragmented information of the performance indicators with a holistic view of an institution's (or: a department's) quality. Peer review, in that view, *promises* to deal with any 'hole' in the measurement of quality left by performance indicators. In this paper, I aim to address the question whether peer review *can* do that. In other words: *What are the limitations on the reliability and validity of peer review in external quality assessment procedures?*

In order to do so, in the next section I propose to present a theoretical model of peer review, highlighting the position of the reviewers in the 'production process' of science.* I shall then, in § 3, illustrate the working of this theoretical model in 'classical' peer review situations with the aid of empirical research of peer review. Finally (§ 4), a theoretical analysis, informed by the empirical research results, will be given of the problems and pitfalls besetting quality assessors in the external quality assessment procedures developed in the 1980s.

A Note on Some Key Concepts

As we are not interested in concepts in a vacuum, but in why things are as they are in the world we know, the relationships among concepts are more important than explicit definitions of the concepts. As will become clear in this paper, for example, 'peer review' in external quality assessment procedures is in many respects different from 'peer review' in the classical sense. Therefore, I shall not devote too much attention to essentialist definitions of what a concept 'really' is. Some attention to what I mean by certain terms is, however, appropriate here.

For the purpose of this paper, 'peer review' is any method of judgement of (a portion of) someone's work by one or more other individuals who are supposed to be knowledgeable about that field of work, usually from working in the same field, and that relies solely or predominantly on the judge's (or judges') statements. Some well-known forms of peer review are: refereeing of manuscripts for scholarly journals, assessment of doctoral dissertations, review of proposals for research project grants, and peer review of persons in decisions regarding tenure or promotion, or awards. These are the forms that have existed for quite some time, hence I shall call them 'classical' peer review, in contradistinction to the 'new' peer review, i.e., the judgement of the quality of study or research programmes, departments or institutions by way of external assessment procedures. The new quality assessment procedures are 'external', especially in the sense that the initiative for these procedures usually is not taken by the institutions concerned. Although voluntary accreditation organizations certainly do exist in the United States, and some European faculties have joined, the impetus for the nation-wide quality assessment procedures that have arisen in the 1980s in Europe, e.g., in Great Britain and the Netherlands came from the respective governments —not from the higher education institutions themselves (see, among others, Goedegebuure, Maassen & Westerheijden (eds.) 1990).

'Performance indicators' are quantitative and/or qualitative empirical data that describe the extent to which an actor accomplishes his goals (based on: Dochy, Segers & Wijnen 1990: 136–137).

Another key concept, 'quality', has eluded efforts to define it in a substantive sense (Westerheijden 1990: 184). In current definitions of quality, the proportion of goal attainment is mentioned as the characteristic element. An equally current definition, that for all practical purposes is equivalent, stresses the degree to which the 'product' does what it is intended to do. Also the ISO definition ('The totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs'), can be interpreted in the same way. These

p. 132

---

* In the remainder of this paper, I shall take the term 'science' to mean not only the physical and life sciences, but also what is usually called 'scholarship' (as in the social sciences and the humanities).

definitions link quality with needs or goals (see also: Cave *et al.* 1988: 29; De Weert 1990: 59 *ff*). For that reason, quality has been called a relative concept: it can be defined operationally only in relation to a set of goals. Goals are held by actors, or stakeholders. Hence the quality of higher education will be viewed differently by different stakeholders, the most important of which in this case are the government, funding agencies and other intermediate organizations, the national subfields of relevant disciplines, staff of higher education institutions, and higher education 'clients', i.e., students, employers and research contract partners (see figure 1; see also: Pollitt 1990: 63; Westerheijden & Weusthof 1990).

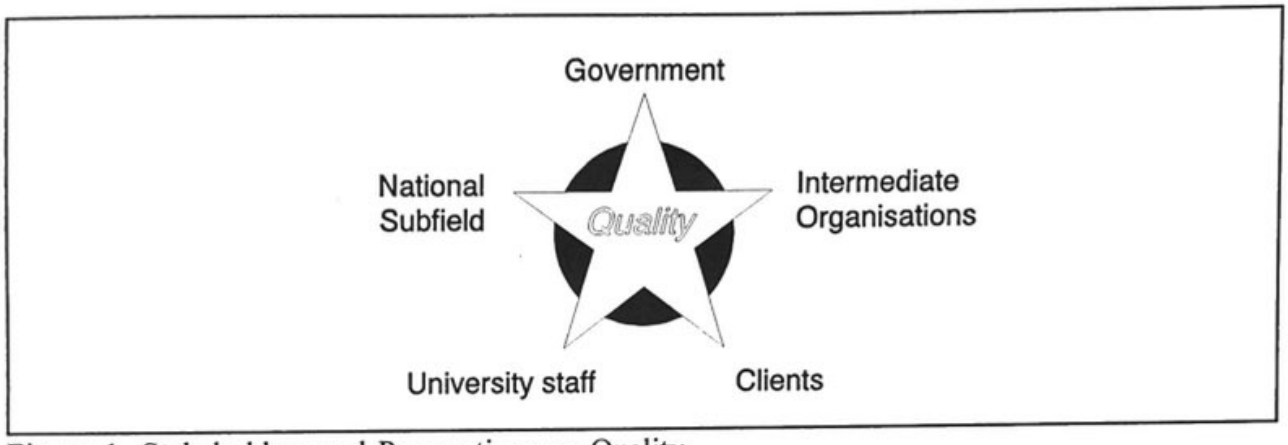

Figure 1 Stakeholders and Perspectives on Quality

Finally, under 'external quality assessment procedures' I shall subsume the procedures introduced in Britain and Holland in the 1980s to ascertain the quality of higher education and/or research on the level of study/research programmes, departments or institutions. For the moment, I do not try to generalize further than that, although I expect that much of the analysis has wider application.

§ II A Model of 'Classical' Peer Review

The Social Organization of Science and the Need for Reputation

Peer review is part of the social organization of science.* It is a mechanism of collegiate control (Johnson 1972), in which scarce resources, such as journal space, research grants, or government money, are distributed in an intentional manner among scientists or teachers. Scientist are dependent on others (e.g., other scientists, government officials, university administrators) for these scarce goods. Dependence of one party, implies influence or power of the other party. How do these relationships of dependence and power operate in the social processes of science?

If science is seen as a production process, its prime products are knowledge-claims. These claims can be recorded in manuscripts, which can be circulated or multiplied to be used by other scientists. Such dissemination is crucial to the discussion of theories, tests, and results which, in its turn, is crucial to the advancement of our knowledge. Two remarks should be made here. First, scientists in this process occupy both the positions of producers and of consumers. This affects their behaviour: sometimes they behave like producers, sometimes like consumers. Second, the process is cyclical (or spiral): knowledge-claims are produced, and then they are input into the next phase of the production process of science, resulting in new or modified knowledge-

p. 133

* The next paragraphs are inspired by, i.a., discussions with Prof. A. Rip; see also Rip 1988.

claims. This cycle has not only an epistemic, but also a social meaning. The social aspect will be elaborated next —the epistemology is not this paper's topic.

The cyclical nature of the social process of research was first emphasized by Latour & Woolgar (1971; see also figure 2). What they stressed in particular, was that 'doing' science resulted in differential opportunities for doing more science. For only if researchers' knowledge-claims become known, can other researchers use these claims. Usefulness of knowledge-claims in producing more innovations is the basis of valuation in the scientific community (Whitley 1984: 12). If the claims are accepted, the researchers gain *credibility* (the *reputation* of being competent scholars), which may lead to their obtaining tenure, more research grants, etc., in short: may lead to more opportunities for them to produce more knowledge-claims. Obtaining a high reputation is, as a consequence, a very important (intermediary) goal for all scientists.

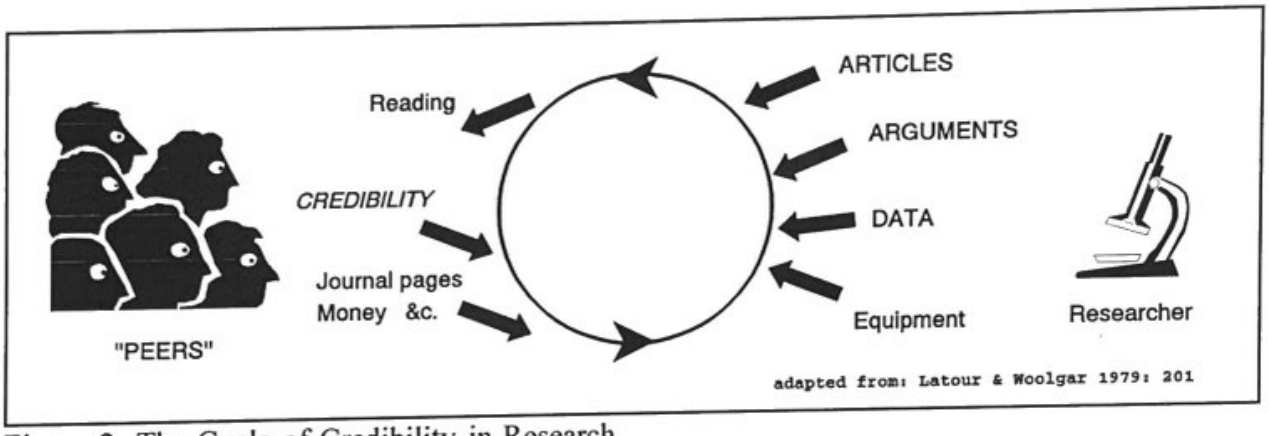

Figure 2 The Cycle of Credibility in Research

The processes by which knowledge-claims come to be known by the consumers are, therefore, of special importance to researchers, especially because becoming known to others is a scarce good. It must be a scarce good, since no one's time is unlimited: one cannot know all about all claims to advances in knowledge during a single life-time. But the scarcity is more evident in other phases: there is scarcity of space in journals. And there is scarcity of financial opportunities to do research, because research requires investment, often of considerable sums, in equipment, laboratory or survey work, etc. It is, accordingly, to be expected that scientists are willing to invest effort, time and other resources into processes that make them known among their consumers —especially those consumers that have other scarce resources to offer in return. The most generally important process in this respect is the publication of articles (or, in some disciplines, books), because this is most immediately relevant to obtain a high reputation. On the other side of this relationship, the consumers of science use these articles for their own science-production; they have, therefore, an interest in knowing that the articles they use are genuine knowledge, not bogus. This explains the rise of peer review in publishing: it is more efficient to have a control on the quality of articles by a few experts before the articles are published, than to have everything published and every scientist —even those less expert in certain matters— sort out the quality of it all. Commenting on the historical evolution of peer review in the seventeenth century, Zuckerman & Merton (1971: 74) put it this way:

p. 134

> In their capacity of producers of science, individual scientists were concerned with having their work recognised through publication in forms valued by other members in the emerging scientific community who were significant to them. In their capacity as consumers of science, they were concerned with having the work produced by others competently assessed so that they could count on its authenticity.

The latter part of this mechanism has been explained succinctly by Stigler: 'Reputation economizes on search' (1961: 224). This points to the great importance of *reputation* in the social organization of science —a concept I italicized before (page 4). Differences in status among scientists are not visible on the outside: one simply has to *know* that Prof. A *is considered to be* the foremost expert in a certain specialism. Human judgement and social organization are inextricably involved in this, for what counts is not whether Prof. A 'really' is the best, but whether the other members of that particular scientific community believe this person to be so —especially the other high-reputation members (a case of recursion?). This requires communication among those scientists, and a certain degree of consensus as to what counts for obtaining a high reputation. Publication of articles and getting citations are important means to gain reputation. But although reputation is acquired through performance, it then tends to be ascribed to a person for an indefinite period: non-performance does not immediately mean erosion of reputation (Zuckerman & Merton 1971: 81).

From the point of view of obtaining resources for further science production, scientists are not equally interested in gaining a reputation in the eyes of all other scientists: what counts most is one's reputation in the eyes of those who have most influence in the distribution of reputations (and other scarce resources) in the disciplinary field: 'Thus, scientists who seek the highest reputations . . . have to convince *powerful* colleagues . . .' (Whitley 1984: 12; emphasis added). And powerful colleagues, more often than not, are those with a high reputation, for they are the ones the others listen to.

High-reputation scientists are asked more often than low-reputation scientists to perform peer review activities: judge manuscripts, research grant applications, personnel, etc. (see, e.g., Zuckerman & Merton 1971: 87–88). Whether they do this in more or less permanent positions, e.g., as members of a grant-awarding body, or on an *ad hoc* basis, e.g., as a referee for a journal, this means that high-reputation scientists have more chances to distribute scarce resources. In short: higher reputation leads to more power. In this way too, a high reputation becomes a valuable good in itself.

Goal Displacement and the Functioning of Peer Review

Reputation is important for scientists in several ways, as appeared in the previous section. It is not just a by-product of 'producing' science, but it also is a very important intermediate goal that all scientists must strive for —whatever be their ulterior goals— if they want to continue as members of the disciplinary field.

This mechanism produces a kind of *goal displacement* in the scientific process, in that scientists no longer are (exclusively) interested in producing 'good science' (new, corroborated knowledge-claims), but become interested in gaining reputation (also). And unto this new goal, it may be just as effective and efficient to marry the son of a Nobel-prize winner as it is to produce a thorough article. Goal displacement effects can be expected to be the stronger, the higher is the dependence of scientists of the resource allocation process under consideration. What keeps the system deteriorating too much, even under conditions of high dependence on a certain resource allocation process, are the norms and criteria that define a minimum level of acceptability for products of science (articles, etc.) in the discipline. Since these norms and criteria are clearer to all participants and less contested in some disciplines than in others, the effects of goal displacement will be less pronounced in disciplines with a low level of technical and strategic task uncertainty (Whitley 1984: 120 *ff*), such as some natural sciences, than in disciplines with high levels of task uncertainty, such as some social sciences or humaniora.

p. 135

The tension remains, however, in all disciplines, between 'purely' scientific behaviour and 'scoring' behaviour. This applies in the first place to scientist as producers of knowledge-claims in articles, etc. But it has consequences as well for the functioning of peer review. For the peers who review will know that their decisions may have effects for their fellow-scientists: publication

or not, research grant or not, hire or fire, etc. Such effects may already mean an intrusion to the 'open' ('*herrschaftsfreie*') nature of scientifically-based decisions, for these decisions will have consequences for other scientists' probabilities to participate in next rounds of the scientific process. Peers may make use of these effects in a more or less deliberate way, e.g., deny a scientist a research grant because he is a member of another 'school' —the money had better be used for research that is better (i.e., in the tradition of this peer's own school). Moreover, peers may in other situations be in the position that they apply for a grant, or want an article to be published, where they are dependent on other peers, so a mechanism of well-understood self-interest may be at work to refrain from too harsh decisions ('mutual backscratching'). Anyhow, the very real consequences of decisions in peer review processes entail the danger that the peer review processes will incorporate other grounds for decisions than only the scientific quality. In the next section, I shall present some of the findings of research into such questions.

§ III Problems and Pitfalls of 'Classical' Peer Review

In the field of tension that exists for peers described in the previous section, the promises of peer review seem beset with problems and pitfalls. But what happens in fact? How does peer review really operate? In this section I shall give some of the evidence collected in the literature about the functioning of peer review as regards the publication of articles and the distribution of research grants. This may not be conclusive evidence, but enough at the very least, I think, to give an inkling of the various problems and pitfalls peer review encounters even in its most accepted applications.

Manuscript Review

Within the problems peer review encounters in manuscript reviewing as it is practised for publication in journals I would like to distinguish three categories: bias associated with the intellectual organization of the discipline, bias associated with the social organization of the discipline, and (random) error. The former two categories call the validity of peer review into question, while the latter calls the reliability into question.

Intellectual Bias

Intellectual bias would occur if systematic relationships existed between the intellectual organization of a discipline and the acceptance or rejections of papers submitted for publication.

On a high level of aggregation such a phenomenon exists, as seen in the highly different acceptance rates of manuscripts in 'soft' and in 'hard' sciences. As Hargens (1988) illustrates with a comparison of the *American Sociological Review* and the *Physical Review*, immediate acceptance in the sociological journal is ±10%, while it is ±65% in the physics journal. He explains this large difference partly from the decision structure these journals use, but the most important variable appears to be the different degrees of consensus prevalent in the relevant disciplines (Hargens 1988: 149). Such differential publication patterns should, however, not be termed 'bias', because they do not, on this level of aggregation, imply systemic variations in probabilities for publication for scientists in their daily life. If it is usual for sociologists to have a 10% chance of acceptance for their manuscripts, this is a fact *all* sociologists have to live with; it does not favour *some* sociologists over others.

p. 136

What certainly should be called bias, however, is when papers of good 'quality' (whatever that may mean in a given discipline) are rejected because the reviewers did not like the methodology, the results, or the political implications of papers. In disciplines that are cognitively fragmented, this is partly common-place, but with a pluriform set of journals most scientists can find a channel where they will have a fair chance of publication. For example, in the social sciences

separate journals exist for 'schools' that favour qualitative research, or conversely, quantitative research. But less innocuous forms of bias may exist too, that are less visible and which cannot be circumvented by choosing appropriate journals. In particular, research has been published about 'response bias'. 'Response bias' means that referees favour papers showing a particular (favourable) view of a discipline over those that call the discipline into question. Peters & Ceci (1982) find that previously published articles, resubmitted for publication in the same high-reputation psychology journals, stand only an 11% chance of getting accepted (one out of nine; three other journals discovered the ploy). Their results indicate that papers affiliated with high-reputation departments have a higher probability of being reaccepted than those from low-reputation departments. In an article that is controversial from the points of view of methodology, of the author's disposition to draw far-reaching conclusions nevertheless, and of ethics of research, Epstein (1990; see also the comments in the same issue) endeavours to establish the existence of response bias in social work journals. Although he fails to do so (resulting from methodological problems), traces of response bias cannot entirely be ruled out, and he gives devastating, albeit anecdotal evidence of the poor level of peer review quality.

Social Bias

Another source of bias occurs when papers are not selected on the basis of their —to the mind of the referee: pleasing— content, but on the basis of personal characteristics of the author (in relation to characteristics of the referee). In the previous section some results of Peters & Ceci (1982) were mentioned already. They state that such bias does indeed exist. These results run counter to older research results, in particular those of Crane (1967) and Zuckerman & Merton (1971). These researchers note that referees disproportionally are selected from higher strata, either defined as individual reputation (Zuckerman & Merton 1971) or as the standing of the department the referee works in (Crane 1967). Further analysis leads these authors to conclude, however, that such 'elitist' selection of referees does not result in an equally 'elitist' bias in the selection of papers. More signs of social bias are found by Blume & Sinclair (1973: 135–136), who report that rewards distributed by way of peer review are distributed in a skewed way: educational, social and institutional backgrounds do play a role, even though quality of work of the reviewed scientist is the most important single variable. The results of Peters & Ceci (1982) mentioned before do point in the direction of social bias as well.

From the point of view of validity, then, the performance of peer review does not lend itself to unambiguous statements —that in itself is doubtful enough.

Random Error

However good or bad peer review may perform on average —which is more or less what is measured in the research I presented in the previous section, this does not help the individual scientist very much, if the deviation for the average is large. As Roy —a long-time opponent of peer review as it operates now— stated it (1985: 75):

> Statistical studies are carried out on large numbers of proposals to establish that *on the average* the system is fair. One wonders if a plaintiff in a robbery case would be satisfied by a similar statistical argument that, on the average, no one was robbed in New York.

Research shows that, indeed, looking at the average only is not enough: there is a high degree of variance. Quoting an article by Scott, Cicchetti notes that in many previous research projects levels of reviewer consensus had been found that '. . . have been appropriately described as "significantly above chance, though far from substantial" ' (Cicchetti 1980: 300). Marsh & Ball presented an overview of about a decade's research into reliability of reviews. The (unweighted) mean level of agreement in 10 studies of reviewer reliability in the fields of sociology, psychology and education is .27, with a .12 standard deviation (March & Ball 1989: 153). In their own

research, March & Ball obtained a .30 agreement. The authors point out that low agreement among reviewers is not the only factor in the decision-making process leading to the acceptance or rejection of a manuscript. Specifically, the editor makes the final decision (peer review of peer review?) and the manuscripts reviewed by several referees (i.e., the ones included in research) are probably the more problematical ones (March & Ball 1989: 167).

One can also refer back, at this point, to the findings of Peters & Ceci (1982) that only one out of nine times a previously published article was readmitted in the same journal, while three out of twelve discovered that it was a previously published article. And Epstein reports that only two out of 53 journals found out that his 'article' was a plagiarism of a well-cited older article, while four others rejected it because of double submissions (Epstein 1990: 17–18). This may be higher than the chance of getting caught after robbing someone in the streets of New York or Amsterdam, but with so many 'policemen' present (one editor and usually at least one referee per paper), one would expect something better.

Grant proposals

Not only the working of the peer review process surrounding the publication of articles in journals has been researched, but also the functioning of peer review in the distribution of research grants. Using 150 grant applications from the American NSF files (half accepted, half rejected, 50 each from chemical dynamics, economics and solid-state physics), Cole, Cole & Simon find that being rewarded a grant depends to a large extent on chance: when the applications are reassessed, some 25% to 30% of the decisions would be reversed. Not even the 'top' and 'bottom' ends of the ratings are safe from such reversals (Cole, Cole & Simon 1982). Some American scientists have drawn the radical conclusion that a lottery would then be equally 'just' —and much more efficient ('US research may drop peer review for lottery', *THES*, 22.2.91).

In sum, the functioning of peer review in its more or less 'classical' applications to manuscripts and to research grants do not stand up well to the tests devised by students of the operations of science. Although '[m]ost referees do a good and painstaking job' individually, for no other immediate reward than 'virtue' (E. Hunt, cited in Cicchetti 1980: 300), and although effects of wilful bias are not as visible as is sometimes feared, the reliability of peer review is so low that serious doubts about even these 'classical' applications may exist.

p. 138

§ IV Peer Review in External Quality Assessment Procedures

Let us return now to a more theoretical level, and ask in what ways peer review in the new quality assessment procedures differ from 'classical' peer review. The differences will be analyzed in two categories: the position of the peers, and the subject of the review.

Identification of Peers: Between Principal and Field

Peers do their reviewing for a certain purpose, namely, the making of decisions about scarce resources. As a rule, the reviewing peers do not make those decisions themselves. The decisions are made by what may be called the '*principal*' for whom the peers are the '*agents*' (for an introduction to the *principal–agent theory* see: De Alessi 1983; Moe 1984). The principal contracts out most of the preparation phase for the decisions to the agents, because the principal lacks the information, the authority, and/or the time to distribute the scarce resources in a way that is acceptable to the 'clients'.* In other words: the expected costs of preparing the decision

---

* 'Clients' are not part of the vocabulary of the principal–agent theory; the concept is used here to

('information costs') all by himself are higher for the principal than the expected costs of contracting agents to do so. Or, more concretely, in 'classical' peer review: journal editors cannot spare the time needed, nor do they have the expertise, to review all manuscripts by themselves; hence, other reviewers are employed to assist in deciding what will be published. And, in 'new' peer review: the government, when 'distributing' budget reductions, lack information on the quality of research or teaching in the country's university departments, and it does not have authority accepted in the scientific field (viz., scientific reputation); therefore, peers are used to provide both the information and the legitimacy.

The fact that agents depend on the principal for the reward from their efforts, will have effects on their behaviour: for them, a certain amount of utility is associated with complying with the principal's preferences. When this becomes internalized into a disposition to act (an attitude), it may be said that the peers develop an *identification* with the principal. On the other hand, being scientists, they also perceive utility in behaving according to (*identify* with) the norms in the disciplinary field. In a certain sense —though stretching the term beyond its original meaning— the disciplinary field can be seen as an abstract principal for individuals performing peer review. For, seen from the field (the scientists as consumers), it is useful to have some agents that assure the quality of publications (see also the citation of Zuckerman & Merton on page 4), or to have some agents that can act as a buffer against too detailed state intervention, such as selective budget reductions.

In 'classical' peer review situations the peers do not have too much problems in identifying with their principal *and* with the disciplinary field, for the principal is part of the field as well; the two identification-'forces' pull in the same direction (see figure 3). In the 'new' external quality assessment procedures, however, there is a distinctive difference between the disciplinary field and the principal. This also implies that quality assessors are not automatically seen as 'peers'. For example, in the Netherlands 15 so-called 'reconnaissance committees', which can be compared with external quality assessors, and which are set up by the Minister of Education & Science, have operated. Only one of these felt it could do without other backing for its assessment statements than its status as peers, but exactly this appeared to be debatable. And this resulted in this committee not gaining much support in the field; moreover, it is seen as one of the failures in the set of reconnaissance committees both in the field and in government circles (see Van der Meulen *et al.* 1991).

Who is the principal in the external quality assessment procedures depends on the specific institutional arrangement per country. In the Netherlands it is the VSNU, the Association of Cooperating Universities in the Netherlands, an umbrella organization of the thirteen Dutch universities, or the HBO-Council, the VSNU-counterpart for non-university higher education. Both organizations are 'owned' by the higher education institutions. The Dutch Minister of Education & Science is a very important —and interested— part of the public for the results of the quality assessments, even though the ministry's present 'philosophy' emphasizes its reticence as regards intervention. In Great Britain the principal is (or at least, used to be, before the White Paper *Higher Education: A New Framework* was published in May 1991) the UFC (University Funding Council) for the universities and the CNAA (Council for National Academic Awards) for the so-called public sector.* These two intermediate organizations are not 'owned' by the higher education institutions like the Dutch ones, but neither are they very closely connected with the government. Common to these cases, however, is that the results of the quality assessments will

p. 139

---

denote the third party in the relationship, namely, the scientists who hope to become the recipients of the scarce resource to be distributed in the process.

* Recently, the AAU (Academic Audit Unit) has initiated institutional reviews. These are, however, different from the education or research oriented peer reviews meant here; the AAU reviews had perhaps better be called 'meta-evaluation'.

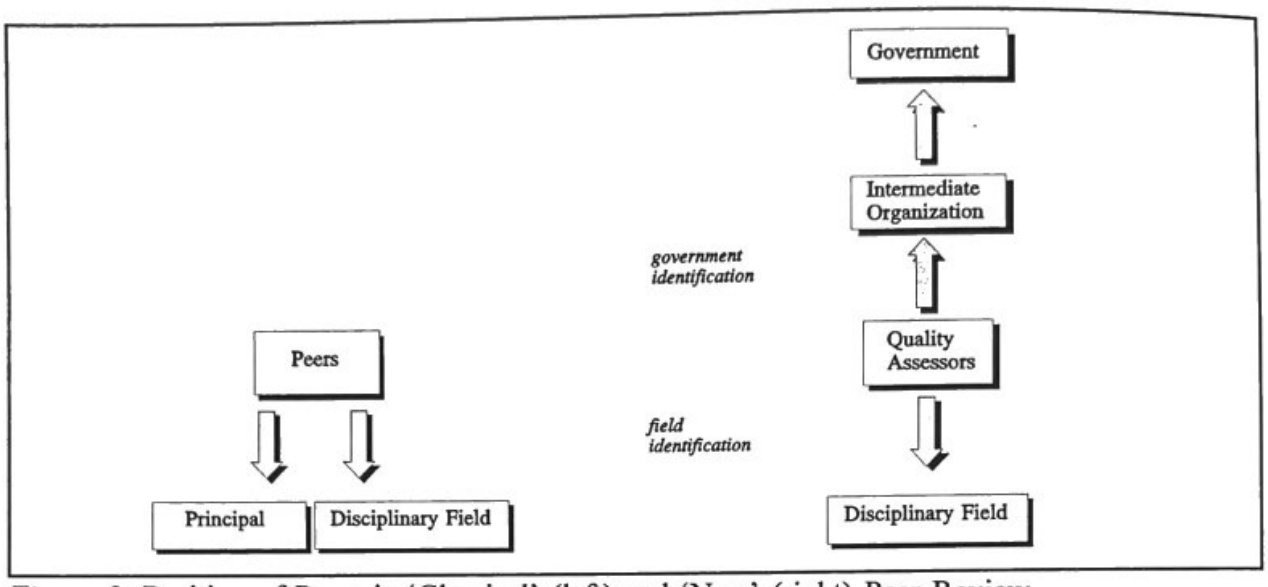

Figure 3 Position of Peers in 'Classical' (left) and 'New' (right) Peer Review

(or may) be used by the government for the purpose of budget allocation —in the present financial circumstances in Britain and Holland that means budget cuts.

It may be expected that the quality assessors will be aware of the possible financial (and other governmental policy-associated) consequences of their activities. In their standpoint they can accordingly take up a position on the continuum between complete *field identification* and complete *government identification*. Quality assessors in the external quality assessment procedures therefore are in an ambiguous position. If they identify completely with the field, the quality assessors will act as a buffer and try to defend the basic disciplinary organizations from the 'malevolent' government. If, to the contrary, they identify with the intermediate organizations (behind which one can feel the presence of the government and its purse), they will perceive the desirability to allocate tax money to the 'best' or otherwise most deserving groups in the field. An extremely pronounced choice for either government or field identification lowers the legitimacy of the quality assessors in the eyes of the other party. Both are important for these peers, so a certain degree of compromise can be expected. This can be combined in reality, however, with a recognizable choice without impairing the peers' effectiveness. For example, out of the 15 reconnaissance committees in the Netherlands 7 identified mostly with the government, 6 with the field and 2 took up an indeterminate intermediate position (see Van der Meulen *et al.* 1991). The identification by the committee played a crucial role in the alignment of the goals the peers adopted, the methods they used and the conclusions they reached. For example, committees that identified with the field were more inclined to interpret their task assignment freely, were less likely to use questionnaire methods and more likely to make their own 'peer' judgements, and did not make budget reduction recommendations. In sum, they functioned more as a buffer for the field than as a policy instrument for the minister. This also had consequences for the way they were viewed by the field and by the government, although that did not transpire in statistical correlations between identification and degree of acceptance of recommendations.*

p. 140

For peers in 'classical' peer review situations, questions regarding which principal to identify with hardly ever occur, and certainly not in such pregnant terms. Obviously, the position of quality asessors is more complex than the position of 'classical' peer reviewers.

---

* Please note that the number of committees was only 15. However, some other (expected) correlations did reach statistical significance (Spearman's rank order correlation, $\alpha = .05$).

Review of Collectives: Problems of Aggregation

Classical peer review implies in its 'ideal type' applications a review of a single scientist's performance. In the new peer review procedures the subjects of the review are usually collectives: research groups, study programme staff, faculties, or even whole universities. This means that performances, often 'produced' by individual researchers or a small group of them, are aggregated into assessments of larger collectives. The organization and implementation of a study programme is, indeed, a team performance. Yet it is contentious whether assessment of the collective is always possible. For example, if a high-reputation scientist or teacher moves out of a study programme, what will then be the consequences for the future quality of this study programme?

Such problems of attribution also complicate the application of performance indicators in the new external quality assessment procedures. Which faculty should the publications be ascribed to that are written by a scientist who moved to another group, or written by members of different faculties? And if these problems undermine the validity of performance indicators, what happens to the validity of a peer review that is based on such 'objective' data?

Answers to these questions I cannot give. The point I want to make, however, is only that in the 'new' peer review procedures, another 'layer' of problems is introduced, which must make its validity more questionable than the validity of previous peer review practices.

§ V Conclusion: Few Promises But The Best We Have

Empirical evidence about the practice of peer review even in its 'classical' applications is diverse, but in general does not portray a very satisfactory state of affairs. Unanimity exists about the low reliability of such peer review. Somewhat more contested is the question of bias, which is a matter of validity. Yet in that matter too the majority of data are disquieting.

Even more problems exist regarding the 'new' quality assessment procedures. Most important of these is the problem of identification for these peers: they face the opposing forces of *field identification* and *government identification*. The existence of this choice already points out that their 'peer-ness' is questionable (at least: not self-evident). The choice they make has consequences for the way they will operate and the acceptance of their assessments with the stakeholders (i.e., the field and the government). Moreover, another 'layer' of questions that lower the validity of peer review is introduced in the external quality assessment procedures with the aggregation of (individual) assessments to departments and other collectives. p. 141

Peer review obviously is not the ultimate answer to the problem of assessment of scientific products and producers, but as appears from the developments in at least Great Britain and the Netherlands, as well as from the literature, assessment methods principally based on peer review enjoy more legitimacy —which can be interpreted as a form of validity! (see e.g., Dochy, Segers & Wijnen 1990)— than assessment consisting of the application of performance indicators. Peer review accordingly promises to be the best we have, even though the best we have is fallible.

References

Banta, T.W. & Fisher, H.S., 1989: 'An International Perspective on Assessing Baccalaureate Program Outcomes' in: Banta, T.W. & Bensey, M.W. (eds.): *Proceedings of the International Conference on Assessing Quality in Higher Education* (Knoxville, TE: Center for Assessment R&D, University of Tennessee)

Blume, S.S. & Sinclair, R., 1973: 'Chemists in British Universities: A Study of the Reward System in

Science', *American Sociological Review* **38**: 126–138
Cave, M., Hanney, S., Kogan, M. & Trevett, G., 1988: *The Use of Performance Indicators in Higher Education* (London: Jessica Kingsley)
Cole, S., Cole, J.R. & Simon, G.A., 1981: 'Chance and Consensus in Peer Review', *Science* **214**: 881–886
Crane, D., 1967: 'The Gatekeepres of Science: Some Factors Affecting The Selection of Articles for Scientific Journals', *American Sociologist* **2**: 195–201
De Alessi, L., 1983: 'Property Rights, Transaction Costs and X-Efficiency: An Essay in Economic Theory', *American Economic Review* **73**: 64–81
Dochy, F.J.R.C., Segers, M.S.R. & Wijnen, W.H.F.W., 1990: 'Selecting Performance Indicators: A Proposal as a Result of Research' in: Goedegebuure, L.C.J., Maassen, P.A.M. & Westerheijden, D.F. (eds.): *Peer Review and Performance Indicators* (Utrecht: Lemma)
Epstein, W.M., 1990: 'Confirmational Response Bias Among Social Work Journals', *Science, Technology & Human Values* **15**: 9–38
Hargens, L.L., 1988: 'Scholarly Consensus and Journal Rejection Rates', *American Sociological Review* **53**: 139–151
Johnson, T.J., 1972: *Professions and Power* (London: Macmillan)
Latour, B. & Woolgar, S., 1979: *Laboratory Life* (Beverly Hills: Sage)
Marsh, H.W. & Ball, S., 1989: 'The Peer-review Process Used to Evaluate Manuscripts Submitted to Academic Journals: Interjudgmental Reliability', *Journal of Experimental Education* **57**: 151–169
Meulen, B.J.R. van der, D.F. Westerheijden, A. Rip & F.A. van Vught, 1991: *Verkenningscommissies tussen veld en overheid* [Reconnaissance Committees between Field and Government] (to be published by the Ministry of Education & Science)
Moe, T., 1984: 'The New Economics of Organization', *American Journal of Political Science* **28**: 739–777
Peters, D.P. & Ceci, S.J., 1982: 'Peer Review Practice of Psychological Journals: The Fate of Published Articles Submitted Again', *Behavioral and Brain Sciences* **5**: 187–195
Pollitt, C., 1990: 'Measuring University Performance: Never Mind the Quality, Never Mind the Width?', *Higher Education Quarterly* **44**: 60–81
Rip, A., 1988: 'Contextual Transformations in Contemporary Science' in: Jamison, A.: *Keeping Science Straight: A Critical Assessment of Science and Technology* (Göteborg: University of Göteborg)
Roy, R., 1985: 'Funding Science: The *Real* Defects of Peer Review and An Alternative To It', *Science, Technology & Human Values, 10, Issue 3: 73–81*
Segers, M.S.R., Dochy, F.J.R.C. & Wijnen, W.H.F.W., 1989: Een set van prestatie-indicatoren voor de bestuurlijke omgang tussen overheid en instellingen voor hoger onderwijs (Zoetermeer: Ministerie van Onderwijs & Wetenschappen) p. 142
Sizer, J., 1990: 'Funding Councils and Performance Indicators in Quality Assessment in the United Kingdom' in: Goedegebuure, L.C.J., Maassen, P.A.M. & Westerheijden, D.F. (eds.): *Peer Review and Performance Indicators* (Utrecht: Lemma)
Stigler, G.J., 1961: 'The economics of information', *Journal of Political Economy* **LXIX**: 213–225
Weert, E. de, 1990: 'A macro-analysis of quality assessment in higher education', *Higher Education* **19**: 57–72
Westerheijden, D.F., 1990: 'Peers, Performance and Power: Quality Assessment in the Netherlands' in: Goedegebuure, L.C.J., Maassen, P.A.M. & Westerheijden, D.F. (eds.): *Peer Review and Performance Indicators* (Utrecht: Lemma)
Westerheijden, D.F. & Weusthof, P.W., 1990: 'International Comparison of Quality in Higher Education', Notes for the presentation to the NUFFIC Seminar 'International Recognition of Diplomas and Degrees', EAIE Conference, Amsterdam, 7 December 1990
Whitley, R., 1984: *The Intellectual and Social Organization of the Sciences* (Oxford: Clarendon Press)
Zuckerman, H. & Merton, R.K., 1971: 'Patterns of Evaluation in Science: Institutionalisation, Structure and Functions of the Referee System', *Minerva* **9**: 66–100