Research Evaluation, volume 14, number 3, December 2005, pages 186–198, Beech Tree Publishing, 10 Watford Close, Guildford, Surrey GU1 2EP, England

Taylorised assessment

Advantages and dangers of 'remote' peer evaluation

Jochen Gläser and Grit Laudel

Under conditions of an increasing scarcity of reviewer time, a 'remote peer review' of research organisations — conducted without meetings between assessors or with the assessed academics at their institution — might be an easier and cheaper solution. This paper explores the impact of 'remoteness' on the practices of reviewers by analysing the recent Quality Review of the Australian National University. A taylorisation of the review process was observed that split the roles of designing the process, assessing the research, synthesising results, and taking responsibility for the outcome of the assessment. This taylorisation of the assessment process, the heterogeneity of individual assessment practices, and the low availability of publications in electronic format led to the conclusion that further organisational innovations are necessary in order to make remote peer reviews feasible.

Jochen Gläser and Grit Laudel are at the Research Evaluation and Policy Project, Research School of Social Sciences, Australian National University, Canberra ACT 02000, Australia;. Email: Jochen.Glaser@anu.edu.au

PEER REVIEW IS THE preferred method among the few evaluation procedures that have both proven practicable and gained at least some legitimacy within the science system,. Since it is rooted in scientific practice, academics are familiar with it, have possibly practiced it themselves and regard it as matching the nature of their work. However, peer review needs peers; that is, academics who work in the field to which the subject matter of evaluation belongs and who are therefore competent to judge. This seemingly trivial necessity is becoming increasingly problematic. Assessing the work of colleagues is a distraction from one's own work, and is therefore commonly regarded as a necessary burden. The growing demand for review work is diminishing the willingness of academics to take part in evaluation exercises.

The emerging scarcity of reviewers is particularly problematic in countries whose science systems are too small to supply a sufficient number of an international elite who have the authority to act as reviewers. These countries must invite assessors from abroad, who have to engage in an even more time-consuming and cumbersome process. Under these conditions, attempts to take advantage of modern technologies and to conduct 'remote' peer reviews seem a natural response. In a 'remote' peer review the reviewers don't meet any other party but evaluate material that is sent to them and submit their judgement. The socially (and therefore the sociologically) important aspect is that reviewers meet neither each other nor the academics whose work is to be evaluated.

Remote peer review has been a very common practice at the grassroots of scientific communities for a long time. The most frequent assessments —

 0958-2029/05/030186-13 US$08.00

those of manuscripts of articles, conference contributions, and project proposals — are conducted this way.

However, remote peer review has usually not been applied in the evaluation of organisations. The standard operating procedure for this kind of evaluation appears to be an on-site assessment by a group of reviewers who interact with both the organisation's management and its academics. In the case of the United Kingdom's Research Assessment Exercise, interactions with the evaluated units are impossible because of the scale of the process but reviewers at least decide collectively in review panels.

If a remote peer review of organisations is emerging, then it is important to know how the differences between 'on-site' and 'remote' practices might affect the outcomes of review processes. When reviewers don't meet each other and don't meet the people whose work or organisation they are evaluating, they might act differently and produce different assessments; that is, assessments that have a different content or a different validity.

While modern information technologies seemingly compensate for remoteness by enabling easy communication and collective decision-making, they cannot create conditions that are identical to those of an on-site peer review. In their comparison of 'virtual collaborations' (collaborations via the Internet) and collaborations that rest on direct interactions between the scientists, Olson and Olson identified five key characteristics of the latter that will be only poorly supported even by future technologies, namely:

- Informal 'hall' time before and after: Impromptu interactions take place among subsets of participants upon arrival and departure;
- Co-reference: Ease of establishing joint reference to objects;
- Individual control: Each participant can chose what to attend to, and change the focus of attention easily;
- Implicit cues: A variety of cues as to what is going on are available in the periphery;
  and
- Spatiality of reference: people and work objects are located in space, (Olson and Olson, 2003, pages 31, 37)

It is by no means obvious which of these aspects are important for evaluations by collective peer review, and what their absence might mean for the outcomes of remote peer reviews. Systematic analyses of remote peer review processes are necessary in order to establish the effects of remoteness. The aim of this article is to contribute to this general task by answering the question of how the 'remoteness' of the recent Quality Review of the Australian National University (ANU) has affected the practices of the parties involved and the outcomes of the process.

Approach

Questions about the specifics of remote peer review are intrinsically comparative. Therefore, a theoretical approach is needed that supports comparisons of peer review processes. Such an approach cannot easily be found. There is a stark discrepancy between the number of empirical 'peer review studies' and the theoretical understanding of the process. The perspectives that have been applied so far can be described as 'theoretical scientism' and 'atheoretical scientism'. 'Theoretical scientism' was characteristic of the Mertonian sociology of science that pictured science as a rational enterprise with scientists guided by the scientific ethos. Peer review was regarded as a scientific activity, and empirical studies analysed the extent to which peer review is universalistic, disinterested, and part of a scientific community's organised scepticism (e.g. Cole *et al*, 1978, 1981; Chubin and Hackett, 1990). After Mertonian sociology of science had been supplanted by the sociology of scientific knowledge, which is by default not interested in trans-local, non-microscopic processes such as peer review, the literature shifts to atheoretical, empiricist approaches by scientists and editors who were mainly interested in the validity and reliability of peer review processes. These studies, too, presuppose that peer review is a scientific process, involving rational decision-making in which objective criteria are consistently applied by various reviewers.

The scientistic yardstick, which is used in most criticisms of peer review as well as in suggestions to improve it (e.g. Cicchetti, 1991), has recently been challenged from a constructivist perspective. Hirschauer (2004) argues that the peer review of manuscripts is a process of collective knowledge construction, in which authors, reviewers and editors jointly construct the published article. The constructivist perspective can be extended to the construction of project proposals. It has gained some empirical support by observations (Knorr-Cetina, 1981, pages 81, 88–89) as well as analyses of manuscripts and reviews (Myers, 1990). According to these analyses, authors and applicants anticipate peer review while writing, and reviewers' and editors' respective funding agencies contribute demands and suggestions that co-shape the outcome of the review process. This process is as 'messy' (idiosyncratic, shaped by personal interests and power constellations) as any knowledge construction process in science. It cannot be 'objective' because the actors who take part in the collective knowledge construction can apply only their individual scientific perspectives, which are shaped by their individual research biographies, interpretations of the existing knowledge, personal networks and local working environments (Gläser, 2004, pages 74–78).

The interpretation of peer review as a process of collective, negotiated knowledge production implies that it is characterised by a specific actor

constellation, and that it is necessary to analyse how varying actor constellations produce different outcomes (reviews) depending on their specific conditions of action.[1] Institutional conditions can be assumed to be among the most important because they specify who takes part in the review process, what are the roles of the different actors, and what power the members of the actor constellation have in the negotiation of knowledge claims.

These considerations suggest the application of a framework that emphasises actors and institutional conditions of action. In our study, we applied the actor-centred institutionalism, a neo-institutionalist analytical framework that has been developed for policy analyses (Mayntz and Scharpf, 1995; Scharpf, 1997), and has been modified to support the analysis of institutions in science (Laudel, 1999). Applying this approach requires answering the following general questions:

- Which actors are involved in the review process, and what are their respective roles, interests, and power relations?
- What negotiations/knowledge construction processes occur in this actor constellation?
- What conditions (institutional, epistemic, and others) affect the negotiations/ knowledge construction processes?
- How do the actor constellation and the specific conditions of action shape the outcome of the review?

For the analysis of the remote peer review, these questions can be specified as follows: *To what extent are the actor constellation, conditions of action, interactions and therefore the outcomes of the ANU Quality Review affected by its conduct as a remote peer review?*

Data and methods

The investigation drew on a variety of data. The *analysis of documents* included the final report (ANU, 2004a), material submitted by the ANU to the review committee(s) (ANU, 2004b), and internal documents that were produced during the assessment process. A second source of data was the *databases* that had been created in the review process, which contained information about the material submitted for assessment, the reviewers, and their assessments. Additional information about publications submitted for assessment was obtained from the *Web of Science* and from the *database of the Research Evaluation and Policy Project (REPP).*[2] More specifically, publication and citation data for academics from selected disciplines were obtained from the Web of Science, and actual and expected citations for selected publications by academics currently working at the ANU were obtained from the REPP database.[3]

In order to learn more about the assessment practices, *interviews with assessors* were conducted in Australia (1), the USA (1), the UK (3), Germany (2) and the Netherlands (1).[4] Additionally, one interview with an ANU administrator about the role of the ANU in the process was conducted.

Textual data were analysed by computer-aided qualitative content analysis, which is essentially a procedure of extracting data by using categories belonging to an analytical scheme (in this case, the scheme described above) and identifying types (in this case types of conditions/ practices/ outcomes) in the material (Gläser and Laudel, 2004). The following categories were used in the content analysis:

Task: attributes of the task assigned to the assessors;
Object: cognitive attributes of the research respectively field that affect the ranking;
Representation: attributes of the material submitted to assessors (publications and context statements);
Criteria: criteria applied in the assessment;
Subject matter: subject matter to which the criteria are applied;
Basis: information which is used in the assessment;
Refusal: refusal to assess;
Hedging: hedging of the submitted assessment;
Competency: self-evaluations of the assessors with regard to their competency;
Relations: prior or current relations between the assessor and the ANU;
Content: any particular content of the assessment that is an effect of the specific process;
Validity: any consequences for the validity of the assessment.

Data on assessors and the fields defined for purposes of the review process were analysed by using descriptive statistics.[5]

Results

Actor constellation and the review process

The most powerful actor in this review process was the organisation evaluated; namely, the ANU. The ANU designed the review process and thus defined the positions of all actors in the process as well as

The ANU designed the review process and thus defined the positions of all actors in the process as well as their tasks, rights and responsibilities. This included the selection of assessors

their tasks, rights and responsibilities. This included the selection of assessors. This case is unusual and had to do with the specific situation of the ANU. A review of ANU's Institute of Advanced Studies was due, and the ANU thought it should be conducted together with a review of the faculties. Furthermore, the ANU felt the need to position itself in the current higher education policy debate (ANU. 2004a, page 3). The ANU's Council therefore initiated a general quality review of the ANU in January 2004, which addressed not only research but also research training, undergraduate and postgraduate education, the impact of ANU's regional and national service, and the strength of ANU's international engagement (ANU, 2004a, page 78). The assessment was not comparative; that is, ANU was the only university under evaluation, albeit the final report compared ANU to other Australian universities wherever possible. The results of the assessment were due in November 2004, which put the whole process under considerable time pressure. Many design features can be ascribed to this limitation.

The major constraint put on the design of the process was that it needed to be a legitimate, that is, an independent assessment. The ANU therefore installed two committees, a moderating committee that prepared a draft report and a review committee that prepared the final report. Both committees consisted of internationally renowned academics and acted independently.

Figure 1 illustrates the assessment procedure, which was designed by the ANU itself. The ANU defined 22 disciplines and 108 subdisciplines, for which assessors needed to be found.[6] Assessors were nominated by ANU's academics, who sometimes informally approached colleagues and asked them if they were willing to take part in the assessment. On the basis of the nominations, ANU formally approached the potential assessors. Figure 2 shows the geographical distribution of the 280 assessors who returned an assessment. Of the 44 Australian assessors, three-quarters were from subdisciplines which can be assumed to have a nationally or regionally specific content. In these fields expertise on ANU's research might be scarce outside Australia or the Asia-Pacific Region.

Organisational subunits of the ANU (faculties, research schools, and centres) were defined as 'units

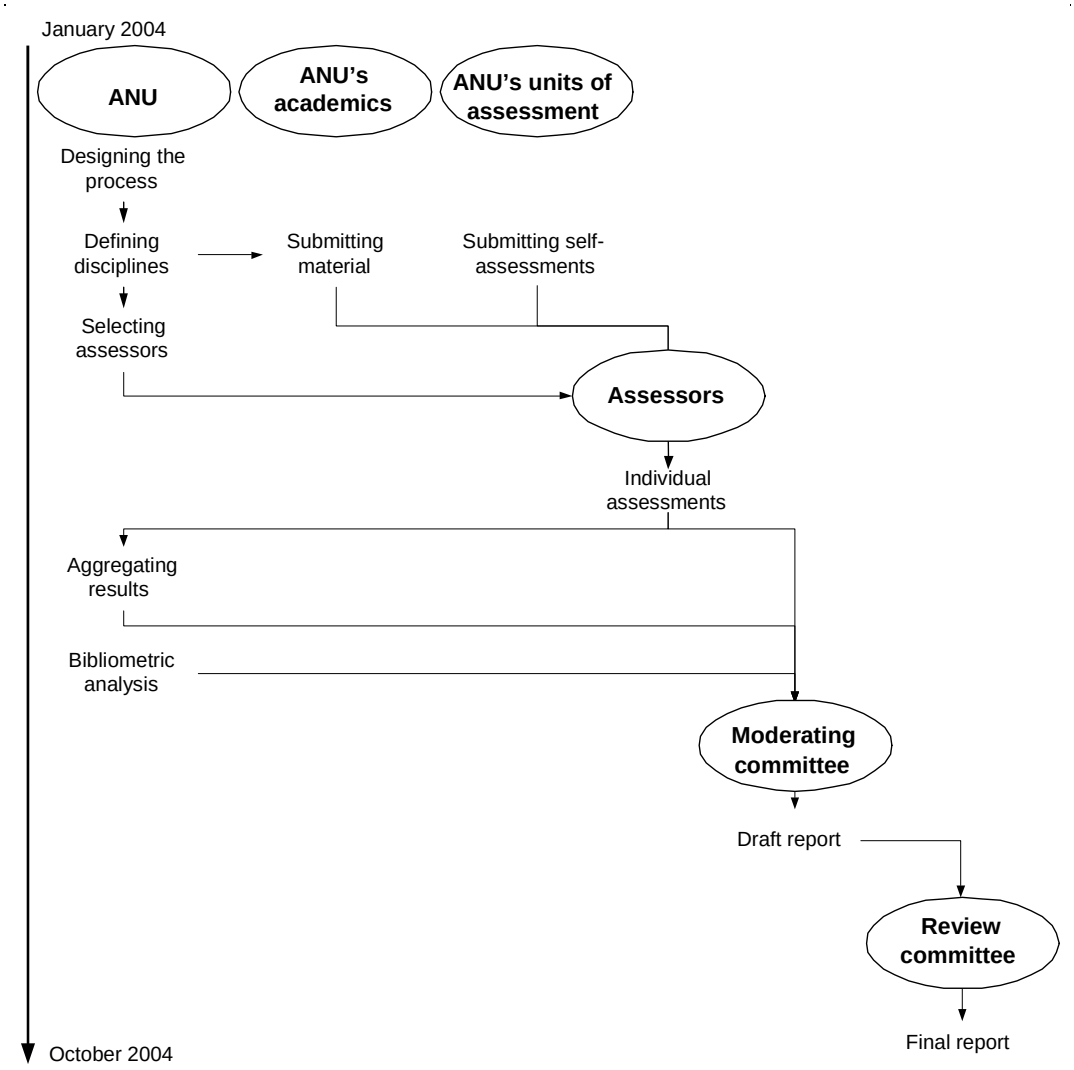

Figure 1. The review process

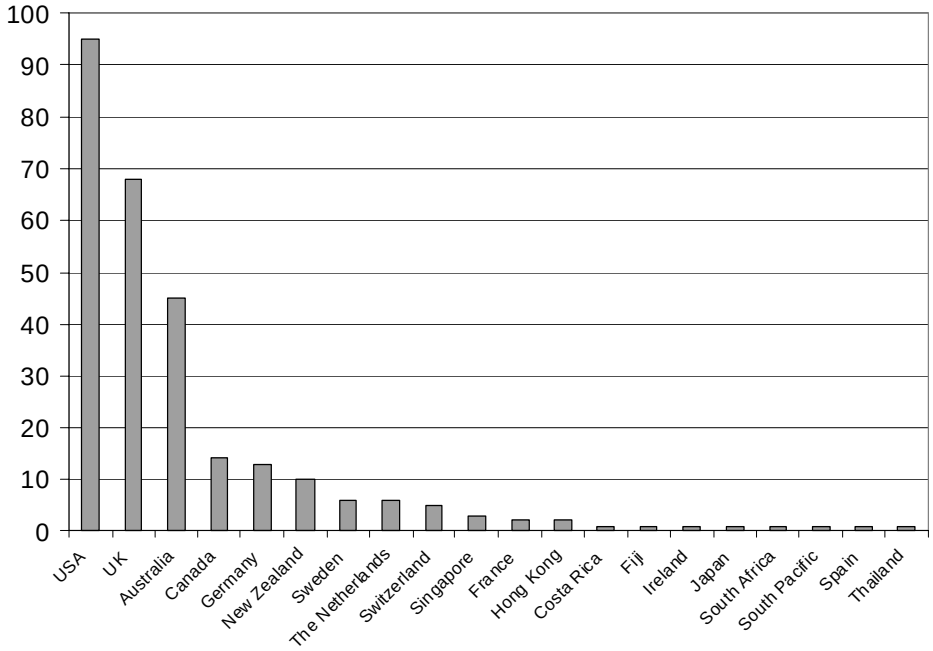

Figure 2. Geographical distribution of assessors

of assessment' (UoAs). The academics of these UoAs were requested to nominate their five best publications since 1995 and to assign each of them to a subdiscipline. Interdisciplinary publications could be assigned to up to three subdisciplines. Early career academics could either be excluded from the submission or nominate fewer than 5 publications. This decision was made by the UoA, which is why the inclusion of early career academics in the exercise varied across the ANU. Electronic versions of the nominated publications were obtained where possible. Self-assessments were requested from UoAs which included data on research profiles, staff, indicators of esteem, collaboration, competitive funding, a self-assessment of publications, and research training.

The assessors received an email which contained a letter describing their tasks and commenting on them; and the material for assessment in their subdiscipline. The material consisted of:

- a list of nominated publications;
- electronic versions of nominated publications in their subdiscipline whenever available; and
- the self-assessments of UoAs that conduct research in the subdiscipline.

The assessors also had access to a secure website containing a database with all ANU publications since 1995; nominated publications and self-assessments were available. If they felt they needed publications that were not sent to them in electronic form, they could request a paper copy of the publication.

Thus, the outcome of ANU's research for the period from 1995 to early 2004 was subject to the evaluation. Each assessor was given two tasks; namely, to assess ANU's research as represented by the publications submitted in his or her subdiscipline, and to assess the UoAs that conducted research in that subdiscipline. The questions were phrased as follows:

1. For the assessment of ANU's research in the subdiscipline:
   - 'What proportion of all submitted research works in your sub-discipline do you estimate to be world class (in the top 25% of international research in the field)? ____%'
   - 'What proportion of all submitted research works in your sub-discipline do you estimate to be exceptionally significant (in the top 5% of world research in the field)? ____%'
   - 'How many works have you sampled?'
2. For the assessment of UoAs that were active in the subdiscipline:
   - 'A recent study of the world's top 500 research universities ranks the ANU within the top 50. Where do you regard the ANU among the world's top universities in your field of research?
     - in top 25
     - in top 50
     - in top 100
     - in top 200
     - outside top 200'

Additionally, the assessors 'were asked for — and in almost all cases provided — a qualitative comment to support their numerical assessment' (ANU, 2004b, page 28). Most of the assessors submitted their judgement by email. In many cases, explicit comments on the task and the submitted material were included.

The ANU aggregated the quantitative assessments and conducted a quantitative content analysis of the comments. The aggregated assessments for each discipline and an independently conducted bibliometric analysis were submitted to a moderating committee whose task was to assess the aggregation conducted by the ANU and to prepare the material for the final report. The final assessment of ANU's research in the report is very positive (ANU, 2004a, pages 21–41).

The task of ANU's academics

Academics nominated more than 6,000 publications and assigned them to subdisciplines. Because of multiple assignments (one publication could be assigned to more than one subdiscipline), 7,521 publications can be found in the 108 subdisciplines. Table 1 gives an overview of the most frequently nominated publication types. If we look at the distribution of publication types in the various disciplines, a familiar pattern emerges (Figure 3).

First, there are 'journal disciplines' in which journal articles comprise more than 80% of the output (natural sciences including biology and biomedical research; mathematics; and business and commerce). Their second most important channel of communication is book chapters.

A second type of discipline is the 'augmented journal discipline'. In these disciplines, journal articles are still the most frequent output, but amount to less than 70%, and other channels of communication play an important role. Information science and engineering is a subtype all by itself here, with conference papers being the second most important channel of communication only in this discipline. The other subtype in this discipline comprises of fields in which books and book chapters are the second important channel of communication. These fields can be distinguished according to their third channel of communication, which is either conference papers (philosophy, environmental research; and Asia-Pacific studies) or research reports and reports for external users (economics; law; Australian and indigenous studies; and policy and political sciences; psychology and mental health; and studies in human society).

Table 1. Most frequently nominated types of output

Type of output	Frequency	Percentage of total
Journal articles	4,741	63.0
Book chapters	1,147	15.3
Books	759	10.1
Conference papers	287	3.8
Edited books	168	2.2
Creative works	158	2.1
Research reports/working papers	116	1.5
Reports for external users	50	0.7
Encyclopaedia entries	22	0.3
PhD theses	12	0.2
Other	61	0.8
Total	7,521	100

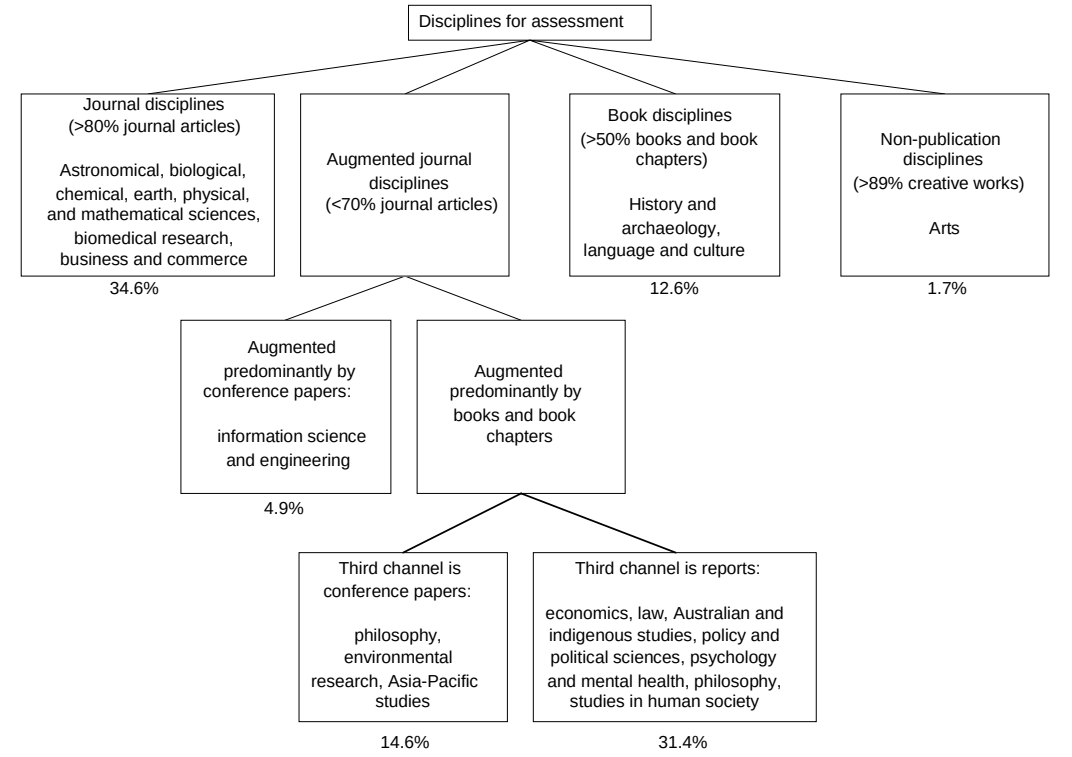

Figure 3. Types of disciplines and their share of nominated publications

A third type comprises 'book disciplines' in which journal articles contribute less than half of the output, and books and book chapters are the most important channel of communication (history and archaeology; and language and culture). Reports are of minor importance in these disciplines because applications are less important. Finally, the arts constitute a type of their own because journal articles are insignificant for them. The major output of this 'discipline' is creative works, which raises the question of whether they should be treated like a discipline at all.

An interesting factor that had the potential to influence the evaluation was the concept of quality underlying the nomination of best publications by ANU's academics. This concept was further explored for four sample disciplines, namely astronomical sciences (journal discipline); information science and engineering; environmental research; and Australian and indigenous studies (augmented journal disciplines). Since the comparison partly draws on citation databases, the book disciplines had to be excluded because no valid information about them can be obtained from those databases. Even the augmented journal disciplines need to be treated cautiously.

The nomination of 'lowly reputed publication types' such as reports could have occurred for reasons that have nothing to do with quality. In particular, academics could have been forced to nominate the reports because they had no other publications or no 'better' publications (journal articles, etc.). This hypothesis was tested for the four selected disciplines. Academics who selected 'lowly reputed' publication types were found to have had alternatives. Most of them had more than five publications and had published journal articles which they did not nominate.

Furthermore, we analysed whether academics regarded their most highly cited publications as their best ones. Citations to all publications of a sample of 10 academics from astronomical sciences and 15 academics from indigenous and Australian Studies were obtained from the Web of Science. The results showed that none of these academics nominated his or her five most highly cited publications, and that three academics from each sample did not nominate their most highly cited publication. This observation confirms findings for chemistry by Porter, Chubin, and Jin (1988) for a wider range of disciplines.

In the case of the astronomical sciences,[7] two intervening factors could be excluded. First, 'strategic nominating' of co-authored articles could have occurred; for example, a nomination by a co-author that would bring the highly cited publication into the sample but would provide the co-authors with the opportunity to nominate another of their publications. This did not consistently happen; many highly cited publications were not nominated at all. Second, nominating academics could have preferred younger publications that would have obtained fewer citations. This was sometimes the case. However, there were other cases in which the nominated publications were not the most frequently cited from their year of publication, and publications exist that are both more recent and more frequently cited.

To further test the relationship between nomination and citation scores we compared the actual and expected citations of nominated ISI items in astronomical science and environmental research to the values obtained for the whole output of ANU in these fields.[8] Actual and expected citations were several times higher than those of the whole ANU output, which means that in the aggregate the academics nominated journal articles whose citation scores were far above the average of their output.

Conditions of assessments

The relevant conditions of action for the assessors were constituted by the task they were given, by the material they had to assess, by any additional information they could draw on, and by the procedure designed by the ANU. An important condition was the extreme time pressure under which the whole exercise was conducted. This time pressure affected the assessors because it limited the time they could spend on the assessment and their opportunities to receive material that was not electronically available.

The assessors were not well informed about the overall review process in which they took part. In particular, they did not know what would happen to the assessments they were asked to submit. They did not know if the publications they were asked to assess were submitted to other subdisciplines as well, and if the academics whose publications they assessed had submitted other publications in other subdisciplines.

A third important condition was the isolation of assessors. The assessors could have known the other assessors in their subdiscipline (from the addresses in emails that were sent out to all assessors simultaneously), but they were not explicitly informed about other assessors. Their only contact at the ANU was the administration that sent out the material for review, answered questions and collected the results of the assessments. The assessors did not communicate with ANU's academics while evaluating their

The assessors were not well informed about the overall review process in which they took part. In particular, they did not know what would happen to the assessments they were asked to submit

research. It must be noted that only one of our interviewees would have preferred to communicate with his fellow assessors during the exercise. The others emphasised that it is much more efficient to work alone, without the need to discuss the assessment and to achieve agreement.

Fourth, the tasks that emerged from the definition of subdisciplines, assignment of assessors and submission of publications varied significantly between assessors. Table 2 provides an overview of the distribution of assessors and publications across subdisciplines.

Table 2 highlights some problems resulting from the definition of disciplines and subdisciplines and the assignment of assessors. Approximately 20% of the submitted publications, in about 40% of the subdisciplines, were assessed by only one assessor. The workload for some of these assessors was rather high. As a rough indicator of the workloads, we calculated the number of publications per assessor in each discipline. The average workload in the subdisciplines varied between 102 and two publications per assessor in the subdisciplines with only one assessor. Subdisciplines with more than one assessor lay in between those extremes.

The content of the task was very tightly prescribed in some respects and very imprecise in others. The most precise part of the task referred to the ranking that was to be returned to the ANU (see 'Actor constellation and the review process' above). The 'Guidelines for External Assessors' gave a short explanation of what the ANU meant by 'exceptionally significant' and 'internationally significant'. However, in this explanation an alternative to the international benchmark was introduced by stating that the research could also be 'an equivalently exceptional contribution to research in an area of particular significance to Australia' (ANU, 2004b, page 213). The alternative between international excellence and significance to Australia caused problems for some international assessors who felt unable to assess the significance of research to Australia.

Table 2. Distribution of assessors and nominated publications across subdisciplines

Number of assessors in a subdiscipline	Numbers of subdisciplines	Number of publications submitted
1	41	1,588
2	30	2,002
3	13	1,114
4	5	482
5	7	671
6	5	511
7	2	422
8	4	608
9	1	123
280 assessors in 108 subdisciplines		7,521

While the outcome of the assessment was highly standardised, the assessors were granted considerable freedom with regard to the way they arrived at these rankings. The 'Guidelines for External Assessors' suggested that the assessors did not necessarily need to read the publications in order to arrive at their judgement:

> We do not expect you to read all the work presented. We expect you only to sample the work. Make your judgement based on your knowledge of the reputation of researchers, the quality of the journals and the quality of the work. (ANU, 2004b, page 214)

A fifth important condition for the assessment was constituted by the material under evaluation. The researchers received a list of publications, the nominated publications belonging to their subdiscipline in an electronic format (wherever they existed in that format), and the context statements of the UoAs. Thus, the assessment took place after ANU divided its research into subdisciplines, and assigned assessors and publications to the subdisciplines in two independent processes. As a result of the division and allocation processes, the following problems occurred:

- *Misallocations*: Assessors declared themselves 'not competent' or 'not fully competent' to judge the publications submitted to them in 'their' subdiscipline. It was assumed by ANU that the limited competency of one assessor would be compensated for by the other assessors in the subdiscipline.
- *Wrong division*: In the social sciences, arts and humanities some assessors complained that the wrong boundaries had been drawn between subdisciplines.
- *Atomisation*: According to other complaints by assessors, the division of the material went too far and left them with not enough material for a judgement. This problem occurred in two different forms. Some assessors were left with very few (sometimes even only one) publications to assess, which made judgements (and particularly the ranking exercise) impossible. The other problem, which remained implicit in many instances, was that publications (and context statements) were not a sufficient basis for a judgement of the quality of research.

Finally, an important condition for the assessment was the availability of publications in electronic format. According to both the ANU database and the packages sent out to assessors, less than half of the nominated output could be sent to assessors in an electronic format. Assessors were encouraged to

request printed versions of publications whose electronic versions were not available. However, few of them did so. Some assessors obtained the publications from their own libraries. How often assessors did this is impossible to tell. Thus, it must be assumed that up to half of the publications submitted for evaluation have not been read by assessors because access was too difficult. The availability of publications varied between disciplines. Table 3 shows the results for our four examples. Astronomical sciences had the highest on-line availability of all disciplines. The values for the humanities were lower than 20 %, and on-line availability for the arts was close to zero.

Practices of assessors

The practices of assessors varied significantly. This was to be expected because assessment practices are field-specific and depend on personal research biographies, styles, and the actual conditions of the assessment (e.g. the availability of time).

While assessors were asked to assess published research results, many of them consciously or unconsciously choose a different subject of evaluation, namely researchers or UoAs. In many of the comments submitted by the assessors, the quality of academics or UoAs rather than that of publications submitted in the subdiscipline was discussed. The inclination to assess academics rather than research was at odds with the atomistic approach chosen by the ANU. Some assessors complained that important publications of the academics whom they were assessing were missing in the material, or that academics had submitted only one or two publications — not knowing that the missing publications might have been submitted to a different subdiscipline. In some cases, assessors circumvented the limitations of the atomistic approach by using the academics' whole publication lists to arrive at an assessment.

A second dimension in which the practices of assessors varied was the criteria they applied in their assessments. As has been described above, the 'Guidelines for External Assessors' gave them considerable freedom concerning the way in which they arrived at their assessment. Reading all or some of the publications and judging them on the basis of their content was only one of the options. While many assessors actually chose this option, the extent to which the publications submitted by the ANU were actually read remains unknown because many assessors did not describe in their comments how they conducted their assessments. Those who did, and the eight assessors we interviewed, applied a wide range of criteria. Criteria that were used to judge the content of publications remained mostly implicit. When they were mentioned, the expected terms ('breakthrough', 'highly original', 'excellent', etc.) occurred.

Beside the criteria for judging content, a wide range of second-order criteria was applied. Second-order criteria do not refer to the content of work but to other properties of the research (grants, publications, awards, etc.), from which conclusions about the content can be drawn. The most commonly used second-order criteria were attributes of publications. Among those, publication types, importance of journals, and reputation of publishers were the most frequently mentioned second-order criteria.[9] Some assessors were critical of researchers who submitted reports, conference papers or other publications of lower reputation. Numbers of publications or specific publication types (e.g. sole-authored books) were also used. With regard to the importance of journals, both criticisms of publishing in 'minor' journals and approval of publishing in important journals occurred. When assessors commented on books, they usually did so positively by emphasising the reputation of the publisher. Few assessors criticised the submission of review articles because they regarded them as publications of lesser importance.

In some cases, the approach to attributes of publications was a quantitative one. Assessors did not simply mention the importance of journals but mentioned impact factors. In other cases, citations to the submitted publications were obtained from the Web of Science and used to assess the publications.

The second-order criteria were not always used as a substitute for the content of publications. They also served as a supplement — either as an additional means of judging publications or simply as an additional argument to support a claim about the quality of the material. In one case the simultaneous assessment of content and type of publication led to the comment that the publications are of good quality and should therefore be submitted to 'better' journals.

A third kind of criteria that were used in the assessment of UoAs drew on information from the context statements. Assessors reasoned about numbers of researchers, external funding, awards, etc. However, it is not clear to what extent these criteria were used to arrive at a ranking. They served as arguments in the comments on the unit, but it did not become clear how they were 'translated' into rankings of UoAs.

Table 3. Availability of electronic versions of publications (sample disciplines and total)

Field	All publications	Electronic publications	Share (%)
Astronomical sciences	123	119	96.7
Information science and engineering	371	286	77.1
Environmental research	209	93	44.5
Australian and indigenous studies	216	44	20.4
All disciplines	7,521	3,526	46.9

A final aspect of the practices of assessors that must be discussed is their response to the problems caused by the division and allocation process (misallocation, wrong division, and atomisation, see above). At least some of the assessors appear to have conducted the assessment as requested regardless of their own critical comments on the conditions of the assessment. Apart from this 'comment and carry on' strategy, the following responses could be identified:

- Assessors compensated for the atomisation and the resulting insufficient information by *drawing on their context knowledge and applying their own information-gathering strategies*. Publication lists were obtained where necessary, Google searches conducted, and the own library was used when publications were not available in electronic format. However, only a few assessors appear to have resorted to actively gathering additional information. For example, none of our interviewees used the website set up by ANU for obtaining full publication lists or other information. The most important resource was the context knowledge of the assessors; that is, their knowledge of the UoAs and their work. It could be concluded from the comments that many assessors had intimate knowledge of the research and researchers they were assessing. This was the inevitable consequence of the selection process. Since ANU's academics nominated their assessors, it was only natural that they suggested colleagues they knew, and who knew them. In our interviews, a variety of relations between assessors and ANU were mentioned. Assessors collaborated with ANU's academics, were visiting fellows either at ANU or elsewhere in Australia, had supervised PhD students from ANU, or had taken part in on-site evaluations prior to the recent Quality Review.[10] Apart from these direct contacts, assessors often knew ANU's work from the literature; that is, they had read some of the publications submitted to them in the course of their own research.
- A specific response to a sense of lack of competence was *resorting to second-order criteria*. These criteria were applied not only in their own right, but also as a response to misallocations.
- In some cases, *assessors refused to produce a ranking*. This happened mostly in cases where assessors felt incompetent to judge the material or where the number of publications submitted (either absolutely or in their area of competence) was felt to be too small.
- In other cases *assessors 'hedged' their results*; that is, cautioned the reader that the validity of their ranking might be limited. 'Hedging' was also practiced by assessors who doubted the validity of the whole ranking exercise as a method of assessment.

From the content of the comments it can be concluded that the assessors did not know what would happen to their rankings and comments. The assumption underlying the comments was that they would be read by a competent colleague and taken into account in the production of the final assessment.

The assumption underlying the assessors' comments was that they would be read by a competent colleague and taken into account in the production of the final assessment

The aggregation process

The task of the assessors was to evaluate the segment of ANU's research in their subdiscipline. In order to achieve an overall judgement of ANU's research, these assessments needed to be integrated. The integration was not conducted by assessors, but by ANU's administration. This aspect of the actor constellation made it necessary to obtain decontextualised assessments from the assessors; that is, assessments that could be handled in a meaningful way by non-specialists. The approach chosen by ANU was to solicit rankings; that is, numbers that could be easily aggregated (see 'Actor constellation and the review process' above). The numbers were aggregated first for the subdisciplines and then for each major discipline. Because the numbers in the subdisciplines were too small, no results for subdisciplines were reported. The final report contained discipline tables listing the share of ANU's works in the top 5% and top 25% internationally (ANU, 2004a, pages 26, 29).

The aggregation weighted the judgements by the number of publications that were submitted to the respective subdiscipline. The number of publications sampled by the assessors was not taken into account, because the question was regarded as too ambiguous ('sampled' could be interpreted as 'belonging to the assessor's area of expertise', 'known to the assessor', 'looked at', or 'actually read'). Equally, all comments on competence, the hedging of their judgement by some of the assessors etc. were ignored. The suggestion in the 'Guidelines for External Assessors' that excellence in an area of 'significance to Australia' can be regarded as equivalent to international excellence did not lead to a specific approach in the aggregation. All rankings were aggregated as reporting 'international excellence'.

While the assessors' comments on the scope and validity of the submitted rankings were lost in the aggregation process, the comments on the research and on UoAs were analysed. A content analysis was conducted by the officer who had organised the initial process of individual assessments. Comments

on the quality of ANU's research were categorised and aggregated. The final report states that 79% of the assessors' comments were 'unqualifiedly positive' (ANU, 2004a, page 30).

When presented with the results, the moderating committee became concerned about the validity of both the rankings conducted by the assessors and the aggregations. They requested an inquiry into the complaints of assessors about the process. They also requested to see the original comments given by the assessors. It turned out that only few of the assessors had criticised the procedure. The moderating committee (and subsequently the review committee) were satisfied that the procedure was valid, and produced the final report.

Discussion

If we estimate the cost of bringing in international assessors to the ANU at 3,000 Australian dollars per assessor, and that the costs of preparing and managing the quality review would have been roughly the same, the remote peer review saved the ANU at least 700,000 Australian dollars. This is a clear advantage of a remote peer evaluation.

The process designed by ANU was an unusual peer review in several respects. First, it was characterised by a *particular actor constellation*. The three crucial tasks of a peer review process — designing the assessment process, conducting the actual assessment, and producing (and therefore taking responsibility for) the final report — were performed by different actors. The ANU designed the process, the 'remote' assessors conducted the assessment, the ANU conducted the synthesis, and a review committee produced and took responsibility for the final report. The process implied that the assessors had no control over the use that was made of their assessments, and that the committee that took responsibility for the final report had no control over the collection of the data on which the report was grounded.

This design is a bureaucratic temptation that comes with 'remote peer review'. Since the absence of assessors is a constitutive feature of 'remote assessments', someone else has to design the assessment procedure. For reasons of legitimacy the procedure must also involve highly reputed academics who take responsibility for the final assessment. Given these requirements (and the time pressure under which the Quality Review was conducted), the described actor constellation was not inevitable but likely to occur.

A second peculiarity of the process was the *taylorisation* of the review process. The extreme division of labour meant that the assessors had to fulfil a very reduced task; namely, to produce rankings of research and organisational subunits on the basis of research done in one subdiscipline. For this task, they were provided information that was similarly reduced; namely, a selection of publications in that subdiscipline and the context statements of UoAs. In many cases this approach was at odds both with standard assessment practices, which are focused on academics or organisations, or with the informational requirements of the assessment process. If it were not for their context knowledge of ANU's research (apparently an unintended effect of the assessor selection procedure), many more assessors would have probably felt uncomfortable with their task.

Because of the taylorisation, all assessors worked in *isolation* from each other. There was no collective decision-making on the assessment of research at the levels of subdisciplines and disciplines. Only a few assessors complained about this fact. On the contrary, it was emphasised by our interviewees that the process designed by ANU was very time-efficient, albeit it was also mentioned that the process was therefore less thorough. It became obvious that the assessors were content because they were experienced in 'remote peer reviews' (from reviewing manuscripts and grant proposals). However, the variety of criteria applied in the assessments indicates that the isolation of assessors favoured the idiosyncrasies of individual judgements. Group discussions of assessors, which are likely to achieve a mutual adjustment of standards and criteria applied, were not part of this assessment process. Assessors were also isolated from the academics whose work they were evaluating. As a consequence, they had no opportunities to gather information or to discuss standards of evaluation. Our analysis showed that in some cases academics and their assessors had different understandings of research quality. The fact that ANU's academics partly disregarded second-order criteria such as publication types and citations indicates that they expected the assessors to read the publications and to judge their content rather than their attributes.

Another consequence of the taylorisation was that the *assessors worked inside a black box*. They were part of a process about which they were not fully informed. It becomes clear from their comments that they expected an integration process that took them into account in a synthesis of individual assessments. The assessors modelled a competent colleague who was able to use their contribution in preparing the final assessment — as is the case in the remote peer reviews of manuscripts and grant proposals — and shifted the responsibility for the use that was made of their judgements to this imaginary colleague. This was clearly at odds with the actual process, in which only explicit comments on the procedure and explicit evaluative judgements were taken into account after they had been standardised.

Standardisation is another effect of the taylorisation. The whole assessment process applied one approach, one set of criteria, and one aggregation procedure to all fields. Problems with the delineation

of subdisciplines, the limited use of criteria of 'international excellence', non-standard forms of output, and the availability of electronic versions occurred primarily in the 'non-journal disciplines'. The assessment process was developed on the basis of an implicit picture of a common natural science and 'journal discipline'.[11] However, this picture applies to less than half of Australian National University's research.

An important problem of the remote peer review that had nothing to do with the design of the process by ANU is the limited availability of electronic versions of publications, which suggests that the 'online version' of remote peer review is currently not feasible. Again, it becomes obvious that the process was modelled after the natural sciences, whose publications have a much higher electronic availability than those of other disciplines. And again it is important to notice that these disciplines do not provide the majority of ANU's output.

The assessors compensated for many of these problems by drawing on their context knowledge; that is, on knowledge about the ANU they had from former visits or from collaborative relationships. Thus, many judgements were made on the basis of information that the assessors did not receive from ANU but had anyway because they were chosen for their prior knowledge about ANU's research.

These effects mark significant differences between the remote peer review conducted by the ANU, on the one hand, and traditional on-site reviews of research organisations or comparative peer reviews such as the UK's Research Assessment Exercise, on the other hand.

Conclusions

The Australian National University designed a remote peer review process of its research under time pressure and with no blueprints of successful remote peer reviews available. Under these conditions, it submitted to the bureaucratic temptation of designing a taylorised process which alienated assessors from the overall assessment process. The assessors were content because the process proved to be very time-efficient, and because they were comfortable with the practice from their everyday work as assessors for journals and funding agencies. Nevertheless, the process produced several threats to the validity of peer reviews that do not exist in traditional collective peer reviews of organisations.

The three most important aspects of the process that counteracted these threats are the context knowledge of assessors about the ANU, the focus of the evaluation on prior research rather than the current potential of ANU's research, and the non-comparative character of the evaluation. We can safely assume that assessors in comparative evaluations cannot have the same context knowledge about all universities. Furthermore, evaluating the potential of an organisation is likely to require personal contacts between assessors and academics. The described process is therefore very unlikely to work for evaluations with these aims.

It is of course impossible to judge the validity of ANU's Quality Review, because an independent judgement of quality would be needed as a reference point. Furthermore, it is important to distinguish between problems that are inevitably caused by the 'remoteness' of the assessors, and problems that are simply caused by the specific design of ANU's Quality Review. To us, only one of the threats to the validity of the analysed process — the very limited availability of electronic versions of publications — is a necessary feature of the 'remoteness' of assessors. Remote peer reviews only provide the opportunity (and thus a temptation) to split the roles and to atomise the review work. It is equally possible to have committees who assess research collectively (interacting e.g. in Internet chat rooms) and design their own assessment procedure.

There are unavoidable limitations to a remote peer review of organisations. Communication among assessors and between assessors and academics will always need to rely on the Internet, and the assessors will not be able to see the work environment in the organisations. Since the potential of electronic communication channels was not fully exploited in the investigated process, the feasibility of remote peer reviews cannot be judged on the basis of this analysis. In order to assess these limitations (as they are indicated by Olson and Olson (2003), cf. the introductory paragraphs), a remote peer review that takes full advantage of the opportunities to support collective, interactive assessments must be analysed. Further experiments are necessary that link responsibility for the outcomes to assessments, take into account the specificity of fields, and support collective decision-making.

An important research theme that emerges from this analysis concerns the actual practices of peer review. The assessors' use of second-order criteria, while encouraged by ANU's approach, is by no means an artefact of this peer review process. Apparently, one of the responses of reviewers to the ever-increasing specialisation in science and to the review overload is to avoid judging scientific content by using second-order criteria. In this context, 'amateur bibliometrics' — the use of ISI's impact factors or of equally questionable raw citation counts — seems to become an important 'quick and dirty practice' of evaluations. These changes in peer practices must be analysed very carefully, because they may have consequences for our approach to peer review. They may lead to a situation in which bibliometrics involuntarily takes over because peers do not judge content anymore but apply amateur bibliometrics, and also do not act as a counterweight to bibliometric measures because, as amateur bibliometricians, they are likely to trust them blindly.

Notes

1. The negotiation process may of course also lead to the suppression of new knowledge. Editors, and sometimes reviewers as well, have veto positions in the negotiation process and may prevent new ideas from being published.
2. The Web of Science is produced by the Institute of Scientific Information (ISI), which is owned by Thomson Scientific. The Research Evaluation and Policy Project has established a database containing all publications listing at least one Australian address that appear in journals in the ISI's four main indices: the Science Citation Index (SCI); the Social Sciences Citation Index (SSCI); the Arts and Humanities Citation Index (A&HCI); and Current Contents (CC). See <http://repp.anu.edu.au/> for more information.
3. Cross-verification with the publication database of the ANU and the REPP database enabled clear allocations of publications to academics because of the additional address and publication information. Academics with uncommon names were selected for individual analyses in order to avoid the problem of homonyms.
4. The low number of interviews was due to both the time frame and the fact that the reviewers were indeed 'remote'. It was important to conduct the interviews as soon as possible after the review in order to secure some recollection by the interviewees. This imperative and the need to economise on travel costs led to a request for an interview between August and October 2004, which is the major holiday time in the northern hemisphere. More than 60 reviewers were approached, but only eight were available.
5. Our data differs from that published by ANU (2004b) for several reasons. While the Quality Review included only assessments that arrived before a certain deadline, our sample includes more reviewers and judgements. Since we were interested in comparisons of disciplines, we assigned publications that were submitted to several panels in each subdiscipline and therefore arrived at a total of 7,521 publications rather than the 'some 6000' (ANU, 2004b, page 27) that were actually nominated. Finally, our content analysis was qualitative rather than quantitative and cast a much wider net. In particular, we searched not only for explicit comments on the process but analysed them with regard to a wider set of variables, thereby taking into account partial occurrences as well. To give an example, we found 58 reviewers commenting on the limitations of their competence, while the analysis undertaken by ANU only found 13 reviewers stating that the work submitted to them 'lay outside their main area of expertise' (ANU, 2004b, page 31). Because of the different categories applied in the content analysis, both figures may be equally valid.
6. The definition of disciplines affected not only the review process but our analysis as well. Some definitions of disciplines and subdisciplines are uncommon (partly owing to the specific profile of ANU's research) and thus cannot easily be compared to established fields. For example, human geography, anthropology, demography, gender studies, social theory, applied sociology, and social policy were included in a field 'Studies in Human Society', while political sciences and history were excluded.
7. Numbers of publications and citations were not high enough for a similar analysis in Indigenous and Australian Studies to be conducted.
8. Numbers of nominated ISI publications were too low in the two other disciplines to enable this kind of comparison.
9. The interviews explored the practices of assessors when analysing the publications. Our interviewees described how they looked up citations, prepared lists of publication types, etc. Thus, the actual use of second-order criteria for judging publications can be established with some confidence.
10. This observation raises the issue of reviewer bias. We found very few examples of judgements that sounded biased because they consisted of sweeping positive comments without any reasoning given. Of course this does not prove that bias was not a problem. Our interviews demonstrated that the assessors were well aware of that danger and had thought about it. Our overall impression is that the assessors assumed a 'reviewer role' that implies impartiality. The subject matter of evaluation — organisational subunits rather than individuals — might also have counteracted biased assessments.
11. Our analysis of ANU's Quality Review suggests that the extension of institutions that are appropriate for the natural sciences to all disciplines occurs not only in science policy (Donovan 2003) but also in science management. This adaptation of institutional scripts that don't necessarily fit the task for which they are adapted but are legitimate in an organisation's environment is a recurrent theme in the neo-institutionalist literature of organisational sociology (Meyer and Rowan 1977; DiMaggio and Powell 1991).

References

ANU (2004a), *ANU: 'University with a difference': The Report of the Committee Established by the Council of The Australian National University to Evaluate the Quality of ANU performance* (Canberra, Australian National University).

ANU (2004b), *Capabilities and Performance Statement* (Canberra, Australian National University).

Daryl E. Chubin and Edward J. Hackett (1990), *Peerless Science: Peer Review and US Science Policy* (Albany, NY, State University of New York Press).

Domenic V. Cicchetti (1991), 'The reliability of peer review for manuscript and grant submissions: a cross-disciplinary investigation', *Behavioral and Brain Sciences*, 14, pages 119–135.

Stephen Cole, Jonathan R. Cole and Gary A. Simon (1981), 'Chance and consensus in peer review', *Science*, 214, pages 881–886.

Stephen Cole, Leonard Rubin and Jonathan R. Cole (1978), *Peer Review in the National Science Foundation. Phase One of a Study* (Washington, National Academy of Sciences).

Paul J. DiMaggio and Walter W. Powell (1991), 'The iron cage revisited: institutional isomorphism and collective rationality in organizational fields', in Walter W. Powell and Paul J. DiMaggio (eds.), *The New Institutionalism in Organizational Analysis* (Chicago, University of Chicago Press), pages 147–160.

Claire Donovan (2003), *Social Science in the Service of Science and Technology: A Case of Mistaken Identity within National Research Policy*. Paper presented at the New Times, New Worlds, New Ideas: Sociology Today and Tomorrow, TASA 2003 Conference, University of New England, Armidale, 4–6 December.

Jochen Gläser (2004), *Produzierende Gemeinschaften: Soziale Ordnung und kollektive Produktion (nicht nur) in scientific communities*. Manuscript.

Jochen Gläser and Grit Laudel (2004), *Experteninterviews und qualitative Inhaltsanalyse als Instrumente rekonstruierender Untersuchungen* (VS Verlag für Sozialwissenschaften).

Stefan Hirschauer (2004), 'Peer Review Verfahren auf dem Prüfstand: Zum Soziologiedefizit der Wissenschaftsevaluation', *Zeitschrift für Soziologie*, 33, pages 62–83.

Karin Knorr-Cetina (1981), *The Manufacture of Knowledge: An Essay on the Constructivist and Contextual Nature of Science* (Oxford, Pergamon Press).

Grit Laudel (1999), *Interdisziplinäre Forschungskooperation: Erfolgsbedingungen der Institution 'Sonderforschungsbereich'* (Berlin, Edition Sigma).

Renate Mayntz and Fritz W. Scharpf (1995), 'Der Ansatz des akteurzentrierten Institutionalismus', in Renate Mayntz and Fritz W. Scharpf (eds.), *Gesellschaftliche Selbstregelung und politische Steuerung* (Frankfurt a. M., Campus), pages 39–72.

John W. Meyer and Brian Rowan (1977), 'Institutionalized organizations: formal structure as myth and ceremony', *American Journal of Sociology*, 83, pages 340–363.

Greg Myers (1990), *Writing Biology: Texts and The Social Construction of Scientific Knowledge* (Madison, University of Wisconsin Press).

Gary M. Olson and Judith S. Olson (2003), 'Mitigating the effects of distance on collaborative intellectual work', *Economics of Innovation and New Technologies*, 12, pages 27–42.

Alan L. Porter, Daryl E. Chubin and Xiao-Yin Jin (1988), 'Citations and scientific progress: comparing bibliometric measures with scientist judgements', *Scientometrics*, 13, pages 103–124.

Fritz W. Scharpf (1997), *Games Real Actors Play: Actor-Centered Institutionalism in Policy Research* (Boulder, Westview Press).