Decision-making in expert panels evaluating research

Constraints, processes and bias

Liv Langfeldt

Dissertation submitted for the Dr. polit.-degree
The University of Oslo
The Faulty of Social Sciences
Autumn 2001

ISBN 82-7218-465-6
ISSN 0807-3635

GCS AS – Oslo - 2002

Preface

This thesis on decision-making in expert panels evaluating research has been part of research training programme at NIFU financed by the Research Council of Norway. The institute would like to thank Liv Langfeldt for the painstaking research that has gone into her doctoral dissertation and for the fact that it has not kept her from contributing generously to our professional environment. We would also like to extend our gratitude to the Research Council of Norway for the funding that has allowed this expansion of the knowledge basis in the field. Thanks also to her supervisers Professor Knut Midgaard of University of Oslo and Researcher Karl Erik Brofoss at NIFU.

Oslo, Juli 2002

Petter Aasen
Director

Randi Søgnen
Reseach Director

Preface

This study of decision-making in expert panels evaluating research is a product of a doctoral fellowship from the former research training programme 'Research on Research' at the Norwegian Institute for Studies in Research and Higher Education (NIFU), financed by the Research Council of Norway. With the exception of a one-semester sojourn as Visiting Scholar at the Department of Science & Technology Studies, Cornell University (Spring 1994), the work was carried out at NIFU. I have been affiliated with the Dr. polit.-program at the Department of Political Science, University of Oslo.

There are many persons I wish to thank. First of all I am indebted to my informants: the interviewed panel members and the secretaries of the panels who willingly sheared their experiences. Their co-operation and interest in the project has been invaluable for my learning about the decision-making process in expert panels that evaluate research.

My principal adviser for the dissertation work has been Professor Knut Midgaard at the Department of Political Science, University of Oslo. His thorough reading of drafts and penetrating comments and advice helped me to go deeper into the material. My adviser at NIFU, Researcher Karl Erik Brofoss, helped me design the project, shared his comprehensive insight in the field with me, and encouraged me to go on. I express my gratitude to both.

NIFU has been a fruitful work-place for the project, with an excellent library and inspiring and insightful colleagues and foreign guests that at various stages of the work have taken the time to discuss the project with me. My sojourn at Cornell likewise brought new dimensions to the project. My drafts have been discussed in various forums, foremost the 'Research on Research' colloquia and other NIFU-meetings, and the Norwegian Political Science Conferences. Again, I express my gratitude.

Furthermore, I thank NIFU, in the first place for offering me the Fellowship, and in the second place for giving me the opportunity to finalise the project, which proved too ambitious for a three-year fellowship.

Thanks are due to John G. Taylor for his proof-reading of the final manuscript.

Finally, a special thank to Torgeir, the political scientist with whom I share my life, for patiently being my scholarly sparring partner throughout the process.

Oslo, November 2001

Liv Langfeldt

Contents

Tables

Figures

1 Introduction

This study deals with the constraints on, processes in and bias of expert panels evaluating research quality and research priorities. The kinds of evaluations studied are expert panel evaluations of research fields, research programmes and research institutions. How do such evaluation panels work? How do they know, or find, the answers to such questions as what is good and worthwhile research, promising research areas and good research groups? How do we know that the evaluations are properly done and that the conclusions are not biased? Do the panel/group setting and the organisational context influence the assessments?

These issues are analysed broadly, including discussions of such different problems as the implications of different ontological views on 'good research', and implications of group effects and of various constellations of interests on the panels. Six different evaluation processes are studied to provide an insight into what influences expert panels' conclusions. Central findings are that there is little overlapping competence on the panels, a high degree of task division and that the composition of an expert panel, the organisation of its work, and the (lack of) group interaction may be decisive for the conclusions of the evaluation. Moreover, 'bias' is found to have many different meanings, and not all kinds of bias in research evaluation are necessarily unacceptable. With regard to the relation between the task and the capacity of the panels, there seems to be a serious disparity between the capacity and resources of actual evaluations and the demands that ideally should be met when judging scholarly quality.

This chapter deals with the 'roots' and foundation of the study. Section 1.1 deals with the background and paths leading to interest in the topic. Section 1.2 gives a conceptual introduction to research evaluation, and contrasts expert panel evaluations of research fields, research programmes and research institutions with the traditional peer review systems. Section 1.3 points out the research questions and approaches of the study, while Section 1.4 discusses methods and data sources.

1.1 Why a monograph on the decision-making aspects of evaluation of research?

The interests in the decision-making of expert panel evaluation are manifold. From the point of view of studies of public policy and decision-making, research evaluations may illustrate both trends and problems in research policy. Evaluations in various forms have become central in public policy. Large parts of scholarly research are publicly funded and state agencies are increasingly concerned with setting priorities for allocations of funds. From the point of view of gaining insight into the central characteristics of the scientific community and the borderline between politics and science,[1] research evaluations are particularly interesting. Evaluations of research are a critical and 'politically sensitive' kind of activity for the scientific community, a kind of activity which may pinpoint general characteristics that are not as visible in its more day-to-day activities, and a promising setting for studying decision-making processes on the borderline between science and politics.

1.1.1 Evaluation in public policy

Routines and practices for planning, gathering information, learning and control within bureaucracies have varied both geographically and historically. In the post-WWII period evaluation has become a central concept. The use of public resources and the effect of public initiatives are to an increasing degree expected to be evaluated as a matter of routine, and the more special or important cases are subject to separate, often 'external' evaluation. Social scientists do 'evaluation research' for public administration (Albæk 1988; Weiss 1972), consultancy firms are contracted to evaluate organisational efficiency and expert committees are appointed to undertake various kinds of evaluations. Evaluation research grew at a 'meteoric rate' in Western Europe and North America during the mid-sixties and seventies (Hellstern 1986:279). Means and effects of government programmes to achieve social change were assessed. Since the late seventies and early eighties, with economic constraints, the focus turned to assessing utility and costs of public expenditure, and development of account-

[1] 'Science' and 'scientific' are in this study used in its general sense ('Wissenschaft') and not restricted to the natural sciences.

ability and quality control procedures. From the mid-eighties evaluation has been more integrated into administrative routines and a variety of approaches are used (Hellstern 1986:305). It should be noted that in addition to general and acknowledged purposes of public policy evaluations – such as accountability, improvement and increased knowledge – strategic motivations may play a central role. Such motivations may include 'to gain time, to show up a front of rationality, and to disseminate an overly handsome view of the executives' work' (Vedung 1997:111).

Emphasis on evaluation also affects the research sector (Rip 1990). During the eighties and the nineties we have seen a large number of evaluations of research programmes, research institutions and research fields/disciplines in Norway as in other OECD countries. Various modes and combinations of expert panels, organisational studies and bibliometric analyses[2] are used (Hansen & Jørgensen 1995). Practices vary both according to discipline and nationally. Also within the Nordic countries practices are manifold (Christiansen & Christiansen 1989).

This increase in evaluations of research succeeded the growth in the public finances spent on research, the channelling of such resources through research programmes,[3] and an increased 'awareness of the consequentiality of a greatly-expanded science' (Cozzens 1990:282, see also OECD 1987). Both research programmes and 'externally' initiated research evaluations are ways of public authorities to communicate with and control the research communities. Programmes are ways of allocating resources and of setting research priorities, while evaluations control characteristics of the research activities and output (quality, productivity, goal attainments etc.), and are also potentially a way of learning (e.g. about effects of policy measures).

The more formal side of evaluation initiatives should also be noted. The ideas of 'Management by Objectives' which have influ-

---

2 Including publication counts, citation indexes, journal impact factors etc.

3 'Research programme' in this study refers to a time limited research effort in a particular research or problem area, organised as a grant programme by a research council. Other meanings of the term (not used in this study) are ad hoc 'departments' at universities (sometimes also named centres) and schools/directions of a research field, e.g. the so-called 'strong programme' in the sociology of scientific knowledge.

enced public policy demand control and monitoring and increases the need for evaluations. Agencies are responsible for showing how public means given as block grants are spend. In Norway, focus on systematic evaluation is now required by Government decision (*'Økonomireglement for staten'*, Royal Decree of 26 January 1996).

Within the research sector, the demand for evaluations has been met with extended use of the traditional way of evaluating research quality: peer review.

1.1.2 The decision making aspects of expert panel evaluation

Evaluation by researchers competent in the field under review (peer review) is the predominant method used for research evaluations, and normally seen as an 'unavoidable' method which cannot be replaced by quantitative methods:

> *'quantitative measures cannot form an alternative method to peer review. Indeed, reference to peer judgement is necessary to develop and test their usefulness in the first place. For example, in constructing influence measures, only the peer community can identify which are the most significant journals in that field; determining what should be considered to be 'measures of esteem' is also dependent on peer judgement. In addition, peer-judgement is needed to interpret bibliometric and other data in individual cases' (ABRC 1990:13).*

Within the scientific community peer review is normally seen as the only legitimate method for valuing scientific quality. To assess the quality of scientific research one has to be a 'peer' of the researcher under review (i.e. an expert in the area). At the same time this method is being questioned. It is claimed that peer review is partial, biased and unreliable, and it takes time away from research activities[4] (Chubin & Hackett 1990; Wood 1997; Campanario 1998a and 1998b; Langfeldt 2001b). I.e. the growth of the research sector and the increasing emphasises on evaluation and control have lead to both broader reliance on evaluation of research by peers *and* growing scepticism towards the objectivity of this kind of evaluation.

This is a particularly interesting context for the study of situations where science and politics meet, and points to a need for studying the characteristics of expert panel evaluation of research. Theory-building and research that might uncover central mechanisms are particularly

---

[4] The UK research councils have estimated that the time used for their peer review processes amounted to 115 researcher work-years annually (ABRC 1990:38).

important if we want to understand the more complex aspects of expert panel evaluation, like the 'scholarly' constitutive versus 'politically' contingent bases of peer judgements, the group dynamics of peer panels, or more generally the situational constraints of evaluation processes (see Chubin & Hackett 1990:47). Hence this monograph is devoted to the decision-making aspects of expert panel evaluation of research. Expert panel evaluation as a research policy instrument is studied by focusing mainly on the basis of peer judgements and the decision-making of expert panels. The overarching research question is: *What affects decision-making processes when research programmes, research institutions and research fields are evaluated by expert panels?* More specifically, the influence of structural and organisational characteristics of such evaluation processes on the content of the evaluation reports, is analysed. Throughout, the focus is on peer judgements on research quality and research priorities. Expert panel evaluation may also encompass judgements on other aspects of the units under review, such as organisation, users' opinions and market adaptability. Judgements on such aspects are not studied here as they do not demand peer judgement. (Other kinds of expert analysis and judgements may be just as legitimate.)

1.2 Peer review and research evaluation: concepts and development

1.2.1 Peer review and related concepts

> *'Peer review is the name given to judgements of scientific merit by other scientists working in, or close to the field in question. Peer review is premised upon the assumption that a judgement about certain aspects of science, for example its quality, is an expert decision capable of being made only by those who are sufficiently knowledgeable about the cognitive development of the field, its research agendas and the practitioners within it' (OECD 1987:28).*

This OECD-report further makes a distinction between direct peer review and modified direct peer review. Direct peer review is 'carried out specifically for the purpose of determining and confined to

questions of scientific merit', while modified direct peer review addresses a broader range of questions (OECD 1987:28).[5]

Expert evaluations of research programmes, institutions and fields are normally modified peer review; in addition to scientific merit they may concern working conditions of the researchers and other organisational questions, socio-economic impact of the research or potential for utilisation of the results. Such evaluations may also be 'modified' peer review with regard to the expertise of the evaluators. When the evaluation not only addresses questions of scientific merit, not all members of the evaluation panels are necessarily 'peers' or active researchers in the relevant fields.

I will reserve the term '*peer review*' for the more traditional review systems of scholarly communities (e.g. manuscript reviews for scholarly journals, review of applications for academic positions, and review of grant applications to research councils). The term '*expert panel evaluations*' will be used for the kind of evaluations under study – research council commissioned (often ad hoc) reviews on the institutional, program or discipline level. '*Peer evaluation*' denotes such evaluations when the panel consists of researchers qualified in the area under review. In cases where expert panels consist of both 'peers' and other experts I refer to '*mixed panel evaluations*'.

Expert panel evaluations of research can be seen as the result of the meeting of traditional (micro-level) peer review with the growth of, and demand for evaluation in public policy. In contrast to traditional peer review it aims at assessments of research on the meso-level (the institutional level) and the macro-level (the national level), whereas traditional peer review makes assessments at the micro-level (single manuscripts, applications or applicants).

1.2.2 The origin and development of referee systems

The origin of peer review can be traced back to the appearance of scientific journals in the seventeenth century. As scientific societies started to organise communications of discoveries, various review or censor mechanisms developed. Such reviews served a number of

---

[5] A more common term for modified direct peer review is merit review. The OECD-report also deals with *indirect peer review*, which uses 'historic peer review judgements made primarily for purposes other than the evaluation in question' (OECD 1987:28), and includes bibliometrics.

missions. They gave the initiating 'societies' a measure of control over the contents of their publications, they motivated to raise the standards of submitted manuscripts and they 'certified' the manuscripts and gave faith to their contents (Zuckerman & Merton 1971). Until after World War II, there was no general 'movement' to adopt formal peer review practices. Practices were adopted piecemeal and independently in each journal in response to idiosyncratic conditions (Burnham 1992).

In the post-WWII period, journal peer review has become more formalised, e.g. 'double-blind review'[6] of all submitted manuscripts by two or more reviewers. Research councils have formalised procedures for assessing grant applications, including standing or ad hoc panels of experts in the field, mail review by individual experts, or both panel and mail review. Both journal peer review and grant peer review processes are now more or less regularly subjected to studies, debates and refinements (for example Chubin & Hackett 1990; Speck 1993; Fisher et al. 1994; GAO 1994; Garfunkel et al. 1994; Laband & Piette 1994; Nylenna et al. 1994; NIH 1996; NSF 1996; Wood 1995 and 1997; Campanario 1998a and 1998b).

As already mentioned, evaluation by peers is initiated and also used by actors outside the research community. In the following, we look at this new kind of evaluations in contrast to the more traditional forms of peer review.

1.2.3 New kinds of research evaluation vs. traditional peer review

Central characteristics of non-researcher initiated macro- and meso-level expert panel evaluations of research are displayed by contrasting these with traditional micro-level peer review. Both purposes and contexts vary considerably between the two. Macro-level expert evaluations are normally aimed at decision-making processes outside the research community, whereas micro-level peer review may be part of the research process itself. As mentioned, the referee systems of scientific journals, for instance, serve functions as improving the

---

[6] Double blind review means that the author does not know the identity of the reviewer and the reviewer does not know that of the author. A substantial number of scholarly journals also practice simple blind review – the reviewers are not known to the authors, but the authors are known to the reviewers.

manuscripts and providing confidence in the research results (Zuckerman & Merton 1971).

By further contrasting the new forms of expert panel evaluation with traditional peer review, we easily reach the conclusion that the former differs from the latter in various ways that indicate that it will produce *vaguer or more positive* judgements:

- *Firstly*, expert panel evaluations of programmes, institutions and fields are expected to be instruments of national research policy. This purpose – external to the research community – may make peer evaluators insecure and sceptical about the potential use of their evaluation reports and make them very careful about what to write and what not to write when the reports are put together. Paradoxically this may apply especially if the purpose is vaguely stated. If the evaluators do not get a clear answer to what the evaluation will be used for, they will either presume that it is purely ritual – in which case they do not need to do a thorough evaluation – or they will suspect that an evaluation may yield so serious implications that the commissioning body refuses to inform them about the purpose (as the commissioning body fears such information would scare peer evaluators from participating in such an evaluation). In the latter case a peer evaluator with 'decent' loyalty to his/hers colleagues is likely to be very careful not to put anything into print that may be of disadvantage to the evaluees (i.e. the researchers/research units being evaluated).[7]

- *Other* important differences between micro-level peer review and evaluations of programmes, institutions and fields, are the fact that the reviewers are identifiable, the public status of the reports and the scope of the evaluations. While micro-level reviewers are mostly anonymous, confidential and limited to one project or manuscript, macro-level reviewers are visible to those being evaluated, the reports are normally public and encompass a large number of projects and publications. These are all characteristics that may lead to less outspoken and more vague reports. Visible evaluators writing public reports are made personally accountable

[7] A Nordic interview study of evaluees found that scientists criticised in public evaluation reports were 'met with decreased status and reputation and suffered adverse psychological impacts' (Luukkonen 1995:364).

for the content of the reports, and are therefore likely to be more cautious about what they put into print (Grigson & Stokes 1993:176). The large scope of the evaluation means that the review will be more superficial.[8]

- *Moreover*, the character of the 'good' that is distributed is substantially different. In manuscript review and grant review, scarce goods – in the form of journal space or money – are distributed. The evaluees are marked and ranked, and losers and winners are identified. It is a zero-sum-game, whereas the kind of evaluations studied here is not. All evaluees may, in principle, be just as happy or unhappy with the content of the report: no winners or losers have to be identified in these evaluation reports. The evaluators distribute comments, honour, advice and recommendations for future research and investments, 'goods' that may be seen as non-limited. It should be noted that practices vary among countries. The argument above only applies when the evaluations do not explicitly compare and rate the evaluated units and results are not directly linked to allocations of resources. For the UK practice of rating departments (see Hansen & Jørgensen 1995), for instance, this argument does not apply. Another point is that it might be *unclear* what kind of good is distributed (refer the first point above). Evaluations may, for instance, provide good arguments for more resources or better research conditions for all evaluees and the situation may be perceived, not as a zero-sum game, but as a variable-sum game (more resources for research will gain the society).

- *Finally*, the panels are expected to write a report that they all agree upon. When there are different views on the evaluation panel and no ranking of the evaluated units has to be made, compromises will often mean vague formulations.

---

[8] It is not realistic to opt for macro-evaluations with the same possibilities for thoroughness as micro-evaluations, e.g. a discipline evaluation spending the time used for a manuscript review multiplied by the number of manuscripts produced in the discipline during (e.g.) the last five years.

Table 1.1 *Micro-level versus macro-level evaluations of research*

	Micro-level (peer review)	**Macro- and meso-level** (expert panel evaluation)
Object of evaluation	One researcher, one manuscript or one application is assessed separately and ranked or graded.	One ore more research programmes, institutions or disciplines.
Goods distributed	Scarce goods/zero-sum	Non-limited goods/unclear what goods are distributed
Decision-making arena	Intra-scientific	Extra-scientific
Public reports?	Seldom public	Mostly public
The evaluation process and the anonymity of the evaluators	Manuscript referees write separate reports and are normally anonymous. Committees for screening applicants to chairs (in Norway) write common reports, and are not anonymous. Practice for grant proposal reviews vary: anonymous individual mail review and/or non-anonymous panel review.	The evaluators write one common, unanimous report, which are never (or seldom) anonymous.

Table 1.1 gives an overview of central differences between traditional peer review and expert panel evaluation of research on the meso- or macro-level. To some degree the characteristics are simplified, micro-level peer review encompasses a variety of review processes for a variety of purposes and to varying degree may be directed solely at 'intra-scientific' decision-making. Macro- and meso-level evaluations are less institutionalised and might vary even more, especially between countries. The content of the table is a compromise between setting up characteristics valid for Norwegian expert panel evaluations, and characteristics with more general validity. Expert panel evaluation may be directly associated with the allocation of scarce resources, evaluators may be anonymous, and some evaluation reports will be confidential. In Norway, though, such evaluations do not (hitherto) distribute scarce resources, are never anonymous, and always unrestricted.

1.3 Research questions and approaches of the study

As there is little research on decision-making in research evaluation, there is no sound basis for very specific research questions or hypothesis. Yet, as partly demonstrated in the previous section, social studies of science and studies of peer review can help us to see what may affect evaluation processes. Chapters 2 and 3 deal with the relevant literature and outline a theoretical frame for the study. The overall research questions addressed are open-ended and explorative and try to cover the central aspects of the context of the decision-making:

- Are there neutral criteria of good research? What does 'unbiased' evaluation imply? Is it attainable, and is it definitely desirable?
- How is good research identified, and what 'professional' and social norms affect judgements?
- How may the group setting of expert panels evaluating research affect the outcome?
- How may the commissioner (of the evaluation) influence the outcome through organisational means, e.g. panel composition, terms of reference and signals about the planned use of the evaluation report?

Based on the theoretical groundwork of the previous chapters, Chapters 4 and 5 present, analyse and compare six evaluation processes. Questions here are more concrete and deal with how the research evaluations were conducted and what influenced the evaluators and their evaluation reports. The effects of organisational design and other structural constraints, as well as cognitive bias, personal and scientific interests are considered.

These are the questions deemed the central ones for more extensive insight in the context and mechanisms of research evaluations by expert panels – questions that previous studies have not confronted, or pointed to as the challenges of future research (see Chubin & Hackett 1990). To understand the decision-making processes in question we need to look both at the micro-level contexts of

group decision-making and at how the institutional environment is likely to affect the process. Focus on a broad group of factors in a detailed analysis of concrete situations is the proper basis, furthermore, for an explorative study.

The approach chosen for the theoretical discussions in Chapters 2 and 3 is open-minded with regard to the various 'paradigms' in political science. There is no encompassing theory within the social sciences aimed at understanding complex decision-making settings and exposing the central mechanisms of social processes. The discussion draws on literature from various disciplines and 'schools', and different theoretical 'tools' are seen as complementary rather than competing. This also includes such traditionally 'antagonistic' approaches as norm-driven or interest-driven actors as the explanation of human actions.

The empirical analysis tries to uncover the actors' own understandings of their motives. Informants may very well not provide honest answers when interviewers try to find out, for example, whether particular actions were driven by self-interest or institutional norms. This should not stop us from trying to answer such questions. With multiple data sources and opening up for a wide rage of reasons and explanations, we should be able to provide a 'thick description' giving good insight in the phenomenon studied. A thick description allows us to consider all relevant contextual factors and include the various actors' 'reasons or rationales for acting the way they do in their own situations' (Farr 1985:1090), and may combine theoretical approaches as 'rational actors' and 'social norms'.

Rationality approaches and social norm approaches aim at explaining different kinds of actions or might be seen as explaining different aspects of actions. Game theory is 'formal' theory for understanding decision-making situations as such, whereas organisational and institutional theory allows 'substantial' hypotheses on the constraints on the situation/the actors. Organisational and institutional theory may, therefore, very well serve as inputs to models set up by rational theories, and the two approaches may easily be combined for the purpose of empirical study.

Still, the fundamental underlying ideas of the approaches have traditionally been seen as conflicting. In sociological institutionalism actors' preferences are formed by institutions, while in rational choice theories preferences are exogenously given, generally as the actors self-interests. Problems may arise in empirical research, when different

causes lead to the *same action*. For example, scholars may conform to a norm of scholarly communities, saying that they should do a literature review before writing an article, (1) because they know it is to their own advantage – they know the article will be reviewed by referees with good knowledge of previous research, and in order to get the article published and avoid a 'slaughter' by the referees they need to undertake a literature review (self-interest). Another reason for complying to the norm may be that (2) the scholar *thinks* it is to her/his advantage, erroneously believing that a lack of a literature review will be detected (self-interest and constraints on information/bounded rationality). Yet another reason for the same action may be that (3) the norm is internalised, it is 'the thing to do', and the scholar wants to do it regardless of the chances for getting the article published, and without the personal interest of gaining time by neglecting the literature review being even considered (norm oriented/non-outcome oriented action).

If we want to explain such phenomena as to why scholars undertake literature reviews, we need some evidence (or 'theory') for choosing between the theories.[9] One method, is to argue that human actions are partly norm-driven (and non-outcome oriented), and partly interest-driven and outcome-oriented, and trying to integrate the two theories in some way as done by Elster 1989, for example. A combination of interests and norms seems a reasonable supposition, but it does not help us to point out the motives of particular actors in particular settings. Dealing (partly) with other approaches, Knott & Miller promote a 'parsimonious theory' for consolidating theories of 'cognitive limits' and 'bureaucratic incentives', a theory that 'shows how a single set of assumptions can lead to the kind of behaviors predicted by *both* models' – explaining 'why organizations sometimes create incentives that lead individuals to remain ignorant, use biased information, and satisfice' (1987:179). A theory that integrates funda-

mentally *different motives* is far more difficult. Such a theory also needs a 'tool' to separate situations where norms are followed because of cost-benefit calculations from situations where the norm is internalised. Motives are substantial for understanding human action, and good social analysis should be able to distinguish norms and interests as motives, even when resulting in the same action.

One way of resolving such questions is to make priorities between the explanations based on what seems the most normal motive. March and Olsen (1989:23–24) for example, think that norm-driven ('obligatory') action 'seems more to describe action' in institutional settings than outcome oriented ('anticipatory') action. In the case of studying a *specific* action or decision, such suggestions are of little help, and thick descriptions, the empirical way out chosen here, are more recommendable – especially for explorative studies.[10]

1.4 Methods, selection of cases and data sources

1.4.1 The research design and strategy

As mentioned, the research questions are explorative. They point to a design allowing thick descriptions of specific decision-making processes to uncover central mechanisms of expert panel evaluations of research. This entails an intensive research design, that is, to study many variables and few cases. One unique strength of such case studies is the ability to draw on several different data sources to

---

[9] I concentrate on interest-driven versus norm-driven behaviour as the question of bounded rationality (whether actors have full information of consequences or not) does not really matter for answering research questions concerning the *motives* of the actors. In many cases it can be taken for granted that the kind of actors in question generally base their decisions on uncertain information. Alternatively, it may be possible to *answer* the question of faulty information. The question of detecting lacking literature reviews may, for instance, be answered by empirical study.

[10] Empirical analysis may also find that the most plausible explanation includes much more complex reasons than self-interests or norms. That distinction may as well turn out to be trivial or irrelevant.

analyse the same event, e.g. archives, interviews with actors, field research and direct observation (Yin 1989:20). A major disadvantage is the limited possibilities for generalising the results when studying only a few cases. To study one evaluation process may, for instance, have very limited value with regard to drawing conclusions about more general reasons and mechanisms. Extensive designs,[11] on the other hand, allow more conclusive generalisations of the results, but are not good at answering 'why' questions in a little researched area (see Yin 1989).[12]

A multiple-case design may, however, provide a more general insight than separate studies of single cases. Multiple-case studies aim at conclusions which may be generalised, but which are based on a different logic than the sampling logic of extensive research designs (Yin 1989:52–59). They use a comparative logic when selecting cases fit to answer the specific research questions.

There are two major comparative research strategies. One strategy is to hold all variables constant except the (two or more) variables that are being tested to find any relation/covariation (called method of difference or most-similar systems design). The other strategy selects cases which are similar regarding the variables that are tested to find a relation/covariation, but maximise differences on other variables (called method of agreements or most-different systems design). A third possibility is a mixed strategy: some variation and some similarities in all (relevant) variables, allowing both kinds of comparisons. According to Frendreis a mixed strategy is superior with respect to

---

[11] That is, a quantitative study: a study of many cases, but a restricted, predefined set of variables.

[12] How a quantitative study may yield meaningful answers to the research question set out in Section 1.3, is hard to imagine.

generability,[13] it is more feasible and adaptable to actual research settings and it allows a good basis for comparisons even when variation in the interesting variables cannot be correctly estimated *a priori* (Frendreis 1983:268). In Yin (1989) we find arguments for a similar strategy, combining what is here called 'literal replication' and 'theoretical replication'.[14] The logic of literal replication is used within groups of similar cases, while 'theoretical replication' may be used across different groups of (similar) cases. In this way multiple-case studies may provide a broader spectrum of empirical evidence.

For our research questions, there is little previous research which points to predefined theories or promising hypotheses, and consequently not a limited set of factors to be tested. Furthermore, the research evaluations in question *vary* with regard to a large number of factors. The composition of the panels vary: they may consist of only peers, or both peers and users, they might have Nordic panel members or a broader international representation. The kind of disciplines involved differ, and evaluations may encompass one discipline or be multi-disciplinary, they may evaluate basic research, applied research, or both, and the units being evaluated might be programmes, institutions or research fields. In addition, there are (i.e. in 1992) five different Norwegian research councils which may organise such evaluations in very different ways. As spelled out above, there is very little previous research to tell us which of these factors are the most

---

[13] Frendreis' argument for the better generability of mixed strategies is that they allow the researcher to select a more representative sample of cases than do the most similar or the most different systems. In the two latter systems, cases are systematically excluded either to hold the dependent or the independent variables constant. This is not needed in a mixed system, which therefore allows a more representative sample of cases. To me, this seems a weak argument based partly mistakenly on the sampling logic of quantitative designs. When a mixed strategy allows more generability than another comparative design, this would not be because of its similarities to a 'universe' of cases, but rather because it allows the researcher to test out competing hypotheses, for example, by using *both* the logic of most similar and most different systems, and to check for all kinds of potentially deviant cases (as, explained by Frendreis, no particular kind of 'real life' cases needs to be systematically excluded).

[14] These concepts are somehow related to most similar and most different systems, but not well defined. Literal replications are cases which yield the same 'results', while theoretical replications yield 'contrary results but for predicable reasons' (Yin 1989:53).

crucial to analyse, and a design opting for similar *and* different cases to control for all such factors is hardly feasible – at least not as a one-researcher non-life-time project. One way to tackle this would be to select cases that vary considerably regarding the mentioned factors, and then, when analysing the data, to look for interesting characteristics and mechanisms common to all cases. This strategy suits a central object of the present study, to understand what characterises expert panel evaluations of research, as such, regardless of varying contexts. In contrast to the normally prescribed comparative designs, this approach does not test specific hypotheses, and if we do not find any characteristics or phenomena common to all cases, the comparative design has not been much help with regard to generability.[15] As an extra 'chance' for conclusive data, we should then also seek some similarities in central factors, that is, a mixed strategy which allows using both the logic of the most similar and the most different systems.

What number of cases is needed for such a strategy? According to Yin (1989:58) the required number of different cases depends on the number of external conditions that produce variation in the phenomenon studied, and then within each different kind of case, a minimum of two or three similar cases is needed. As the phenomena to be studied are rather complex, and there is a substantial number of 'external' conditions that may affect them, a considerable number of cases seems to be needed. The need for and possibility of variation on central factors is discussed below.

Studying variations between fields of learning

Do central aspects of evaluation processes vary between fields of learning? If so, in order to draw conclusions about differences, a solid basis for studying each discipline is needed. A design for studying disciplinary differences of expert panel evaluation of research including all research disciplines would be a vast project. If we take for granted that differences, if any, follow a simple distinctions between the humanities and social sciences on the one hand, and natural, medical and technical sciences on the other, the task is much more manageable. Previous studies of disciplinary differences may

---

[15] While the prescribed designs always provide some substantial conclusions either by confirming or refuting a theory based hypothesis.

also be used as a basis for interpreting results and strengthening conclusions (Whitley 1984; Becher 1989; Gulbrandsen & Langfeldt 1997).

A study limited to gross differences is also more realistic given the limitations of finding a proper empirical basis for reliable conclusions on disciplinary differences. When restricted to Norwegian cases suited for study expert panels' decision-making at the time of data collection for the present study, there are far from enough cases available for studying the different disciplines. The older the cases, the more limited details interview data may give. As interviews will frequently be the only sources for data on the decision-making processes in the panels, I excluded evaluation reports which were more than three year old from my list of suitable cases. Among the evaluations remaining, there were very few cases in some disciplines, and in the 'hard' natural sciences like physics and mathematics, there where no evaluations at all. However, the final choices represent a broad rage of fields of learning. Fields in the humanities, natural, technical, social and medical sciences are included (see below).

The various kinds of research units subjected to evaluation

The purpose of evaluation may vary between the various kinds of macro- and meso-level expert panel evaluations of research.[16] Evaluating temporary 'units' such as programmes may, for instance, include advising on whether to continue or terminate the programme, while evaluations of university departments may aim at ranking and distributing status to the departments. Aims may also be multiple, and vary for the different actors. Evaluation reports may not give very much information on such purposes, and a selection of cases on the bases of such aims is consequently difficult, particularly if the desired criteria for selection are how the evaluators or the evaluees perceive the purpose of the evaluation. When searching for cases for this study no evaluations supposed to provide information for specific budget cuts, reorganisations or similar, were found. All the evaluations studied therefore have rather general (official) aims. Going in more detail, the aims still differed between the evaluations of fields, institutions and programmes.

---

[16] The contrast to the specific micro-level decisions at which grant reviews and journal refereeing aim, is provided in Section 1.2.

The evaluations of research *fields* were intended to give a broad overview of the standing of the field in an international perspective, and provide general information and documentation for the research councils which commissioned them. The main purpose was to learn what could be done to improve the quality of the research in the research areas. The evaluations of research *institutions* asked whether the institutions performed their tasks satisfactorily, and did not, at least not explicitly, aim at ranking institutions nor at providing recommendations for reorganisation processes. The evaluations of research *programmes* studied, indirectly asked whether it was worth investing more money in the programmes. The main reason for the evaluations was that the government required that the programs should be evaluated as they had invested a large amount of money in them. The research councils that organised the evaluations were eager to assure the public that it was worth investing money in the programmes. The cases were selected with the expectation that the different purposes of these three kinds of evaluations may help us understand differences in evaluation process and decision-making.

Conclusions of the evaluation reports

The outcome of the decision-making of the expert panels, that is, the written report to the commissioning research council, may be seen as the dependent variable of the study. What affects central characteristics of the report, such as the explicitness of the assessments, and critical or praising conclusions? To study such questions, evaluations with varying degrees of explicitness and 'positive' or 'negative' assessments were selected. With hindsight, more variation in the conclusions of the selected evaluation reports would have been

preferable. However, the cases available at the time had a limited range of variation on such factors.[17]

Strategy: focus on similarities, 'controlling' for differences and developing ideal types
The cases studied include both similar and different evaluations with regard to design, approach and purpose, and they deal with a wide range of academic disciplines (see next section). What they all have in common is that they in some way are ad hoc evaluations. They were all the first evaluation of their kind of the specific field, program or institution, which means both that no routines for evaluations were established and that the implementation and potential effects of the evaluation were uncertain.

Both the limited number of cases and the ad hoc character of the cases imply that systematic differences found between variables may have limited claim to validity for other cases than those studied. It may be difficult to make substantial general conclusions of the nature of for instance evaluations in the humanities in comparison to evaluations in engineering, or evaluations of institutions in comparison to evaluations of research fields. What these different cases turn out to have in common, however, should yield good basis for conclusions on general features of the making of public ad hoc evaluations of research by expert panels within the research areas covered by the study (as mentioned, the study does not cover 'hard' natural sciences like mathematics or physics).

---

[17] The prevalence of positive and vague conclusions was explained in Section 1.2. One report with conclusions differing from those chosen appeared at the end of 1992 after I had made the final choice of cases and commenced data collection: '*Informatikk*: Research and Teaching in Norway. A Critical Evaluation' (NAVF 1992). This evaluation makes clearer judgements on the evaluees and picks 'winners and losers' to a greater extent than the evaluations which are part of the present study. The information contained in the evaluation report and the reactions from the evaluees contained in the 'hearing documents' form valuable data on a deviant case. The evaluation of Norwegian work research (NORAS 1992) is another example of a rather critical evaluation of Norwegian research. However, as this evaluation was conducted by one expert and not a panel, it does not deal with the kind of decision-making studied here. In recent years, more evaluation reports clearly deviating from those studied here, have appeared, e.g. 'Physics research at Norwegian universities, colleges and research institutes' (The Research Council of Norway 2000). See also Section 7.2.

Consequently, the main analytic strategy is to describe and explain common characteristics and mechanisms. In addition, a 'mixed' comparative strategy should allow (more tentative) conclusion also about differences between various kinds of evaluations.[18] Furthermore, to summarise the theoretical insights gained, ideal types extracting central factors and relations are developed. These are analytical constructs to pinpoint the logic and mechanisms of expert panel evaluations of research. This additional way of presenting conclusions from the study serve both analytical and communicational purposes. Ideal types are pure and extreme cases, without direct basis in (more complex) real life situations.[19] The ideal types, which extract theory from the empirical findings expressed in a purified abstract form, demand simplifications of contexts and relations, and should make conclusions clearer to the reader.

1.4.2 The cases

The study is based on analysis of six ad hoc panels (of mostly non-Norwegian experts) appointed by various research councils to evaluate research in Norway at the end of the eighties and the beginning of the nineties. There are two evaluations of research fields, two of institutes and two of programmes.[20]

- Two of the cases are evaluations of research *fields* – one from the humanities and one from the natural sciences. These are mainly evaluations of basic research at university departments conducted by international peer panels.
- Two other cases are evaluations of research *institutions* – one of an engineering research institute and one of three social science

---

[18] As explained in the first part of this section.

[19] 'An ideal type is formed by the one-sided accentuation of one or more points of view and by the synthesis of a great many diffuse, discrete, more or less present and occasionally absent concrete individual phenomena, which are arranged according to those one-sidedly emphasised viewpoints into a unified analytical construct.' 'This procedure can be indispensable for heuristic as well as expository purposes' (Weber 1949:90).

[20] Because of confidential data-material the cases are presented anonymously. See below.

institutes. These are evaluations of mainly applied research, conducted by Nordic experts. The panels were mixed, i.e. consisting of both researchers in relevant fields and sector representatives/potential users, and evaluated both the applicability/relevance and the scholarly quality of the research.

- The two last cases are evaluations of research *programmes/priority areas* – one natural science programme and one multidisciplinary programme, including applied, strategic and some basic research. One of them was conducted by an international peer panel, the other by a Nordic peer panel.

Altogether, a broad variety of fields of learning are represented. Half of the cases include either natural or medical sciences, two include engineering, and two include social sciences, one includes the humanities.[21] Of the six evaluation panels, four are peer panels, two are mixed panels, three are international panels and three are Nordic panels. Four different initiators/commissioning research councils are covered by the cases. One of the evaluation reports has clearly praising conclusions, two are more moderately praising, two are both praising and critical, and one has rather vague conclusions. Details about the characteristics of the various cases are found in Tables 5.2 to 5.7.[22]

1.4.3 Data sources and data collection

The data sources used were the files on the evaluations in the archives of the commissioning research councils, interviews with the participants in the decision-making processes and in some cases, their private notes and drafts. All except one of the members of the selected evaluation panels have been interviewed (27 out of 28 panel mem-

---

[21] Two cases involve more than one of the categories.

[22] Several factors co-vary. The evaluations of fields assessed basic research in one discipline, were organised by the same research council, and were carried out by international peer panels. The evaluations of institutes were undertaken by Nordic mixed panels and assessed applied research. The research programmes were multi-disciplinary, consisting of both basic and applied research, and were evaluated by (Nordic or international) peer panels.

bers).[23] Five panel members were Norwegians, ten were from other

Nordic countries and twelve were non-Nordic. Four of the evaluators were 'non-peers' whose function was to assess extra-scientific relevance, applicability and use of the research (the two panels evaluating institutes).

In-depth semi-structured interviews were used, that is, open-ended questions posed in a certain order, and with possibilities for follow up questions. The order of the questions was frequently changed to facilitate the dialogue. Eighteen of the interviews were person-to-person, nine were conducted over the telephone. At least half of each panel was interviewed person-to-person. Phone interviews are, of course, not optimal neither in terms of obtaining sensitive information, nor in terms of preventing misunderstandings, but as the evaluators were from all over the world it was impossible to reach them all within a reasonable travel budget.[24]

The interviews commenced by asking the evaluators why they thought they were selected to undertake the evaluation, and their motives for accepting the job. These questions provided much background and network information, and proved to be a good way to start a dialogue and develop a rapport with the informants.

The core of the interviews dealt with the sources of information provided and used by the panel, the criteria for assessment, the way discussions were conducted in the group, and disagreements between the panel members. This included questions about the panel members' prior information about the research and researchers they were going to evaluate, the criteria for different kinds of reviews, and for the evaluation in question, and what criteria they thought the other panel members had adopted.

The interviews usually ended with more general questions about the evaluators' opinions on the purposes, usefulness and weaknesses of the kind of evaluation they had participated in. Most interviews lasted between 1½ and 2 hours.[25] The interviews were recorded and transcribed.[26] Interview data were seen in relation to the other data sources – the research councils' files on the evaluations, oral information from the secretaries of the panels,[27] and in some cases the evaluators' private notes and their drafts for the evaluation report. When informants' accounts on the same question diverge, the various versions are presented in the case descriptions. When reaching conclusions, conflicting statements are analysed in relation to each other and the context in which they were stated (informants memory

of events differed, as well as their will to speak openly about sensitive matters).[28]

The cases studied are presented anonymously. To secure informants' confidentiality was seen as necessary in order to acquire the necessary information on the decision-making of the panels, especially on disagreements among the panel members and other kinds of information systematically excluded from the evaluation reports. As the identity of the panel members of the evaluations is public knowledge, I cannot disclose which evaluations are studied without revealing the identity of the panel members and my informants. Consequently both informants and cases are anonymous. A reason for full confidentiality, in addition to the access to data that informants was expected to otherwise be very reluctant to provide, was not to affect those involved in the evaluations. If, for instance, the non-official conditions for the evaluation reports had been made public, it might have affected both the credibility of the reports, and the reputation of the evaluees.

Informants were told that they would not be cited by name, and that they would be allowed to read text in which they were (anonymously) cited before publication. The major drawbacks with confidential data are that these cannot be checked by the reader, and the reader cannot draw on other information he or she has about the case (Yin 1989:142).[29] The only 'external' check on validity of the presentation of the cases was made by the actors themselves (panel members, co-ordinators and secretaries). They were all asked to comment on my draft on the description of the evaluation. The drafts presented the various accounts of the panel members and (when necessary) tried to 'reconcile' them into a coherent story.[30] All secretaries/co-ordinators, and seventeen of the panel members provided comments. Most of them just gave a short message saying 'no objections' or 'OK'. A few reacted to particular formulations in the accounts which did not affect the description as such, in which case I reformulated the phrases in question. When informants wanted to change the formulation of their statements in a way affecting the meaning, this is included in footnotes (Chapter 4). I also received inspiring feedback from informants, especially concerning those cases with more 'intricate' decision-making, stating that they had gained some insight by reading the draft (Case 5 and 6).[31]

There are obvious problems with collecting relevant data from confidential decision-making processes like expert panel evaluations. Written documentation is limited, and the participants have incentives to give a picture of their decisions as more neutral and thorough than they actually were. Apart from problems with getting appointments with some of the panel members, the data collection has been easier than expected. Some of the interviewees have been quite outspoken and have provided me with information which has been useful when interviewing less outspoken persons to get their views on controversies and similar in the group. Nevertheless, it is all second-hand information. The data material might have been better with direct observation of the decision-making. However, taking the ad hoc character of the evaluations into consideration, direct observation might easily have affected the work of the panels. A more pressing problem would be getting permission and access to observe on-going evaluations (see Section 7.2 for further discussion).

2 The problems with identifying good research

A central focus of the literature and debate on peer review has been the problem of biased assessments. It is, for instance, more or less directly, claimed that peer review is essentially an 'old boy system', that it is full of scientists feathering their own nests, that it stifles innovative research because assessments are done by well-established researchers rejecting ideas differing from their own, that it discriminates against scientists working in low-prestige institutions, or fails to screen out grant applications of questionable merit (Cole et al. 1978:11ff; Turney 1990:39; Chubin & Hackett 1990; Wood 1997).

The *concept* of bias is seldom discussed, but interpreted in various ways. Some studies finding disagreements among peer reviewers interpret this as some sort of 'cognitive particularism' (Travis & Collins 1991) or 'confirmatory bias'[32] (Mahoney 1977), while others interpret disagreements as 'real and legitimate differences of opinion among experts about what good science is or should be' (Cole et al. 1981:885). Such divergent interpretations reveal a lack of common understanding not only of the notion of bias, but also about what are legitimate considerations when evaluating research. This chapter deals with perspectives on constitutive and contingent properties of the *basis* of research evaluation as a starting point for discussing the notion and researchability of 'bias' in research evaluation. The first section discusses the existence of, and need for, criteria and indicators of good research. In the second section formal and informal rules for evaluation processes are discussed and the ontology of 'research quality' is seen from the point of view of realism, idealism and pragmatism/nominalism. The third section elaborates on the properties of research evaluation by discussing the norms of scholarly communities and scientific activity. The last section discusses the notion of bias and develops a classification of bias in research evaluation.

2.1 What constitutes 'good research'?

To deal with the notion of bias in research evaluation we need to know at least something about what an unbiased evaluation is supposed to be based on. Is there something that is unrefutable as the basis of research evaluation – something we all must accept as the nature or essence of 'good research'? Are there identifiable impersonal and neutral criteria for good research, or factors that should *not* influence research evaluations?

While scientific research is traditionally looked upon as a truth-seeking activity, we may say – considering the extensive focus on bias – that fair judgements or neutrality seem to be an overall requirement for the evaluation of research. Separating the concepts of truth and neutrality, it may be expressed as follows: When *doing research* truth is an objective, while neutrality might be the way leading to truth.[33] When *evaluating research*, neutrality is an objective, while truth, methodological stringency, refutability, and similar, may be neutral or scientific criteria for assessing research.[34]

Not everybody would agree to such a focus on neutrality and impartiality in research evaluation. Judging the technological or societal relevance or effects of research involves taking a political position: what technology and what kind of society do we want? Discussing criteria for selecting research projects or effort areas, Weinberg (1963) puts forward both internal and external criteria. These can be summarised as follows:

> Internal criteria for scientific choice:
> - Is the field ready for exploitation?
> - Are the scientists in the field really competent?
> External criteria for scientific choice:
> - Technological merit: Is the technical end worthwhile?
> - Scientific merit: Relevance to neighbouring fields? [35]
> - Social merit: Are the social goals worthwhile?

This list illustrates the wide spectrum of considerations relevant to *ex ante* research evaluation: competence, researchability, interdisciplinary relevance, societal and technical ends.[36] Are the assessments of these aspects expected to be neutral/impartial in some way? Technological and social ends are as mentioned normally not considered neutral or impartial.[37] Yet, procedures for impartial assessments may be construc-

ted (at least theoretically, see John Rawls' *A theory of Justice,* 1971). As for the internal criteria, impartiality is expected to be a far more central characteristic of the assessments. The concern with bias deals specifically with intra-scientific evaluation (i.e. peer review). Personal interests or other biases are not welcome as the basis for the assessments of the competence of a researcher, of whether a field is researchable or not, or of the quality of a research report.

What are such assessments supposed to be based on then? According to empirical studies there is a certain set of aspects and attributes of research quality that researchers have in common and use for scientific assessments. The most emphasised aspects are the research problems, the methods and the results. Stringency, correctness, novelty, depth, breadth, intra- and extra-scientific relevance and productivity are examples of attributes of good research – some of the attributes are presumably more relevant for one aspect than another (Hemlin 1991).

Combining Weinberg's list with these findings, including some later issues in research policy debates (environmental merit, ethical acceptability) and also considering possible indicators of good research, we may end up with the following sketch of what constitutes good research and how it may be assessed:

Table 2.1 *Possible aspects, attributes and indicators of good research*

Aspects	**Attributes** (examples)[38]	**Indicators** (examples)
Intra-scientific quality		Ex ante indicators Expert judgements of: • Reseachability/fruitfulness. • 'Talent'. • Previous formal reviews (book reviews, review of candidates for chairs, etc). • Reputation (institutional affiliation, posts, international network). Ex post indicators: • Peer judgements of the various attributes. • Frequency/amount and forum of research reports/ publications/-communications. • Citations.
Research questions	Researchability, fruitfulness, stringency, originality.	
Methods	Fruitfulness, stringency, correctness.	
Theory	Explanatory power, consistency, simplicity, generality, cumulativity, originality.	
Reasoning	Consistency, stringency, profundity, completeness, originality.	
Results/effects	Correctness, theoretical and empirical contributions/ novelty.	
Extra-scientific relevance/effects	Societal merit Technological merit Environmental merit Ethical acceptability	Actual or potential applications, patents, feared or experienced societal/environmental consequences, ethic board assessments.
Productivity	Cost efficiency Organisational efficiency	Intra- or extra-scientific results/effects (output) according to resources (input). Relevant input factors: funding, equipment, organisation (academic freedom, group size, etc.).

Underlying the kind of model presented in Table 2.1, and most philosophy of science, is what may be termed a 'constitutive' perspective on good research. There are certain characteristics that *constitute* good research as such, and restrict what may be meant by 'good research'. Table 2.1, for instance, assumes that a question needs specific attributes (e.g. researchability, fruitfulness, stringency and originality) to be a good research question, implying that a question lacking all these characteristics cannot be a good research question. When it comes to *indicators* of good research, the claims of a constitutive perspective may be more vague. The kind of 'evidence' of good research listed in the last column is not constitutive in the same

sense as the attributes. The talent and reputation of a researcher is not, for instance, said to constitute the quality of his/her research questions. Nor do patents constitute technological merit, though patents may be socially constituted as a 'valid' measure of what should count as technological merit.[39]

Strictly speaking a constitutive perspective, defining the nature of research quality, is required for a meaningful concept of biased evaluation. To separate a biased from a non-biased evaluation, we need a concept of good research that specific suspicions about biased judgements may be measured against. *If there is nothing that constitutes good research as such, everything is equally valid and it does not make much sense to speak of bias.* A constitutive perspective allows a meaningful concept of 'neutral' (i.e. non-biased) judgements on research quality. Yet, if there is no general agreement on specific indicators of the constitutive attributes of good research, a constitutive perspective will not directly challenge the conclusions of any specific evaluation – as long as these conclusions do not refer to criteria in a way that violates what is said to be constitutive of good research (e.g. judges a theory as good because it has low explanatory power, or a research question as bad because it is researchable). In other words, a constitutive perspective is a necessary but not a sufficient condition for distinguishing biased from non-biased research assessments.

2.2 The two faces of research evaluation: constitutive versus contingent aspects

Prevailing theory within science and technology studies implies that by studying constitutive aspects of science from an 'armchair philosophy' point of view, we get only part of the picture – only one of the faces of science. We get the ideal, ignoring its relation to 'real life science'. Likewise, if we focus on the ready made, black-boxed side of scientific research, we easily ignore 'science in the making' including controversies, the art of arriving at a convincing result, and all kinds of social, political, economic and psychological aspects that philosophers of science traditionally have paid little attention to (see Latour 1987). Science is said to have one 'official' face, suitable for the public, and one 'informal' face reserved for insiders.

One way of studying these two faces is to analyse scientific discourse, focusing on the divergent repertoires that scholars use in different situations. Gilbert and Mulkay (1984) identify a *contingent* repertoire as opposed to an *empiricist* repertoire used in the formal research literature.[40] The contingent repertoire is used in informal discourse and gives accounts diverging from the accounts of the empiricist repertoire. In a chapter studying scientific humour Gilbert and Mulkay exemplify this with what they call the scientific proto-joke, of which they obtained various versions from several biochemistry research groups. The joke consists of one list of phrases used in the formal research literature, and another list of their informal equivalents. An example from the formal lists is 'Accidentally strained during mounting', the informal counterpart being 'Dropped on the floor'. Some of the other 'couples' are:

> *'Handled with extreme care throughout the experiment' – 'Not dropped on the floor'*
> *'It has long been known that . . .' – 'I haven't bothered to look up the reference'*
> *'Fascinating work . . .' – 'Work by a member of our group'*
> *'Of doubtful significance' – 'Work by someone else'*
> *(Gilbert & Mulkay 1984:176)*

Categorising this joke, it may be said to be the kind of humour that makes us laugh because it pushes to extremes something partly taboo that we all easily recognise. In this case, the joke shows us the way we disguise our formal communication with the proper rhetoric in order to gain scientific credibility. It 'comes close to being a satire directed at the official discourse of science' (Gilbert & Mulkay 1984:178).

The existence of such double repertoires, or two divergent faces, is not unique to science. In fact, the example of two divergent faces most relevant to research evaluation panels that I have found, is from the legal realm. In Garfinkel's presentation of official and informal rules for jurors we find a version of dual rules easily 'translated' to the context of research evaluation. Substituting 'evaluator' for 'juror',[41] the *official* rules appear as follows:

Table 2.2 *Official rules for fair and thorough evaluation processes*

Impartiality requires:
I1. For the impartial evaluator, personal preferences, research interests and paradigm, i.e., his/her perspectival view, are suspended in favour of a position that is interchangeable with all positions found in the concerned research community. His/her point of view is interchangeable with that of 'Any Researcher' in the area.
I2. Assessments vary independently of sympathy.
Thoroughness requires:
T1. For a good evaluator, 'criteria' and 'evidence' are the only legitimate grounds for an assessment.
T2. The good evaluator delays judgement until all material subjected to evaluation has been investigated.
T3. For a good evaluator the expression of a position that involves an irrevocable commitment is withheld. A good evaluator will not take a position at a time that will require him to defend it 'out of pride' instead of 'on the merit of sound criteria and available evidence'.

(Moderated from Garfinkel 1967:109–110)

In contrast, focusing on the non-official face of research evaluation the following rules appear:

Table 2.3 *Informal rules for good evaluations processes*

An evaluation process is good:
P1. If it keeps to its time limits.
P2. If it does not require the evaluator – as a condition for making judgements – to act as if he knows nothing, i.e. it does not require the evaluator to make no use of What Any Competent Member of the Research Community Knows that Anyone Knows.
P3. If the number of variables defining the problem (and thereby the adequacy of a solution) can be reduced to a minimum by trusting that the other persons on the panel subscribe to the same kind of common sense.
P4. If the opportunity and necessity for looking behind the appearance of things is held to a minimum.[42]
P5. If only as much of the situation is called into question as is required for a socially supportable solution to the immediate problem in hand.
P6. If the evaluators emerge from the inquiry with their reputations intact.

(Moderated from Garfinkel 1967:108)

The two sets of rules bring out the incompatible requirements confronting an evaluator. On the one side there are formal (and maybe unrealistic) requirements, on the other side there are informal

pragmatic requirements. The formal rules focus on aspects that may be said to be constitutive of research evaluation: impartiality and thoroughness. If research evaluation is not supposed to be impartial and thorough, it becomes somewhat meaningless. Even as means in a political power struggle the conclusions of an evaluation lose a crucial function (credibility/authority) if it is clear to everybody (and everybody knows this) that the evaluation cannot be said to be impartial and thorough – or at least as impartial and thorough as expected/possible for the specific purpose. The impartiality requirements are common to various contexts of judgements, both legal and professional. The rules of thoroughness have their parallel in standards of scientific research and are part of what is meant by 'scientific'.

The informal rules focus on pragmatic aspects, on various simplifying or cost reducing strategies, that moderate the requirements of impartiality and thoroughness. The informal rules say that time limits, common 'knowledge', a socially supportable conclusion and the reputation of the evaluators are more important than impartiality and thoroughness. They may, however, be taken to be pragmatic either the way that they tell you to do your best within the given confines, or to choose the easiest way out. In the first case the informal rules tell you to moderate the formal rules as little as possible. In the latter case they tell you that you may disregard the formal rules.

Anyway, the double set of rules pictures an ambiguous situation – there are two divergent sets of rules for 'good' evaluations. In Garfinkel's study the informal rules are presented as the decision rules of everyday life and perceived by the interviewed jurors as seemingly unacceptable in a legal context. The jurors gave idealised accounts of their decision processes in line with the formal rules and became anxious 'when during the interviews, their attention was drawn by interviewers to the discrepancies between their ideal accounts and their 'actual practices" (Garfinkel 1967:113).

That informal rules are not officially 'promoted' or admitted does not deny their existence. Yet, to count as rules they must in some way be *accepted* (at least informally) as pragmatic strategies – and not seen as mere imperfections. In Table 2.3, P1 and P4 *prescribe* cost reducing strategies, P2 and P3 *prescribe* the use of common 'knowledge' and P5 and P6 *prescribe* attention to what is socially acceptable.

There may be a more 'nuanced set of rules, reconciling the formal and informal rules, by specifying in what way and to what degree the formal rules may be modified, or in what situations the various sets of

rules apply. When such rules do not exist, each decision-maker is 'free' to find his/her own way out of the ambiguities. In other words, discretion is essential, as for most professional decision-making.

2.2.1 Divergent rules in light of ontological perspectives: realism, idealism and pragmatism/nominalism

Statements on how research quality is constituted or on the status of 'attributes' of good research imply an ontology of research quality. As we shall see, the understanding of bias in research evaluation heavily depends on the ontology of research quality we rely on. Different ontological views on research quality also have different implications for the kind of rules suitable for guiding evaluation processes. The formal and informal rules (Tables 2.2 and 2.3) do not only differ in their requirements for good evaluation processes, they also have divergent meanings for the basis of research evaluation, i.e. divergent implications on the content of 'research quality'. The constitutive aspects of research evaluation relate to how we understand the *ontological* status of 'research quality'.

In discussing how 'research quality' may be said to be constituted, we shall use the terms 'idealism' and 'realism' in their old philosophical meaning. *Realism* denotes the view that reality exists independently of being experienced or conceived. Realism of research quality means that there are standards constitutive of good research, unrelated to what evaluators might define as good research. In this view, good research might be something quite different from what the research community defines as good research. *Idealism*, on the other hand, means that reality is constituted by experience and thought, and therefore there is no reality independent of human consciousness (Lübcke 1983:362, 204). An idealistic concept of research evaluation implies that the meaning of 'good research' is constituted through the evaluation process.

Both realism and idealism may be seen either from an optimistic or pessimistic point of view. Being optimistic, realism can be taken to mean that there are independent standards for good research obvious to the evaluators. Pessimistic realism, on the other hand, takes the existence of independent standards of good research for granted, but says that such measures are not obvious, and an evaluator may easily reach false judgements. Optimistic idealism may for instance say that evaluation processes constitute research quality thoroughly and

impartially, and therefore authoritatively. Pessimistic idealism may for instance say that evaluation processes are heavily influenced by non-relevant factors, power strategies and the like, and research quality is therefore 'constituted' in a partially and non-authoritatively manner.[43]

'Research quality' might also refer to something that is neither independent of how evaluators may define it, nor culturally or socially constituted through evaluation processes. It might refer to the conclusions of single evaluations as such, regardless of how these conclusions relate to independent or socially/culturally constituted standards. In such a perspective 'research quality' has no particular content, and any conclusion is equally valid. Standards left for judging an evaluation then, are standards like whether the evaluation process is efficient and its conclusions unambiguous. This may be called a pragmatic perspective on research quality. The perspective is also related to nominalism – the view that general terms have no content and that the particulars they denote have nothing in common except that the same general terms are used about them (Lübcke 1983:316). A nominalistic view on research quality implies that 'good research' is an empty concept meaning that various good research projects have nothing in common except that *we say they are good.*

Three major different understandings of 'research quality' have been outlined above: realism, idealism and pragmatism, summarised in Table 2.4. Realism and idealism have both an optimistic and a pessimistic version.

Table 2.4 *Perspectives on 'research quality'*

Realism:
There are independent standards for good research.
Optimistic realism:
These standards are obvious to competent evaluators.
Pessimistic realism:
These standards are *not* obvious to evaluators.
An evaluation panel may easily reach *false* judgements.

Idealism:
Standards are culturally and socially constituted, i.e. research quality is defined through evaluation processes.
Optimistic idealism:
Evaluation processes define research quality thoroughly and impartially, and therefore authoritatively.
Pessimistic idealism:
As evaluation processes are heavily influenced by non-relevant factors, power strategies and the like, research quality is defined in a partial and often non-authoritative manner.

Pragmatism:
'Research quality' has no generalisable content: various good research projects have nothing in common except that we say they are good (nominalism).
The main object of research evaluation is reaching a conclusion.
To succeed, decision-making rules constructing consent are essential.

'Realistic' evaluations mean assessments according to specific non-socially constituted standards. The legitimacy of evaluations in a realistic perspective, depends on the competence of the evaluators – their insight into the independent standards and their ability to reach the *right* conclusions. Idealism implies that the felicity of conclusions depends on characteristics of a collective decision-making *process*. Realism does, however, not exclude the idea that some kinds of decision-making processes are more likely to lead to the right conclusions than other processes (in Plato's terms this would be a process leading out of the cave).

Realism and idealism are in line with a constitutive perspective. Realism says that there is something that *is* constitutive of 'good research', idealism says that 'good research' is constituted culturally and socially.[44] Pragmatism/nominalism on the other hand, rejects both these views – saying that 'good research' has no content except as conclusions of particular evaluations. An exclusively pragmatic/-nominalistic view on research quality renders research evaluation

meaningless. If all conclusions are equally meaningful or good, there is no need to appoint experts to evaluate research. The outcome might as well be determined by a lottery or another cost-effective method. Such a random method would never be accepted as an *evaluation*. When research quality is 'constituted' by a 'nude' decision it appears random and meaningless. Randomness contradicts the whole concept of evaluation.

We shall not attempt to make any clear-cut conclusions on the status of 'research quality', here. Questions like 'Is research quality unrelated to how it may be perceived or defined by evaluators – research quality *per se?'* need not have a definitive answer. One possibility is that realism, idealism and pragmatism are supplementary ways of understanding research quality, not necessarily contradictory perspectives. Consistency may, for instance, be seen as a realistic aspect of research quality, while novelty and cumulativity may be seen as idealistic aspects, constituted through consensus-making processes. On other aspects, like fruitfulness and extra-scientific merit there may be no broad common understanding. Such aspects may be 'resolved' by pragmatic decisions when evaluation panels are expected to give unanimous judgements on them.

What about the divergent rules for research evaluation? What implications do a 'realistic', an 'idealistic' or a pragmatic view on 'research quality' have on the understanding of official versus informal requirements for evaluation? In a *'realistic'* perspective procedural rules are in themselves irrelevant for whether an evaluation is 'right' or 'wrong'. The requirements of thorough evaluation processes in Table 2.2., for instance, cannot guarantee the right outcome. A lottery may by hazard also provide a correct conclusion. Yet, rules specifying standards for 'correct' assessments (i.e. standards of good research) and how to reach the 'right' conclusions, may be essential as a means to a correct evaluation. In such a perspective the official rules, emphasising criteria, evidence, impartiality and thoroughness definitely seem better suited than the informal rules emphasising cost-reducing strategies, common knowledge and social acceptability. If, for instance, an evaluation is not thorough enough to detect inconsistencies in a research report, it will reach faulty conclusions regarding the consistency of the report, which – given that consistency is a 'real' aspect of research quality – have implications for whether the assessments of the research quality are correct. Informal rules promoting common knowledge and social acceptability, on the other hand,

are not the way to detect inconsistencies and would therefore not serve realism. Furthermore, realism claims that 'research quality' is independent of the kind of social factors emphasised by the informal rules.

From an *'idealistic'* perspective we may claim that the best rules are those that best serve the process defining research quality, i.e. the rules should reproduce authoritative definitions of research quality or redefine research quality authoritatively. The question then is whether the official or the informal rules best serve this process. If we say, under the 'idealistic' perspective, that impartiality and thoroughness are essential (social/cultural) requirements for acceptable evaluations, the answer is that the idealistic perspective brings about the official rules. It should be noted that this answer is culturally conditioned, and not given *a priori*. In principle, any procedural rule – also the informal pragmatic requirements – may help define research quality, depending on what kind of evaluation processes yields authoritative results in a given context. In this regard, the requirement of social acceptance (included in the informal rules) is likely to be able to overrule requirements of thoroughness and impartiality.

From a *pragmatic* perspective 'research quality' has no content, and the most efficient rules are the best. As the informal rules are likely to be easier to follow and yield conclusions faster than the official rules, they are preferable from a pragmatic perspective.

Summing up, the official rules seem to be the most appropriate from a 'realistic' perspective on research quality, while the informal rules are the most appropriate from a pragmatic view on research quality, and idealism in principle may apply both sets of rules. Idealism therefore may be the best suited for combining the two sets of rules.

2.3 Norms of research evaluation?

This section further discusses the status of divergent rules and the scope of acceptable evaluations by looking at the two sets of rules of Tables 2.2 and 2.3 (official rules and informal rules for good evaluation processes) in relation to norms of the scientific community.

2.3.1 'The ethos of science'

The classic text on the norms of the scientific community is Merton's *'The Normative Structure of Science'*, written in 1942 (at least partly as a response to the Nazis' attempts to control science). According to Merton, there is an 'ethos of science' containing four sets of institutional imperatives: universalism, communism, disinterestedness and organised scepticism. Versions of universalism, disinterestedness and organised scepticism may be recognised in thoroughness and impartiality of the 'official' rules (Table 2.2), while communism has no parallel in these rules, nor is it directly applicable to the evaluation of research.

Universalism says that 'the acceptance or rejection of the claims entering the list of science is not to depend on the personal or social attributes of their protagonist; his race, nationality, religion, class and personal qualities are as such irrelevant.' Universalism further demands that careers be open to talents. In terms of evaluation procedures, universalism demands that 'truth-claims, whatever their sources, are to be subjected to preestablished, impersonal criteria' (Merton 1942/1973: 270–272).[45]

Communism, here meaning common ownership, says that 'the substantive findings of science are a product of social collaboration and are assigned to the community.' Scientists may compete and have controversies over priority, but the products of competition are communicated and common property. The producer's reward is recognition and esteem, not private property of the findings. Merton describes patents as incompatible with communism and a threat to the scientific ethos. As a response to this threat 'some scientists have come to patent their work to ensure its being made available for public use', Merton writes (ibid.:273–275).

Disinterestedness is a norm aimed at preventing fraud and misconduct in science. Merton is not specific about the content of this norm. He says that it is not a question of 'distinctive motives' that the scientist should have, 'it is rather a distinctive pattern of institutional control of a wide range of motives which characterizes the behaviour of scientists'. The norm is 'effectively supported by the ultimate accountability of scientists to their compeers' (ibid.:276). The interpretations of Merton's 'disinterestedness' vary from Mulkay saying that it 'requires researchers to pursue scientific knowledge without considering their career or their reputation' (Mulkay 1977:98), to the

rather strict 'without having any reward in view, whether financial, emotional or social' (Barnes & Dolby 1970:4) or more general interpretations like 'curbing of personal bias' (Zuckerman 1988).[46]

Organised scepticism is the mandate of the scientific investigator to wait for evidence before making judgements, and to ask questions about any aspect of nature and society regardless of 'the cleavage between the sacred and the profane, between that which requires uncritical respect and that which can be objectively analyzed' (Merton 1942/1973:277–278).[47]

According to Merton, the norms and values of the scientific ethos 'is held to be binding on the man of science' and 'are expressed in the form of prescriptions, proscriptions, preferences and permissions'. The ethos has not been codified, but 'can be inferred from the moral consensus of scientist as expressed in use and wont, in countless writings on scientific spirit and in moral indignation directed toward contraventions of the ethos' (ibid:269). This does not mean that the norms are not violated. The desire for recognition may lead to fraud and plagiarism, violating all sets of norms (Merton 1957/1973). Nevertheless, the idea seems to be that the ethos is either violated *or* obeyed. There is no room for the kind of 'negotiability' expressed by the 'informal' rules, the ethos being roughly in line with the 'official' rules.

Merton's 'proposal of norms in science provoked [a] prolonged and heated discussion', which was one factor leading to the division of sociologists of science into 'several contending groups' advocating different approaches to the studies of science (Zuckerman 1988:516).[48] My argument here is that the existence of the ethos is *not* a question of whether it is violated or not. Examples of violation of the norms are easy to find and probably most scientists know of some violations. The real test of the social significance of (potential) norms – of whether they are adhered to or not – is the moral reaction to such violations. In our case the question is: are there examples of behaviour contrary to the ethos – particularism, non-communication of results, personal bias or 'dogmatism' – that will not be sanctioned or entail moral indignation among colleagues, if detected?

As I see it, the answer depends on how categorically the ethos of science is interpreted. If we take universalism to be incompatible with all kinds of 'old boys networks' and institutional and personal loyalties, communism to be a norm not only for academic research,

but also for industrial research, organised scepticism to forbid the kind of dogmatism a scientific paradigm entails, and disinterestedness to be incompatible with the kind of personal bias resulting from having personal and institutional loyalties and complying to a paradigm, then we do not even need empirical studies to answer the question. With such an interpretation there are obvious situations in which one would be expected to be (and rewarded for being) 'particularistic', 'dogmatic', 'personally biased' or to not communicate results.[49] On the other hand, if we adopt a 'soft' interpretation, and say that the only violations of the ethos are those including behaviour that are clearly understood as fraud or misconduct, we come close to a tautological argument, saying that all violations of the ethos entail moral indignation because everything defined as fraud or misconduct entails moral indignation.

So, with a categorical interpretation the ethos is obviously not binding – it can be violated without provoking moral indignation or negative sanctions – but with a 'soft' interpretation it is binding per definition. In deciding whether scientists are 'controlled' by professional norms or not, however, we do not have to take a stand on which of these interpretations is the best (and Merton is reasonable enough to have meant something in between). *The fact that some behaviour is defined as fraud or misconduct and sanctioned, is evidence enough to claim that the scientific community has some norms.* This does not mean that Merton's proposal is an adequate description of those norms however, nor that the norms are the same for all scientific periods, environments or situations.[50] As mentioned, several studies find that there are reasons to question 'the ethos of science' as put forward by Merton. Whether 'the ethos of science' provides an adequate description of the actual behaviour of scientists or not, it seems to be a good account of the way scientists like to portray themselves (Jasanoff 1990:63; Mulkay 1977:108). The ethos is the proper rhetoric for maintaining the prestige of science in society, which means that we should be somewhat critical about scientists' own accounts on their adherence to such norms.

Thus, we cannot conclude this section by pointing to a set of clear and specific norms that guide behaviour in scientific communities. Studies of science have not yet fully grasped the balance between fraud and misconduct on the one side, and acceptable loyalties to persons and paradigms, and acceptable secrecy of results on the other side. The reason for this might be that there are no clearly set borders, but *continuous negotiations* about the definition of specific cases as to which

side of the border they belong – a set of unclear, unspoken, context-dependent and changing informal 'rules' that moderate or assist in the interpretation of the 'official ethos'.[51] The recent establishment of ethics committees, both at the institutional and national level, would be one indicator that the norms of the scientific community are felt to be too ambiguous to be interpreted solely by individuals on their own.

Some analysts have dealt explicitly with the ambiguities of the norms of scientific behaviour, however. Merton himself stresses that 'the social institution of science ... incorporates potentially incompatible values' (Merton 1963/1973:383). On one hand, for example, there is a 'value set upon originality[52] which leads scientists to want their priority recognised' (loc.cit.). On the other hand, there is the norm of 'selfless dedication to the advancement of knowledge for its own sake' (op.cit:399), and a 'value set upon due humility, which leads [scientists] to insist on how little they have in fact been able to accomplish' (op.cit.:383). Merton uses these ambivalent norms to explain the *contradicting statements and behaviour* of scientists. Scientists involve themselves passionately in priority debates, while in other contexts they tend to trivialise or even reject such priority debates. The idea of ambivalence caused by norms and counter-norms in science is further developed by Mitroff in a study of the Apollo moon scientists. He finds counter-norms to all the norms of 'the scientific ethos', and concludes that the dominance of norms is *context dependent*. In some situations the norms of the 'ethos' will dominate; in other situations the 'counter-norms' will dominate. Mitroff suggests that the nature of the scientific problem addressed will be among the factors influencing what set of norms will dominate a specific situation, but concludes that much more study is required before we can understand the ambivalence of scientists (Mitroff 1974).

2.3.2 Basis and norms of quality evaluations

Given the sparse and inconclusive empirical studies of general norms in science, trying to infer specific norms for evaluation behaviour might be a bit too optimistic. However, the 'ethos of science' might serve as a fruitful starting point. Translated into imperatives for evaluating research quality, Merton's scientific ethos could be summarised like this:

> Your evaluation shall not depend on personal or social characteristics of the researcher, nor shall your personal motives influence your evaluation. Your evaluation shall be based on preestablished impersonal criteria, and be rigorous enough to uncover any fraud or misconduct. You shall reward critical, undogmatic inquiry and free communication of results.

There are two critical problems with this ethos for evaluation. Firstly, as mentioned above, there is no evidence that 'the ethos of science' is binding on the scientist, and certainly no evidence saying that my 'translation' of the ethos to the context of evaluation is binding. Several studies of peer review indicate that assessments are neither rigorous, nor impartial (Ceci & Peters 1982; Peters & Ceci 1982; Chubin & Hackett 1990; Cicchetti 1991; Cole et al. 1981; Mahoney 1977; Travis & Collins 1991; Wood 1997). The focus on bias, however, and the concern to bring the problems into the light, indicate that it is common view that research evaluation ought to be rigorous and based on impersonal criteria. In terms of idealism, evaluations need a minimum of rigour and impartiality to be authoritative. Furthermore, if one were to argue for the opposite, that evaluations should be lax or partial, one would contradict the whole meaning of evaluation. In terms of realism, there is some intrinsic idea in the notion of evaluation saying that a lax or partial evaluation cannot be a good evaluation. Consequently, we can conclude that such an 'ethos of evaluation' is likely to be in accordance at least with scientists' official statements about evaluation (the official rules), but not necessarily a good description of what norms peer evaluators adhere to (the informal rules), i.e. one may violate the 'ethos' without provoking any moral indignation if all those competent to detect the violation conceive the 'ethos' as mere rhetoric, and that no one (or very few) try, or have the capacity, to live up to it. In addition, there might be a large 'grey area' between a rigorous and impersonal evaluation, and a lax and partial one, and as long as one does not move outside this area – over in the area of clearly lax and partial evaluation – the evaluator might not need to fear any negative sanctions.

Secondly, even if there are norms saying that evaluations shall be impersonal and rigorous, and reward critical inquiry and free communication of results, these norms give no guidelines for determining the quality of research. How is research of good quality distinguished from research of inferior quality by journal referees, committees for academic appointments and grant application reviewers? What are the

standards or criteria adopted for evaluation? As we saw in the previous sections, these questions have no apparent answers.

Ravetz emphasises that the assessment of research quality 'involves the making of a number of subtle, indeed tacit judgements, which depend on an intimate craft knowledge of the work under review' (Ravetz 1971:274, see also Appendix A for an explication of 'tacit'). The techniques are too subtle, the criteria too specialised, and the materials too rapidly changing, for any formal categories of quality to be feasible. It has also been underlined that the attempts by philosophers of science to formulate criteria of good research (i.e. corroboration, explanatory power, predictive power, simplicity, and similar) are 'not intended as practically applicable measures for appraisal within science policy' but are aimed at understanding the nature of science, and are part of 'the dynamic and critical self-reflection of the scientific community on the criteria of good science' (Niiniluoto 1987:22). This reliance on tacit knowledge, craft skills, and the lack of explicit criteria,[53] underscores that there will normally be *a large grey area* of acceptable evaluations, i.e. evaluations not clearly definable as lax or partial.

As mentioned, numerous studies of peer review have focused on the reliability and possible bias of peer review and found a low degree of agreement between referees and various kinds of bias (academic and institutional status, nationality, gender and research field of the author/applicant influence judgements, as well as different kinds of cognitive bias).[54] Contrary to such studies, the studies of Hemlin referred to in Section 2.1, focus on criteria. He finds a common 'language' in evaluation of scientific quality, a certain set of criteria (aspects and attributes) to which researchers pay attention. He also finds significant differences between 'hard' and 'soft' sciences in the emphasises on the various aspects. The humanities and social sciences, for instance, put more emphasis on theory, while the natural sciences put more emphasis on the results. There is also generally more variation in the emphasis on the various criteria in the humanities and the social sciences, than in the natural sciences (Hemlin 1991).

As Hemlin did not address the question of inter-reviewer agreement on the assessments of the candidates, his studies do not contradict the studies that find low reliability in peer review. If both the findings of low reliability and Hemlin's findings of a common set

of criteria for evaluation of research quality are correct, this can be taken to mean that while there is a certain set of criteria to which attention is paid – more or less explicitly – by peer evaluators, *these criteria are interpreted or operationalisised differently by various evaluators.* This is in accordance with Ravetz's account of subtle, tacit judgements and lack of formal categories of research quality.

Such a common set of criteria (or conceptual basis) with a wide range of interpretative possibilities, can also be read into the findings of a Norwegian study of peer review looking both at criteria and bias. Fürst (1988) studied the basis for evaluations of candidates for professorships in Norway. According to the rules a certain amount of 'breadth' is required for these positions, and Fürst also found that this criterion was emphasised in the evaluation documents. There were no clear norms for the assessments, however, and there was a large variation in what criteria were emphasised – and *how* they were emphasised. It may seem therefore, that what determined each case was accidental. Nevertheless, Fürst sees some patterns in the evaluation documents. As criteria have no standard operationalisation or interpretation there are ample possibilities to choose interpretations that promote the personal favourites of the evaluators. The study concludes that these interpretations tend to be biased against the female candidates. The choice of words for describing the breadth and depth of a candidate's research production, is one of the major examples. The research of female candidate is typically described as either *narrow* and one-sided, or *spread* over many areas, while the research of her male competitors either goes in *depth*, has thematic *coherence*, or good *breadth*. As a result of the nature of the research topic, Fürst was not able to prove whether such differences in the descriptions of candidates were a result of reviewer bias or not simply caused by differences in the research production between the male and female candidates. This illustrates an important characteristic of peer review – and the main problem for students of bias in peer review – quality criteria are not standardised and their interpretation is the privilege of the reviewer (see Ravetz op. cit.). Outsiders, then, seldom have the possibility to 'prove' bias (see the 'researchability' discussion below).[55]

The lack of pre-set standards and formal hierarchy of control, and the large scope of acceptable outcomes of peer review, in no way necessitates the conclusion that peer review is accidental and unreliable. According to Cole a certain degree of consensus is assured, even at the research frontier, both in the natural and the

social sciences. Consensus is created and maintained by social processes. One main process is the training provided in graduate school, another is the dependence on evaluations made by others:

> *'In making evaluations, scientists depend heavily on standards internalized in graduate school. Most scientists have been educated at a relatively small number of prestigious graduate departments where they studied with eminent scientists. To some extent the views of these teachers influence the standards adopted by their students and the subsequent evaluations made by the students' (Cole 1983:136).*

This is the way tacit knowledge and craft skill is learned. When executed, evaluations depend on the informal hierarchy of the research field:

> *'In the process of evaluation, some opinions count more than others. Generally, the stars of a particular discipline occupy the main gatekeeping roles. By their acts as gatekeepers and evaluators, they determine what work is considered good and what work unimportant' (Cole 1983:138).*

As most people in the research community are willing to accept the judgements made by others, the system works:

> *'We give people more credit for publications in prestigious journals. We think more highly of people who have received grants, fellowships, awards, memberships in prestigious organizations – all based on the evaluation of others' (Cole 1983:137).*

Dependence on reputation/eminence and internalised standards facilitates consensus, but studies of peer review still show low inter-reviewer agreement. The diversity and the individualism in science leading to diverse views seem unavoidable. In addition to the fact that different groups promote different standards of quality and relevance, each member of the scientific community tends to have his/her own particular ideas about quality and relevance. The problem is amplified by the highly specialised character of scientific work. In some cases there may be no 'peers' to conduct peer review:

> *'For the "best" scientists peer review is unlikely. Scientists are at the mercy of peer review systems that may offer neither "peers" nor "review." Instead, applicants must compete with others' intellectual capital, positional advantage, and political clout. Luck of the reviewer draw or mere chance may matter nearly as much as measurable features of the manuscript or proposal. Under current conditions of high competition for research funds and space in first-rate journals, such nonmeritocratic criteria make a decisive difference at the margin' (Chubin & Hackett 1990:194).*

Yet, such a tendency of fragmentation and 'mere chance' should not be overstated. Research areas overlap, depend on each other and compete with each other. Those at the top of the informal hierarchies of science not only influence the standards of their own area; as they are obliged to satisfy each other's expectations to ensure funds and facilities for their area of speciality, there is also a tendency towards some degree of unity. This points towards a 'system of quality control ... throughout the whole of science' (Mulkay 1977:107).

Summing up the discussion on the basis and norms of peer review, we may conclude that peer review depends on tacit knowledge and craft skills internalised through socialisation processes – rendering some unity in the basis of peer review. Informal hierarchies, 'gatekeepers', dependence on judgements made by others, and the overlap and dependencies between research areas also contribute to unity. Moreover, there seems to be a common 'language' for peer review – a certain set of criteria that reviewers (more or less explicitly) pay attention to.

On the other hand, studies of peer review find low inter-reviewer agreement, indicating that evaluators either use different criteria, emphasise the various criteria differently or interpret the criteria differently – all leading to divergent assessments. Various characteristics of scientific research and scientific communities may account for such findings – a major one being the tacitness of the basis for assessments. Tacit basis means a large scope of possible assessments. The tacit basis of evaluation also means that the status of the rules (in addition to the rules themselves) may be tacit. Other factors pointing against unitarian standards or norms for evaluation are the inherent uncertainty and controversy in scientific research and the individualism of the scientific culture.

In consequence, we cannot specify rules, standards or exhaustive criteria of peer review. There are still limits as to what may be seen as acceptable evaluations. Criteria are set more or less explicitly by the 'stars' of the discipline, internalised through graduate school, *et cetera*. We may add that to keep to such limits an evaluation at least has to appear to be rigorous and based on impersonal criteria of scientific merit.

2.4 The concept of 'bias' in research evaluation

In Section 2.1 it was stated that a constitutive perspective is necessary for distinguishing biased from non-biased evaluations. If there is nothing that is constitutive of research quality, we have no basis for saying that an evaluation is influenced by prejudices or partiality. In Section 2.2 we saw that philosophical studies of the constitutive aspects of research evaluation are likely to give a one-sided account, leaving out possible informal rules allowing conduct that might be called biased according to 'official rules' found in a non-empirical/-philosophical approach. Section 2.2 furthermore emphasised that 'constitutive' may have various meanings, depending on whether we rely on realism or idealism. Section 2.3 emphasised the tacitness of research evaluation opening a large scope of possible outcomes. In this section we shall identify different kinds and levels of bias in research evaluation, trying to understand the notion of bias in light of the preceding sections.

Bias may have its source on different *levels*. Research evaluation conducted by expert panels may be biased due to factors on either the organisational level, the panel level, or the level of the individual evaluator. We shall examine these possibilities from the two discussed perspectives on research quality that may identify bias: realism and idealism (see Table 2.4).

On the *organisational level* there is the organiser (e.g. the research council) setting the conditions for the endeavour. From a 'realistic' point of view the appropriateness of the organiser's decisions depends on whether these decisions promote 'correct' evaluations or not. Rules, criteria and evaluation forms developed by the organiser may facilitate the task of finding the 'right' conclusion or they may cause 'wrong', i.e. biased, conclusions. Anyway, as far as such bias is standardised, the bias caused by the organiser is not random and might promote fairness in the sense that it provides a common basis for the judgements which may prevent more random bias. According to realism however, such 'fairness' considerations are not relevant. Biased standardisations yield a predictable but arbitrary outcome. There is no help in predictable conclusions if they are not the right conclusions. On the contrary, standardised biased rules and criteria are likely to

persistently impede right conclusions. Under idealism on the other hand, organisationally standardised judgements are more likely to be accepted. If standardisation reproduces authoritative assessments or authoritatively contributes to a common basis for assessments, it may be said to be appropriate according to idealism. Organisational bias, according to idealism, are decisions not in line with common opinions: e.g. appointing partial panels, 'ordering' superficial reviews (indirectly by demanding too much material to be reviewed within a limited period of time), or prescribing the use of methods that will not be accepted as the basis for valid assessments (e.g. publication counts).

On the panel level, *group processes* may either promote or impede the 'right' conclusions (concern of realism) or authoritative conclusions (concern of idealism). We may for instance imagine that processes leading to conformity (see 'groupthink', Chapter 3) may hamper the thoroughness necessary for reaching the 'right' conclusions, or that processes encouraging polarisation may frustrate the authority of the conclusions.

On an *individual level* the competence of an evaluator is decisive both for realism and idealism. The meaning of competence differs, though. With realism, evaluators are incompetent if they do not make the 'right' judgements. With idealism evaluators are incompetent if conclusions are not in line with the process constituting 'good research'. In the first case, the cognitive limits of an evaluator are supposed to be the source of bias. In the second case, some sort of social limits are also involved. Other kinds of bias involved at this level, may be various versions of partiality, for example, that assessments vary dependently of sympathy (forbidden by the 'official rules' of Table 2.2).

The constraints of all levels are discussed in separate sections of Chapter 3 which deal with research evaluation as decision-making and asks what may influence the work of research evaluation panels. The rest of this chapter deals with various *kinds* of bias in research evaluation as such, developing a preliminary classification of bias in research evaluation.

One distinction highly relevant to idealism is that between 'structural' and 'non-structural' bias. 'Paradigmatic' bias, for instance, is structural, while personal likes and dislikes need not be related to any structures. Such personal bias may however be shaped by social and cultural environment, and in that way be structural. Definitive dividing

lines between structural and non-structural bias may therefore be hard to draw. It may be easier to make a division between personal bias on the one hand, and professional or scientific bias on the other. However, personal and professional factors may of course be heavily interrelated, blurring such distinctions. Another distinction to be made is between bias due to interests and bias caused by cognitive constraints. This distinction may also be hard to deal with empirically. Both a large part of our cognitive constraints and our various interests are in some way social. They may in some way be socially related, shaped or constructed, and therefore more or less directly related. Interests may also be directly shaped by cognition, or the other way around: cognition may be shaped by interests.[56] Such distinctions may be useful as analytical tools, but easily blurred empirically.

Table 2.5 *Categories of bias in research evaluation*

	Cognitive constraints	Interests
Scholarly/ professional bias	*A: The constraints of a professional platform:* Preconceptions of good and valuable research. *Selective perceptions* = looking through 'the glasses' of your *'school'*/scholarly viewpoint/profession.[57]	*B: Research interests:* Taking effects on economic and political standing of the field/research area into consideration.[58] *Nepotism* = helping 'heirs' or other colleagues because of *'school'*/scholarly viewpoint or research topic.
Non-professional/ personal bias	*C: General or personal cognitive constraints:* Sub-optimal thoroughness and information seeking. *Selective perceptions* = disregarding information due to routines/limited capacity for handling information.[59]	*D: Personal interests:* Taking effects on personal situation or situation of friends, partners or competitors into consideration. *Nepotism* = helping colleagues because of friendship.

Table 2.5 distinguishes between professional and non-professional/-personal bias on the one hand, and between cognitive bias and interest bias on the other, arriving at four main categories of bias in research evaluation. The two upper categories – the constraints of a professional platform and research interests – are likely to be more structural and predictable than the two lower categories. The bias of category A is grounded in a field's traditions for evaluating research and is therefore not likely to be defined as bias from the point of view of idealism. From such a point of view, the constraints of a professional platform are likely to be looked upon as the basis of (authoritative) evaluations rather than as a source of bias. The bias of category B is more difficult to define in such terms. In some contexts research interests might be part of an authoritative evaluation; in other contexts they might have no authority. From the point of view of realism, both category A and B will be bias. Both professional preconceptions and interests are likely to impede the use of the right standards which realism takes to be culturally and socially independent.

In contrast to the two upper categories, the two lower categories are not very likely to count as authoritative bases for assessment, and more likely to be regarded as bias not only from the point of view of

realism, but also from the point of view of idealism. An evaluator or evaluation panel disregarding, or not understanding, vital parts of the material under review, is obviously incompetent both from the point of view of realism and idealism (category C bias). Giving credit for friendship or discredit for rivalry is also likely to result in neither a correct nor an authoritative evaluation (category D bias). Evaluees subjected to the 'non-professional' bias of category C or D have nothing else to rely on than the 'luck of the reviewer draw'.[60] In the case of professional bias, on the other hand, one can increase the probability of good assessments by sticking to mainstream or widely reputed approaches and topics, and trying to fulfil specific quality criteria.

The various claims of bias in peer review referred to in the introduction to this chapter, may illustrate all categories of Table 2.5. That reviewers are feathering their own nests may either mean that they 'invariably argue … for better treatment of the[ir] field: for more money, more people, more training' (Weinberg 1963:161), i.e. category B bias, or that they tend to credit friends and discredit enemies when reviewing research (category D bias). The claim that peer review stifles innovative research because assessments are done by a conservative 'establishment' is a claim of category A bias. That peer review fails to screen out grant applications of questionable merit is a claim of category C bias.[61]

We also see that the referred[62] contradicting interpretations of disagreements among peer reviewers can be said to be a question about whether 'category A bias' is bias or not. Cole et al. (1981) interpret disagreements as legitimate differences of opinion about the definition of good research. Travis & Collins (1991) and Mahoney (1977) interpret disagreements as bias due to different scientific schools of thought/theoretical perspectives. All of them place the phenomenon in category A, but differently from Travis & Collins and Mahoney, Cole et al. do not recognise it as bias. We have said that category A is bias according to realism, but not according to idealism. Do Cole et al. promote idealism, while Travis & Collins and Mahoney promote realism? Not necessarily. There may be other reasons than ontological positions behind opinions on whether category A is bias or not.

In Travis and Collins' article, the reason for defining category A as bias may be taken to be based on arguments of both fairness,

relativism and realism. Travis and Collins say that cognitive particularism (i.e. 'school of thought'/category A) in grant review is much more severe than institutional particularism[63] because it directly influences the overall direction of the research field, its cognitive developments – having consequences that institutional particularism may only cause indirectly and in special circumstances. The authors seem to mean that the review system should not be allowed to discriminate or privilege a school of thought because all 'schools' are equally good (relativism) or must be treated equally well in lack of consensus. Applicants should have equal chances regardless of their school of thought (fairness). At the same time they say that 'grant applications should be judged on universalistic criteria, such as the scientific merit' (Travis & Collins 1991:325), seemingly meaning that there are standards of good research unrelated to schools of thought. 'Universalistic criteria' may be taken to mean that there are socially and culturally independent standards of good research (realism), or that one should aim against standards that are universally agreed on. Nevertheless, their conclusion must be that all research is not equally good, that the definition of 'goodness' shall not depend on particularistic criteria, and that 'school of thought' is a particularistic criterion. This implies that peer review should only use uncontroversial criteria (meaning 'black-boxed' or objective criteria) and not take a stand in ongoing debates. This is far from idealism which presents peer reviewers as the central actors in the definition and redefinition of 'good research'.[64]

Researchability may be another argument for focusing on category A bias. Arguments related to school of thought and controversial assessments may serve as indicators of this kind of bias. Arguments in panel discussions or evaluation reports that indicate that 'school of thought' is part of the assessment are likely to be an easier subject of research than the other categories of bias. Category A and possibly also category B are likely to be more openly stated and therefore easier to detect than non-professional bias. Category C, disregarding or misunderstanding the material under review, will frequently not be conscious to the reviewer, and it may therefore be hard to find indications of this kind of bias. Category D, personal interests, is a more conscious kind of bias, but as there are clear social norms against such kinds of considerations when evaluating research, information on such bias may be very hard to obtain. Reviewers may of course inform on each other, or we may find indications in the form of correlation between friendship and outcome of reviews. On

the other hand, reviewers' motives for saying that their co-reviewers' assessments are biased due to personal interests may be unreliable, and correlation between friendship and outcome may be purely spurious. Such correlation might be a good indicator of bias, but not

necessarily bias of category D. Personal interests may correlate with both professional platform, research interests and personal cognitive constraints. The four categories may consequently be hard to distinguish empirically.

Summing up this chapter, the initially posed questions have been found not to have any conclusive answers. Asking the questions: 'Is there something that cannot be refuted as the basis of research evaluation – something we all must accept as the nature or essence of 'good research'? Are there identifiable impersonal and neutral criteria for good research, or factors that should *not* influence research evaluations?', we end up with the problem of defining bias. As there are no indicators of non-biased or correct evaluation independent of peer judgements, bias in research evaluation is difficult to study.[65] The answer to what is the proper basis of research evaluation depends on whether we rely on idealism or realism. According to idealism, acceptability by the actors involved may serve as a good indicator of properly based evaluations. According to realism, there are no clear indicators of correct evaluations, unless we adopt the optimistic and somewhat naïve assumption that peer review consistently reveals the 'truth' about research quality, in which case there is no need to study bias in peer review.

Two of the categories of bias lined out in Table 2.5 have been said to be likely to be defined as bias according to idealism (category C and D), while all four categories must be said to be bias according to realism. The category least likely to be defined as bias according to idealism (category A) has been found to be the one most easily subjected to research. Idealism therefore complicates the study of bias by making the most 'researchable' category less important. As mentioned above, the 'dynamic' concept of research quality and lack of pre-set standards implied by idealism in itself complicates the study of bias. Idealism, with the view that peer review is part of a continuous process defining quality, also implies that low inter-reviewer consensus on an evaluation panel is no indication of low validity of the assessments. In fact, lack of consensus may indicate that the panel as a whole is highly competent to make valid assessments because the panel represents a large scope of the various views on what is good and valuable research (see Harnad 1985). In this view, the evaluation *process* may be a far better indicator of peer review validity than the outcome of the evaluation. A process based on tacit negotiations and compromises would probably give a far more narrow representation

of the reviewer's opinions, than either a process based on open confrontation of the divergent views or a process based on independent reviews. The decision-making of the evaluation process is the topic of the next chapter.

3 Research evaluation as decision-making

Having discussed the basis for defining good research and found an inherent problem in distinguishing biased from non-biased research evaluation, we turn to the question of possible constraints on decision-making. We will relate the constraints to the categories of bias (Table 2.5), but not deal directly with whether the constraints should be defined as bias or not.

What factors are likely to influence peer evaluators' opinions and statements on research quality? We shall first discuss social considerations and expectations guiding evaluation work in scientific communities (3.1), and then analyse the contexts in which evaluations are made: the group work setting (3.2) and the constraints that the commissioning body (i.e. a research council) may place on the evaluation process (3.3).

3.1 Social considerations and expectations in research evaluation

The discussion on the basis and norms of peer review in Section 2.3 concluded that peer review depends on tacit knowledge and craft skills. Socialisation processes renders some unity in the basis of peer review, as standards set (more or less explicitly) by the 'stars' of the discipline are internalised, for instance through graduate school. There also seems to be a common 'language' for peer review – a certain set of criteria to which reviewers (more or less explicitly) pay attention. Moreover, to be accepted, an evaluation at least has to appear to be rigorous and based on impersonal criteria of scientific merit. There are consequently limits as to what may be seen as acceptable evaluations. On the other hand, studies of peer review find low inter-reviewer agreement and indicate that the scope of possible outcomes of evaluations is rather wide. The tacit basis for assessments, the inherent uncertainty and controversy in scientific research and the individualism in scientific culture may all impede unitarian basis and standards for assessments.

In addition to norms of quality assessments, discussed in Section 2.3, evaluations may also be influenced by more specifically social expectations and considerations. What does loyalty to the involved colleagues, or to the scientific community as such, require? How harsh can you be when writing a public evaluation report? The answer may depend on various kinds of norms or considerations.

The evaluator may consider the possible undesired effects his/her conclusions may have on the evaluees. Such *tactical considerations and guards* to avoid unwanted *effects on the evaluees*, may for instance be due to provisions not to play into the hands of opponents of the evaluees (e.g. make sure not to be involved in any political mission to obliterate the evaluees), or general considerations about how public criticism may endanger the working conditions of the evaluees (reputation and funds, see bias of Category B in Table 2.5).[66] Other kinds of tactical considerations are those dealing with *effects or sanctions on the evaluator* (bias of Category D in Table 2.5). Such considerations may include avoiding unpleasant reactions/sanctions from the evaluees, from other colleagues, or from the commissioning body. An evaluator stating his/her opinions frankly might provoke the involved evaluees being the evaluator's enemies for life. Not pointing out obvious weaknesses, on the other hand, might discredit the evaluator in the views of both the commissioning body and the research community. Thus, there are two diverging considerations to take into account.

Adherence to social expectations may not involve any specific tactical considerations at all. The strongest social norms are those which are internalised and with which people comply for no other reason than that they simply prescribe 'the way things are done'. Evaluators may avoid harsh statements or any kind of overt negative criticism, simply because using harsh statements or distributing overt, negative criticism would be contrary to the evaluator's own ethos. He/she would feel uncomfortable contributing to such criticism ('internal sanctions').[67]

Social groups invoke restrictions on the behaviour of their members, and the reasons for complying to the norms of a group may vary. The rest of this section focuses on the constrains on a peer evaluator due to identity with a scholarly group (community or paradigm), as – due to his/her socialisation – the identity with this group is more likely to yield internalised and 'solid' norms, than the relation to the commissioning body or the evaluees as such. What

would be the norms of a peer group, reasons to comply with such norms, and the effects on research evaluation?

Certain rules (mostly unwritten and maybe unspoken) must be obeyed to remain member of a group. Group membership is important to most people. We normally do not like to stand alone, having no-one that shares our opinions and values (Mullen & Goethals 1987; Lysgaard 1961/1985). It may take time and resources to gain membership of a group – especially the most prestigious and influential groups of the scientific community – and having been defined an insider in one group one is likely to be automatically defined as an outsider by certain other groups. If one is a member of a scholarly community complying to a specific paradigm, there are at least two different sets of reasons for not giving up the membership easily. Firstly, there is the identity and loyalty attached to the membership. Secondly, there is the talent and career invested in this community/-paradigm, and the costs related to the process of gaining membership in a new community/paradigm. Scholars then, have various reasons for being loyal to their scientific community and paradigm, in addition to the group identity internalised through socialisation (Kuhn 1962/1970).[68]

What kind of effects may such loyalty and group identity have on the evaluation of research? As mentioned, several studies of peer review find a low degree of inter-reviewer agreement. The three studies dealt with in Chapter 2 conclude that the lack of consensus is due to divergent professional platforms (Category A bias). Such findings indicate that *'school'/scholarly viewpoint or identity is one of the main bases of peer review.* Divergent bases for assessments on a review panel need not be explicit. Controversies within disciplines resulting in divergent opinions on research quality and relevance, may be tacit. A situation of tacit controversies is open to ambiguous or double identities – a kind of context giving incentives to deliberately avoiding confrontation and consequently reinforcing the tacitness of controversy. In such a context we may expect strong *norms against overt criticism* of both members of own or other paradigms or research communities.

In Section 3.2 we discuss group effects in the more narrow and ad hoc context of the expert panel evaluation.

3.2 Research evaluation as group decisions

Panels are set to solve problems for various reasons of which representational reasons and efficiency reasons may be seen as major categories. A panel may be *representative* regarding the competencies needed to solve the problem and/or regarding the parties to the problem. Representation of competencies is aimed at the quality of the result. The idea is that the group will reach a better solution if competencies on all subjects relevant to the problem are represented on the panel. Representation of interests is aimed at the acceptability of the result. A broad interest representation on the panel is seen as a means to reaching a solution acceptable to all parties.[69]

The *efficiency* reasons for group work may not be as obvious as the representational reasons. The overall idea of group work when it comes to efficiency is that group work in some way yields a better result than individual work with the same resources. This may either be because group interaction generally increases the quality of the work or because it reduces the costs (time and other resources required). In the first case it is expected that group interaction yields better results than individual work, i.e., the sum of the work of the same persons working individually. In the second case it is expected that group work yields some sort of co-ordination profit. This may be that group work makes the most of secretarial/support services, or that groups work faster, e.g., having five separate reviewers would take more support resources and/or more time than having a group of five reviewers. In both cases (group work increases the quality or reduces the costs) a group of for instance five would be expected to work more efficiently than the five persons separately. However, such benefits are not obvious. It might be that group interaction impair the quality of the result (e.g. 'groupthink', see Section 3.2.1) or that the co-ordination of panel work demands more resources than individual work (e.g. meeting costs). There are therefore no unambiguous reasons to organise panels from this cost/benefit point of view.[70]

Seeing efficiency in relation to representation it should be added that both representation of competencies and representation of interests may work against efficiency. Having opposing interests or competing competencies/scholarly viewpoints represented on a panel set to solve a problem (i.e. agreeing on a conclusion) may incapacitate

the group. Conflicts on vital questions may lead to a deadlock of endless debates.

Summing up so far, reasons for appointing a group to solve a problem may be that the group interaction itself is expected to give a better result than individual work, representation of various competencies is expected to give a better solution, or representation of interests is expected to give a more acceptable solution.[71] At the other hand, the group level is one source of possible bias.[72] In what way a group does a better job or comes to different conclusions than the same persons would have individually, depends on the decision rules, the constellation of interests and the group dynamics – all discussed below.

3.2.1 Group effects[73]

The kind of group effects that organisers hope for when appointing panels to solve problems is of course that the group in some way is more than the sum of the participants; that *the interaction itself has qualities that enhances the review*.[74] This may be that the participants gain mutual insight through discussions, or simply that group interaction offers a situation where more information and a larger spectrum of ideas are considered by each member of the group.

Another possible effect of group work is that each member of the group *strives harder* to perform than he/she would have done working alone (Bozeman 1993:88). Translated to a peer review context, this is the situation where peers motivate each other to do a 'good' job – they are each others' 'supervisors', each others' 'heroes' to impress. In one extreme case such pressure to perform may end in a 'slaughter' of the work submitted for review. The evaluators 'may compete with one another to identify the most flaws, mistakes or problems, either real or perceived' (Bozeman 1993:89). This is like the context for appraising doctoral dissertations; the examiners are judged by their peers through their ability to find and communicate any possible flaws or weaknesses of the dissertation. In this case a likely result of panel work, compared to individual work, is a more profound/rigorous review, and lower ratings of the research under review.[75]

Group work may also yield the opposite effect. The shared responsibility for executing the task may yield a situation where no one performs – a situation of *collective shirking*.[76] Shared responsibility

then means no responsibility. In this case the result of panel work, compared to individual work, would be perfunctory review.

Yet another possible group effect is *groupthink* which may be seen as the opposite of the good interaction that organisers of panel work hope for. 'Groupthink refers to a deterioration of mental efficiency, reality testing, and moral judgement that results from in-group pressures' (Janis 1982:9). Loyalty to the group 'requires each member to avoid raising controversial issues, questioning weak arguments, or calling a halt to soft-headed thinking' (ibid.:12). Groupthink may also be seen as 'mental shirking'. Working in a group composed of highly qualified persons makes one feel confident of the quality of the work, and one may not strive as hard to detect possible flaws or weaknesses of the conclusions as one might if being charged with sole responsibility. Such an effect may occur on peer review groups as far as there is overlapping competence.

Another kind of groupthink relevant for review groups is the tendency to suppress minority opinions and arrive at a false consensus. One obvious reason for this tendency is that internal conflicts reduce the group's total efficiency and power, a situation the group tries to avoid (Hernes 1978:126). Going deeper into the social psychology of groupthink, the pressures towards uniformity in groups are explained by a drive for self-validation, i.e. a wish to establish that one's opinions are correct (Goethals & Darley 1987). Such a drive for self-validation may bias the interpretation of information in various ways, overestimating the degree of consensus in the group: group members disregard each other's objections, or they self-sensor opinions they fear wont be validated, and the others assume that silence means consent. We have a false consensus or an 'illusion of unanimity' (see Janis 1982:175). Self-censorship might also be more fundamental and impair critical thinking in a way resulting in 'actual consensus'. In this case, the members of the group censor what they *think*, not necessarily what they say.[77]

3.2.2 What kind of factors would promote the various kinds of group effects?

In the previous section, four possible group effects were outlined: (1) the interaction has qualities that enhances the review work (more ideas/information are considered by each member, or the group members gain new insights through dialogue), (2) the group members

try to impress each other and therefore work harder (or appear tougher) than when working alone, (3) shared responsibility results in collective shirking, or (4) pressures towards uniformity/groupthink. This section briefly discusses the contextual conditions of the four mentioned group effects.

Some contexts yield more opportunities or incentives for specific group effects than others do. For instance a low degree of task division in a 'flat' structure gives better opportunities for *shirking* than formal and explicit task division in a hierarchically organised group. A peer group that is homogeneous regarding research field and paradigm, but heterogeneous regarding academic status (for example containing both eminent full professors and more unknown/young researchers) may give raise to a situation where the 'unknowns' *strive hard* to perform, i.e. to impress the senior members of the group. A homogeneous group exposed to high external pressure, perceiving their task as difficult, may be a very easy victim of *groupthink* (Janis 1982). We may expect an 'especially fortunately' composed group (e.g. homogeneous enough to really take each other points, but heterogeneous enough to have diverse background information and ideas) in an 'especially fortunate' setting (e.g. enthusiasm and confidence; the group has an interesting task with no 'delicate' implications), to result in *fruitful dialogue*.

Consequently, it is possible to imagine some rule of thumb of organising for desired group effects, though such 'rules' cannot be unambiguous. Take for instance the case of trying to avoid shirking and motivating the group members to supervise and impress each other. Division of tasks may prevent the tendency of shirking. Yet, if competence is not overlapping (in which we have the 'natural case' for division of tasks), the panel members cannot really value each other performances and have no particular incentives to try to impress their co-panellists. To make all members supervise and impress each other, a homogeneous and 'flat' structured group would be preferable. If the task is not of a kind making the group members compete, however, this might be the perfect context for shirking. The group members have shared competence and shared responsibilities and may therefore all try to avoid the work, if individual effort entails no 'profit'.

Similar problems may arise when trying to prepare for a 'fruitful' dialogue and avoiding groupthink. A relatively homogeneous group, with an open-ended problem to solve, and no strict division of tasks,

may be a good structure for creative work/good discussions (i.e., new insights through dialogue/more information and ideas considered by each person). Yet, such a structure provides no guards against groupthink.

Janis finds that 'a group whose members have properly defined roles, with traditions and standard operating procedures that facilitate critical inquiry, is probably capable of making better decisions than any individual in the group who works on the problem alone' (Janis 1982:12). According to Janis the problem is that 'the advantages of having decisions made by groups are often lost because of psychological pressures that arise when the members work closely together, share the same values, and above all face a crisis situation in which everyone is subjected to stresses that generate a strong need for affiliation' (loc. cit.).

The *stress* factor that generates a need for affiliation may be highly relevant for expert panel evaluation. The evaluation task might put the peer evaluators in quite a 'delicate' situation. The scientific community to which the peers belong is likely to have high standards for review work, standards that the commissioners of the evaluation (for example a research council) expect them to live up to. At the same time the evaluation is expected to be an instrument of national research policy, but the reviewers do not know in what way the evaluation report might be used. In such a situation reviewers might be reluctant to keep high standards of impartial and rigorous review. They might conclude that the more vague and less criticising the review is, the better for their peers and the better for themselves (Larsen 1985). They are trapped between parties with conflicting expectations to the work, parties they are all expected to serve – the commissioning body and other potential national users of the report, the evaluees, the potential intra- and extra-scientific users of the work under review, and the research community as such. The best way to solve the task may be to make as little out of it as possible.

Another important point here is that the scope and amount of work to be reviewed is too large for the reviewers to live up to the standards of traditional peer review (see Section 1.2). Consequently the peer evaluators are in a situation where it is impossible to succeed, that is, to do what is expected of them. Their only 'refuge' is others in the same situation, their co-reviewers.

In such a context, peer review panels have some characteristics which may expose them to groupthink – the group members are in a difficult situation both regarding the conflicting expectation from outsiders, and regarding the nature and the amount of the work. As their only non-controversial allies are each other – the other group members – the result may be a loyalty and unanimity pressure. Janis lists various characteristics of situations likely to lead to groupthink (Janis 1982:244):

'A. Decision-Makers Constitute a Cohesive Group
B-1. Structural Faults of the Organization:
1. Insulation of the Group
2. Lack of Tradition of Impartial Leadership
3. Lack of Norms Requiring Methodological Procedures
4. Homogeneity of Members' Social Background and Ideology
Etc.
B-2. Provocative Situational Context:
1. High Stress from External Threats with Low Hope of a Better Solution than the Leader's
2. Low Self-Esteem Temporarily Induced by:
a. Recent Failures that Make Members' Inadequacies Salient
b. Excessive Difficulties on Current Decision-Making Tasks that Lower Each Member's Sense of Self-Efficacy
c. Moral Dilemmas: Apparent Lack of Feasible Alternatives Except Ones that Violate Ethical Standards
Etc.'

These characteristics do not feed directly into the context of peer review panels. Yet, seeing these characteristics in relation to the context described above, some are definitely relevant for the analysis of peer review panels. The task is difficult (B-2 2.b.) and there are high external pressures (B-2 1.). The difficulty of the task is related to a lack of methodological procedures for this kind of review (B-1 3.). Evaluation of institutions, programmes and entire research fields, is a new kind of task for which there are even less clear rules or standards than for more traditional peer review tasks. Those of the panel members who have previous experiences from similar tasks may perceive these experiences as failures (B-2 2.a.) – either because they had no effects or because they had undesired consequences.

The two first characteristics on Janis' list may also follow from those already mentioned. The difficulty of the task and the external pressures yield an insulated (B-1 1.) and cohesive group (A). Insula-

tion may also follow from the requirements to conduct an independent and impartial evaluation.

3.2.3 Decision rules/how to handle disagreements

Apart from group effects, the procedures for handling disagreements on an evaluation panel may determine in what way 'the group level' bias the outcome of an evaluation. Rules for handling disagreements may be more or less explicit. The explicitness of decision procedures may be placed on a continuum from open confrontation to sounding. At the confrontation end, we find explicit voting without any preceding exploration of opinions – a process clearly defining winners and losers. At the sounding end, we find the participants tacitly feeling out each other's opinions, systematic use of vagueness, avoiding the definition of clear alternatives, heavy emphasises on reaching consensus, and consequently no explicit voting (Olsen 1972). Confrontation and sounding may have divergent effects not only on the outcome of the decision. The procedures differ regarding the time/resources needed for decision-making and the effects on the participants/the decision-making body. Sounding may take far more time than confrontation, while confrontation may have dramatic consequences on the cohesiveness of the group.

When dissension is undesired, explicit voting or other confrontation procedures are less likely. The actual decision rules of a sounding procedure are tacit and may therefore be hard to define. They may include tacit voting, tacit negotiations/bargaining/logrolling and discussions transforming opinions. When a group has no official decision rules, any group member may tacitly set the procedure, by just behaving as if there is a rule, for example by talking as if opinion A is the opinion of the group when sounding has revealed that a simple majority holds opinion A. (See Appendix A for definitions of the various kinds of tacit decision-making.)

An ad hoc evaluation panel will seldom adopt anything but tacit decision rules. Formal negotiation procedures or explicit voting rules are rare in such contexts. Yet, tacit rules may be decisive for the outcome. A situation where the participants confer to a tacit rule of simple majority decisions, may alter substantially from a situation where agreement is obtained through bargaining and logrolling, or a situation where the unpronounced rule is that disagreements should be resolved through discussions seeking to transform opinions. Such

major categories of procedures for handling disagreements as voting, bargaining and transforming opinions may be combined, e.g. voting may be preceded by logrolling votes or discussions partly transforming opinions (Langfeldt 1998).

It should be noted that there is no direct relation between the explicitness of the decision-making processes and the room for transforming opinions. Transformed opinions may be obtained by open confrontation and discussions of the different points of view, or without confronting the different points of view. An example of the last case would be a process where all participants contribute with pros and cons for a set of alternatives without stating their points of view in advance of the discussion.

Moreover, the characteristics of negotiations may be placed on a scale from *strictly co-operative negotiations* ('the parties entertain a serious preference for joint problem-solving aimed at an efficient and fair outcome' and 'are completely attentive and open') to *pure bargaining/tug-of-war* ('brinkmanship or attempts at attrition are predominant' and 'the possibility of commitment is exploited fully') (Midgaard 1976). This important dimension of decision-making should not be confused with the confrontation-sounding dimension dealt with above. The confrontation-sounding dimension concerns the explicitness of the decision procedures, while the other dimension is related to the means by which the parties further their interests and their openness for changing their opinions.

In general, the dialogue in expert panels evaluating research can be expected to be more at the co-operative negotiation side than at the pure bargaining side of the scale. However, the degree of attentiveness and openness for changing opinions will vary, and panels may often reach a point where there is *no* room for (more) transformation of opinions, but they still have to reach a conclusion. The discussion in the following section focuses on possible solutions to such situations in different kinds of contexts.

3.2.4 Decision games and the group members' influence on the outcome

Given that at least a part of the outcome is determined by tacit negotiations/bargaining/logrolling or rules tacitly set by individual panel members, the resource and interest constellations on the evaluation panel may be decisive. The power of the individual panel members depends on their resources – their competence, the time

they have available for the evaluation, their personality (e.g. reticent versus vocal panel members) and their general rank/status on the evaluation panel (depending, for instance, on academic eminence). They may have various interests in the evaluation *process*, or in the *outcome* of the evaluation. Interests in the evaluation process may, for example, include the status resulting from being on the evaluation panel (i.e. being identified with the scientific elite), and exploring/-gaining information on the evaluation object. Interests in the outcome of the process may include all sorts of stakes in the field being evaluated (personal or scientific).

The type and constellation of interests may be important for the standing and influence of a panel member. Those with the highest stakes in the outcome will normally be the most active in using their resources and therefore the most influential members. Furthermore, 'legitimate interests' may give a right to influence, and/or there may be a norm saying that one should not use one's power in an area where one has no interests (Hernes 1978:38). On the other hand, the panel members may have loyalties or interests that 'devaluate' their rank in the group. Special loyalties to the commissioning body, the evaluees or other specific groups may define a panel member as an outsider, or someone the other panel members should guard against.[78] There may therefore be a substantial gap between a panel member's will and ability of furthering his/her interests. Furthermore, the members' interests in maintaining loyalty to the evaluation panel may hinder them furthering their individual opinions or interests and consequently reduce their influence on the outcome.

Interests may also be decisive for the parties bargaining power. To control parts of the outcome that the other party is interested in is essential for bargaining power. Two opposing parties on an evaluation panel may, for instance, barter 'authority': Party X lets party Y have the final say on the assessments of institute A, in exchange for the final say on the assessments of institute B. *Not* being interested in something the others consider decisive may offer unique bargaining power. If, for example, actor X is far less interested than the rest of the panel in avoiding dissension in the evaluation report – because actor X thinks his/her personal dissension in the report may in fact carry far more weight than the statements from the rest of the panel – actor X may get his/her will on all vital points against 'giving' nothing more

than unanimity. (X would have more to gain and less to loose from standing firm. See Figure 3.2 below.)

The composition of the panel may result in specific constellations of interests that give rise to various kinds of negotiations. Simplified, we may imagine three different situations of two parties bargaining/-negotiating about the content of the evaluation report, crudely sketched in Figure 3.1.

Figure 3.1 *Various possible constellations of interests on an evaluation panel*

A. Evaluators from different sub-disciplines or 'schools' with partly common perceptions and evaluation standards

(Assurance Game)

		Discipline X: compromise/co-operate	Discipline X: insist
Discipline Y	compromise/co-operate	4 4	3 1
	insist	1 3	2 2

B. Peers and non-peers (i.e. non-academic experts/users of the research) on the same panel: opposing interests and no taboo to disagree publicly

(Prisoner's Dilemma)

		User: compromise/yield	User: insist
Researcher	compromise/yield	3 3	4 1
	insist	1 4	2 2

C. Evaluators from different sub-disciplines or 'schools' with strongly opposing perceptions and evaluation standards (taboo to disagree publicly)

(Chicken)

		Discipline X: compromise/yield	Discipline X: insist
Discipline Y	compromise/yield	3 3	4 2
	insist	2 4	1 1

In situation A all panel members are peers, representing two different sub-disciplines or 'schools' with *partly common* perceptions and evaluation standards, realising that to get a good outcome both parties need to have a say. The best solutions for both parties would be that they both co-operate, the worst that the other party alone influences the outcome.[79] As situation A has the structure of an Assurances Game there are two equilibria: that both parties obtain their best solution (mutual co-operation (4,4)) or that they obtain their second worst solution (dissension (2,2)). *Mutual co-operation* will be the outcome if the situation is open enough for both parties to realise and trust that the other party has the same preference structure as himself. This situation provides opportunity for a strictly co-operative game and what above has been called 'fruitful dialogue', i.e. that group interaction offers a situation where more information and a larger spectrum of ideas are considered by each member of the group and the participants gain mutual insights through discussions. *Dissension* will be the outcome if one party (let this be X) expects the other party (Y) not to prefer mutual co-operation to X's unilateral co-operation, and therefore X commits to a non-co-operative strategy to avoid being the only co-operative party – which is X's worst outcome. When X is firm on a non-co-operative strategy, Y will also choose a non-co-operative strategy – realising his/her second worst outcome (dissension) instead of his/her worst outcome (Y's unilateral co-operation). It is hard to see, however, how such a situation can arise on an evaluation panel. It is generally quite clear on an evaluation panel that dissension should be avoided and conflicting opinions may be far from explicit. In a situation where the actors have partly common perceptions and evaluation standards, and realise that to get a good outcome everybody needs to have a say, dissension is not a likely outcome, even when the parties do not realise and trust that the other party has the same preference structure as himself. A more likely situation is some kind of tacit process ending in a vague evaluation report, and maybe including some false consensus (see groupthink above). A tacit process in a situation like A would mean that no opinions are presented as conflicting, and no alternative conclusions are discussed or presented.

In situation B there are both peers and non-peers on the panel, constituting two parties with opposing opinions. Both parties would like to determine the outcome of the evaluation on their own, and as neither party would lose substantial credibility if publicly disagreeing

with the other party, a 'non-unanimity' report is not an unlikely result, but both parties would prefer a compromise to such dissension, as lack of unanimity would reduce the influence of the report. 'Insisting' is the dominant strategy of both parties, and a dissension is consequently the only equilibria of the game. Depending on the circumstances, however, other outcomes are also possible. For instance, the chair of the panel may very well be in a position to achieve a compromise which all parties accept as being better than dissension. Moreover, the options of the parties need not be restricted to those given in Figure 3.1. To simplify the discussion, 'insist' or 'yield' are presented as the only options in Figure3.1, but normally the panel members may opt for different *degrees* of insisting or yielding – options on a continuum from no yielding to full yielding. One option on this continuum is to yield enough to accept that the other party's opinion is presented in the report without stating (written) dissent. That is, a report with *tacit* dissension – an inconsistent evaluation report containing the views of both parties without pointing out the disagreements. In a situation like B, tacit dissension may be a likely result as none of the parties have anything to gain from explicit dissension.

In situation C all panel members are peers, and represent two different sub-disciplines or 'schools' with *strongly opposing* perceptions and evaluation standards. Like situation B, the best solution for both parties would be that their own perceptions and standards dominate the report, but differently to situation B. The worst solution for both parties would be that they do not reach an agreement – peers do not disagree publicly.[80] In this situation the party best at brinkmanship will win. With a tacit process in a situation like C, the first party to get the impression that the other party holds another view and is not willing to compromise, will lose. If the winning party does not perceive such tacit disagreement, the winner may expect the other party to agree, also in ensuing games, and the winner of the first game therefore has good chances of winning *repeatedly*. However, if the initial losing party at some later stage in the process gets the impression that there is a chance that the previous winner may yield in one of the following situations like C, and the 'loser' mediates his/her opposing opinions, the parties may be directly confronted with their opposing opinions and their common fear of dissension. This may also open for *explicit re-negotiations* of previous 'agreements',[81] resulting for instance in a division

of authority over the various conflict issues (e.g. Y gets the final say on discipline Y, and X his/her say on discipline X). This might also be organised by the chair of the panel or some other third party to the conflict (if the chair is not considered a third party).

It should be noted that the symmetrical preferences of the parties in Figure 3.1 are hypothetical. There are no particular reasons why both parties should have the same preference structure. An example of a game where the actors have asymmetrical preferences is sketched below (Figure 3.2). Here the evaluator(s) from discipline Y has preferences as in situation C above (Chicken), while the evaluator(s) from discipline X has preferences as in situation B above (Prisoner's Dilemma). This implies that one party (Y) thinks that dissension is the worst outcome, while the other party (X) would prefer dissension to being the only party that yields.

Figure 3.2 *Game with asymmetrical preferences*

		Discipline X	
		compromise/yield	insist
Discipline Y	compromise/yield	3 3	4 2
	insist	1 4	2 1

If (or rather, when) Y realises that insisting is a dominant strategy of X, Y will yield and the outcome is found in the upper right cell of Figure 3.2 – that is, X insists and Y yields, which is the only equilibria of the game (2,4). This game illustrates the situation mentioned above where actor X is far less interested than the rest of the panel in avoiding dissension in the evaluation report and may get his/her will on all vital points against 'giving' nothing more than unanimity.

3.3 Organisational constraints: the role of the organiser

So far, we have discussed how panel members may act given the incentives put on them by norms, interest and context, but not explicitly how the overall design of the evaluation process influences such factors. In this section we discuss the role of the research council commissioning the evaluation in setting organisational constraints influencing the evaluation process and its outcome. Organisational constraints set the context of an evaluation. The selection of evaluators/composition of the panel and the kind of research subjected to evaluation may for instance directly set the operating norms and interest and thereby the opinions stated in the report. 'Operating norms and interests' may include norms of quality, loyalty, criticism and conflict-solution and the scientific and personal interests getting access to the evaluation, norms that the commissioning body might not be supposed to influence. Yet, more or less willingly and conscientiously those defining the evaluation object and composing the evaluation panel are decisive for the outcome. Units commissioning research evaluations may, however, conceive their role quite differently:

(1) The research council may see its role as *the 'neutral' organiser*, trying not to influence the outcome of the process in any way. If the commissioning body does not contract away all choices concerning the evaluation (which of cause is also a choice), such a role is not feasible. The commissioning body may still be ignorant about its influence. The commissioning body may have no idea about the effects of appointing Professor X, Y and Z for making the assessments of institute A, or the effects of appointing Professor Y as chair of the panel and Ms. W as the secretary. Such effects may indeed be unpredictable, and it may be impossible, even with hindsight, to say that given the organisational constraints set by the council the conclusions of the evaluation report ought to have been predicted. The tacitness included in the evaluation of research and the social processes on a review panel are by nature not (fully) predictable. In the meaning of not foreseeing the influence of one's choices, the commissioner of a research evaluation may therefore be 'neutral', though not in the meaning of not influencing the outcome. The research council may be ignorant, but not without influence.

(2) On the other hand, the council may see its task as controlling the evaluation process and *guaranteeing a particular kind of outcome*. This may be possible, given that the council has enough information about the research being evaluated, and about the evaluation standards/criteria and scientific viewpoints of possible evaluators. If it is also possible to foresee how the evaluators will function as a group and how they will reach decisions, the council may in fact 'order' the conclusions wanted.[82] In this case the evaluation process is symbolic. The council could have written the conclusions of the evaluation report itself, but ordered someone with more authority to write it. Yet another case of symbolic evaluation would be the known but unwanted influence. The organiser may be aware of the likely consequences of their choices, without seeing any feasible alternative way to organise the evaluation, for instance because the definition of the evaluation object may be a result of internal compromises or previous practice/standards of the council, and selecting the most eminent foreigners 'available' for the job in each area to be evaluated, may be the only acceptable way of picking a review panel. If the research council sees that this will result in an evaluation report in which they will not have confidence, or do not want to act upon, but just 'have to' carry out, the evaluation will be more symbolic than the previous case referred – the evaluation report is not even intended for use. (Nevertheless, the council may be forced to use it.)

(3) The research council may, of course, see its own role as neither to guarantee a particular outcome, nor to pursue a hands-off-policy. A third, and more evident, role would be to *ensure the quality and acceptability* of the evaluation. Such a task may for instance include ensuring that the most qualified evaluators are appointed, that all relevant scholarly viewpoints are represented on the evaluation panel, and trying to avoid obvious or unacceptable bias and undesired group effects. Furthermore, more detailed means, like a detailed mandate – including specifications of methods and criteria – or instructions on the division of task between the panel members, may serve to execute procedural control. By sticking to procedural objectives, the council may avoid directly influencing the direction of the conclusions of the report. However, such procedural influence may be a disguised way of pursuing a particular outcome. The distinction between influencing the content of an evaluation report on the one hand, and the quality and acceptability of the outcome, on the other hand, may in fact be

impossible to line up, as the quality and acceptability of an evaluation report are supposed to depend on its content.[83]

We stated in Section 3.2 that a major reason for setting down evaluation panels and not single evaluators, is that the interaction of the group is supposed to enhance the quality of the evaluation. At the end of Chapter 2 we stated that idealism implies that an evaluation panel should have broad representation of the various scholarly viewpoints in the field under review, and that open confrontations securing all 'authoritative' viewpoints a say in the conclusions are better than tacit negotiations and compromises that give a more narrow representation of the opinions on the panel. The latter kind of process may lead to a vague outcome expressing no ones opinion, or even to a false consensus. From this point of view, that game situation presented above (Section 3.2.4) which best promotes good panel work is situation A with a mutual co-operation outcome. In situation A the parties are close enough to each other to have a common basis for discussion, and they also realise that to get a good outcome both parties need to have a say, and there are consequently good conditions for entering into a fruitful dialogue.

To make such a situation possible (or even likely) the commissioning body may adopt various organisational means. First, to get the assurance game of situation A, the composition of the panel is decisive. The panel members need to have partly common perceptions and evaluation standards as a basis for dialogue. Moreover, the subject of evaluation in itself should not provoke conflicts which are incapable of being resolved. A panel set to rank or compare work within various conflicting scholarly traditions is not very likely to find themselves in situation A, if they do not all adhere to the same 'tradition' (in which case the evaluation is very unlikely to be accepted within the opposing traditions).[84]

Finally, to arrive at a mutual co-operation outcome and fruitful dialogue in situation A – and not a dissension resulting from a misunderstanding of the 'opponents' preferences – the co-ordination of the panel work, the time available for discussions and collaborative work may be decisive.

3.4 Central factors to be analysed

This section sums up the chapter, by placing the various factors in a scheme of analysis (Table 3.1). The scheme gives an overview of organisational constraints which may affect the work of an evaluation panel (column 1), of various ways an evaluation panel may approach its tasks (column 2), and of possible characteristics of an evaluation report (column 3). The underlying idea is that there are some more or less clear links between the initial planning or design/organisation of an evaluation, the way it is executed, and the result[85] of the process. Moreover, it is a point of departure that the kind of 'optimistic realism' referred in Chapter 2 *does not* apply to research evaluation processes. Research evaluation processes are more complex than a situation where reviewers simply 'see' the one and only correct conclusion to the questions asked by the commissioner of the evaluation.

For instance, the scope of the work/evaluation material and constraints on time and resources influence the possible thoroughness of the review work and the likely level of the assessments. A comprehensive evaluation object and limited time and resources set clear limits to the thoroughness and level of assessments (unit of analysis) the evaluation panel may adopt.

Another example is that when disciplines have different cultures for review work and emphasise different aspects or criteria when evaluating research, the kind of research being subjected to evaluation may, for instance, directly affect both the approach (method and thoroughness) of the evaluation panel, the criteria adopted, the report's emphasis on descriptions, explanations and assessments, and how vague the conclusions are. Some fields may have rather explicit standards and a tradition of ranking/comparison, while other fields may have tacit standards and a tradition of descriptive or explanatory evaluation reports.

Moreover, the composition of the evaluation panel is likely to influence the division of tasks between its members, group effects, disagreements on the panel and how they are handled, all of which may influence the conclusions in one way or the other. For instance, a homogenous group with no official leader may give a 'flat' structure with ample room for shirking. A group of evaluators with incentives to impress one another, on the other hand, may give a very hard-working group (and might 'slaughter' the evaluees). A group of evaluators with partly common standards in a setting of enthusiasm

and confidence may produce a fruitful dialogue, while homogenous groups exposed to a difficult task and external pressure may easily be exposed to groupthink. Ensuring specific group effects may be hard however, as explained in Section 3.2.

It should be noted that the scheme is open and tentative, and aims at an explorative study (see Chapter 1).

Table 3.1 *Scheme for analysing research evaluation as decision-making*

Organisational constraints set by the research council	**The evaluation panel's way of approaching their task**	**The contents of the report**
Scope of the evaluation: One or more Institutions Programmes Fields Kind of research: Basic/applied Natural sciences Humanities Social sciences Multidisciplinary Mandate: Are focus, methods and criteria specified? Constraints on time and resources. Selection of reviewers: Peers/non-peers The coverage of various areas and paradigms The reviewers' constraints and dispositions: competence, obligations, friends, research interests, likes and dislikes. Signals given concerning the planned use of the report.	Methods: Reviewing publications. Questionnaires to the evaluees. Ordering statistics on input and/or output indicators. Site visits/interviews. Rigorous or superficial reviewing? Division of tasks between panel members or common group discussions, writing, editing? If disagreements on the panel: Transforming opinions Voting (tacit or explicit) Bargaining/logrolling (confrontation or tacit compromises) Group effects: Work harder Shirking Fruitful dialogue Groupthink	The emphasis on: Describing Explaining Assessing Level of the assessments: The individual researcher The research group The institution The community The network Direct or indirect assessments: Based on the evaluators' own views or assessments made by others? Criteria: input, output, quality, productivity, reputation, relevance (for the research community or for society/specific user groups). Conclusions: Vague or sharp. Appraising or criticising.

4 Six ad hoc panels evaluating research in Norway

In this chapter the six evaluation processes under study are presented. For each case the background to the evaluation, the terms of reference (mandate) for the evaluation and the selection of evaluators, evaluation work and strategies, the basis and criteria for assessments, the decision-making of the panel, as well as the reactions to the evaluation report, are described. At the end of Section 5.4 there is a table for each of the cases summarising the central factors.

4.1 Peer evaluation of research fields within the natural sciences

4.1.1 Background

The evaluations studied for this dissertation were part of various evaluation policies. In the first case, an evaluation of four natural science sub-fields, the initiative was part of an evaluation plan of the former Council for Natural Science Research adopted in the mid-eighties. This evaluation policy was a modification of an evaluation model already adopted in Sweden for natural science fields, using international peer panels to assess fields and make policy recommendations to the research council.

In the files of the Research Council a 3-level purpose for the evaluation in question is mentioned: The research groups would have attention drawn to their results, the Research Council should get guidelines for policies and priorities, and national authorities should have control and documentation of public spending. There is no official statement as to why the four sub-fields were selected for evaluation. One informant from the Research Council thought the objective was to put the relative status and value (intra-scientific) of the sub-fields on the agenda. However, this was not perceived by the panel members. They seemed to perceive the selected sub-fields as

just a natural unit for evaluation, and their mandate said nothing about comparing sub-fields.

4.1.2 The selection of evaluators

The research institutions involved in the four sub-fields (mainly university departments) were consulted about the composition of the evaluation panel. The Research Council asked the concerned institutions for their opinions on suitable evaluators. Each institution was asked to propose two foreign researchers as panel members. The Council itself made the selection among the proposed candidates. The result was a group of six foreign scientists, representing the separate research fields to be evaluated, but also having general evaluation competence in the overall discipline of the four sub-fields:[86]

Professor Bowman was mainly involved in sub-field A. He had a close colleague in Norway, but had never visited Norway before.

Professor Bergström was a Swedish scientist with extensive connections to Norway and a good overview of the Norwegian research community in his area which included mainly sub-field B.

Professor Carvin was a scientist whose research area (B) was close to that of Professor Bergström's, who was also an old colleague of his. He took special interest in Scandinavian research and knew about a dozen Norwegian researchers, and took a particular interest in the research of one of these, a person he had also visited in Norway.

Professor Eckard did research within area a of sub-field C[87] and had some prior connections to the Norwegian research community through people he had met at international conferences.

Professor Oswald did research on area b of sub-field C. He had some personal contacts in Norway and had visited Norway several times.

Dr. Philips was a scientist whose main research interest was sub-field D. He also had several personal research contacts in Norway.

All the panel members, except Philips, had been involved in similar evaluations before. As co-ordinator of the panel, the council appointed a Norwegian scientist who was not involved in the fields subject to evaluation. An executive officer of the Research Council's secretariat and a doctoral student served as secretaries for the panel.

4.1.3 The terms of reference

The mandate given to the panel followed the mandate for the previous evaluations of science fields in Norway and asked for an

assessment of the scientific quality of the activity of the Norwegians groups in the field during the last decade, seen in an international context and taking available resources into account. The panel was also asked to comment on methods, give priority proposals and to review the Research Council's role in developing the field.

4.1.4 The evaluation work

The Research Council asked all the research groups in the field to provide specific material on their activity. This material was forwarded as a package to the panel members. The package included a publication list for the last 10 years, reprints of representative articles and reports, reports on past activities and future plans, surveys of available resources and the groups' self-assessments of their resources.

The panel met in Norway and visited all the departments and groups to be evaluated within 8 days. The days were used to talk with the evaluees, the evenings to discussions within the evaluation panel and writing drafts for the report. A division of tasks according to specialities was set by the Research Council by the selection of experts to the panel. As Philips put it: 'I think [the division of tasks] became clear to the group … before we met. I am sure I had a letter telling me who the other members of the group were, and you know their names and you know what their specialities were.'

In fact, the panel members received a letter containing more direct information on the division of tasks than Philips, or any of the other evaluators, remembered while talking to me 4–6 years later. In a letter to the panel members, the Research Council explicitly points out the specific fields of responsibility for each of the panel members. They were sent publications for reviewing from a broader field than indicated here. The three evaluators from sub-fields A and B were sent all papers dealing with A and B and the three evaluators of sub-fields C and D were sent all papers dealing with C and D. At times, while in Norway, the three A-B experts and the three C-D experts of the panel were also divided in two separate working groups.

The Norwegian co-ordinator of the panel had purely administrative tasks and did not take part in the assessments. Much of the report was written during the site visits in Norway. Each panel member prepared statements on the research groups closest to his own specialities. Commenting on and discussions of these drafts were to a certain degree limited to the two sub-groups of A-B experts and C-D experts.

The panel did not meet again after the site visits, and the co-ordinator, with the help of two secretaries, was responsible for putting the report together, circulating it to the panel members and making the final changes on basis of their comments.

4.1.5 The basis and criteria for the assessments

When the panel members were interviewed about five[88] years after the site visits, they were presented a list of possible criteria for evaluating research (see Chapter 5). Commenting on this list, Bergström, Bowman, Eckard and Oswald all mentioned profundity and originality as central criteria for evaluating scientific work. Bergström said:

> *'According to my view, one should never compromise when evaluating. One should always require the best possible international results. Originality, profundity, that is what is most important.'*

When asked about whether any special criteria were used for the evaluation in question, Bergström said that for the kind of institutional evaluation the panel did, one would adopt a broader perspective, a perspective in which: 'the relation between organisation, resource input, scientific quality and scientific output, is extremely important'.

Eckard said that correctness and profundity were the key criteria for evaluating research. Originality also rated high, he said, but he was also willing to let a high score on for instance the quantity of the research production, compensate for a 'certain lack of innovation'. When asked about the importance of extra-scientific relevance he elaborated on how he thought a high score on one criterion might substitute for a low score on another:

> *'I would say that everything that really brings us a good step forward in understanding the problems of life, I would rate as important and as valuable. If something is expensive and not applicable and also along rather uninteresting routine lines, everybody would have known.. would have expected that anyway from the data already available. Then I would say, well here we have a deficit. So, I think there is not a single criterion. You would have to reach multi-sided evaluations, if you wanted to be fair.'*

Oswald was less willing to do this sort of trading between criteria. After stressing the importance of correctness and profundity, he said:

> *'[T]he researcher who is very careful, or in other words very correct, needs much longer time than one who is not so correct, I would prefer the first one. But in our system, and also in yours, you have to produce a certain quantity and the conflict situation is that the higher the quantity is, the less the quality must be, because complex studies need*

time. So, from my point of view, I prefer one publication a year which is really a very good one; brings new information, is absolutely correct and has strong profundity.'

Commenting on intra-scientific relevance he said:

'In the case of [sub-field C] and [my area], I prefer those papers which [deal] with very correctness with new organisms which were not known before. I dislike such publications which are only dealing with [classifying already known organisms]. This is not what I consider an important contribution.'

Carvin stressed intra-scientific relevance as his main evaluation criterion. When evaluating a paper, he was most concerned with whether it addressed an interesting problem, but to be of good quality the research also needed to be well done (i.e. 'pure scientific criteria') and yield interesting results, he said. He had clear opinions of what counted as interesting problems:

'[L]ots of studies are simply ... descriptive studies, which aren't terribly interesting, and the other ones ... are more experimental, [which] I think, are more interesting and more insightful. So, you can just get simple divisions like that, you know. And then look at some of the groups in Norway, particularly, and say look, well these people are basically doing descriptive [B-work] the same sort which was going on 50 years ago. And you know, is this really modern [B-work]?'

Like Bergström, Carvin and Philips also made a distinction between the sheer evaluation of papers and the more encompassing evaluation of a field or a department. When evaluating a whole research field, Carvin said, one would also be occupied with other aspects of the research, aspects he saw as 'coming from a whole other side' than the quality of a paper:

'[T]his is a balanced judgement that comes down to sort of people, the facilities they have available to them, students available to them, or how the university treats them. You know, the structures of everything. You can have the greatest genius in Norway and [if] he or she isn't given adequate structural support, so to speak, they won't get anything done.'

Philips put this differently. He said that intra-scientific relevance and sound methods were the main criteria for evaluating papers, but when evaluating the work of a department he would emphasise the achievement and motivation of the researchers, the space for social and intellectual interaction in the department and to what degree the faculty completed and published their work. This view leaves room for a 'harder' evaluation than Bergström and Carvin argued for,

assessing the characteristics of a department separately, and not just taking the resources into consideration when assessing research output.

The panel did not discuss which evaluation criteria to adopt. Carvin said that such discussions are very seldom, because one seems to assume that everybody uses the same rules as oneself. However, he did not think that all the other panel members used the same criteria as he did. He suspected, for instance, that Eckard and Oswald might grade certain types of papers far better than he would, saying that some rather descriptive 'C-D'-papers that he would rate 1 on a scale from 1 to 10, they would probably rate 5. He also thought that Bowman might put emphasis on the volume of papers published by a department, while he (himself) would not be impressed by the volume of papers unless they dealt with some interesting questions.

The way Bowman expressed his views indicated that the disagreement between Carvin and Bowman was not a question of quantity as a criterion for evaluation, but rather a question of what are important and interesting research subjects. While Carvin stressed that he would not give a paper a favourable evaluation unless it addressed a problem he found interesting, Bowman emphasised cumulativity, a linkup with general questions, and profundity as central criteria, and stressed that something that is seen as irrelevant in current terms might prove to be extremely important.

While there seems to have been little if any discussion at all about the criteria for evaluating the papers, there were discussions about more departmental criteria, like group size, equipment and teaching loads. In addition there were some discussions about the use of citation indexes, as the Research Council had provided some citation and publication analyses which was to be included in the evaluation report. Bowman, Carvin, Eckard and Oswald seemed to think that citation analysis should not have been included in the report because there are so many flaws in citation analysis that they can give a completely false picture of reality. Bergström and Philips were also sceptical to such analysis because of possible pitfalls and biases, but thought that they might still provide some useful information. No one seems to have protested against the inclusion of bibliometrics. On the contrary, at least one of those who was most sceptical to such analysis in general, Carvin, explicitly praised the bibliometric appendix while commenting on the preliminary report: 'Most useful to have this included in the report', he wrote about the bibliometric appendix.

4.1.6 Evaluation strategies

Four of the interviewees from the panel talked explicitly about the politics of writing evaluations of research institutions. Bergström, Bowman and Eckard emphasised that one should try to offer an evaluation that would serve the institution – to point out the positive sides and be very careful with negative assessments. The idea is to help the groups to develop and not cause trouble for individuals. These three all seemed to agree on such a 'soft' strategy, but they expressed various degrees of 'softness'. Eckard said that the idea of evaluation is to stimulate good research by linking financing to research quality. Bowman stressed that there was too little Norwegian research in the field and too few institutions to neglect supporting any of them.

Carvin on the other hand, emphasised that an evaluation report demanding increased support to everybody would not be acted upon by the Research Council:

> *'I think, we felt it in general, that the sort of recommendations that simply said: we want more money for this and more money for this and more money for this, was just going to kind of fall on deaf ears. So, it was better to structure things that minimised the just demands for more money kind of question, without sort of losing it all together.'*

Carvin doubted the effects of these kinds of evaluation reports in terms of the Research Council really enacting their recommendations, but he still thought they had several useful functions. These evaluations give the Research Council information about the problems in the field and they are a channel for the researchers to inform the Council about their problems (in terms of scarce resources). In addition the evaluations signal to the researchers that the Council cares about how they spend their money and intends to check this regularly. Summing up Carvin said:

> *'And so it keeps the whole circle going, which I think is essential for good science, that the administrators, if you like, the higher-ups, really have some contact with what is going on down there, and what the problems are.'*

When composing the report, the panel included the names of the best researchers in the assessments, but did not mention names of any of the other researchers.[89] When asked why this strategy was chosen, Carvin answered that he could not remember how the decision was taken, but that he would have supported such a strategy because it

allowed the panel to point out who was doing the good research instead of evaluating the departments as units, which in his opinion would not make sense. He explained the omittance of negative assessments on individuals by saying that it would probably not be acceptable to 'run down people' in Scandinavia. 'So, we would praise the good, but ignore the bad', he said.

Bowman had a different account of why the panel did not distribute negative criticism to individuals. First he said: 'I think in truth we didn't find any that were bad'. Then he said: 'It isn't in anybody's interest to say: Let us wipe out this particular group and have nobody in that particular area doing what they were doing. … We tried to offer an appraisal that would serve the institutions that were there – those that could improve to improve, and those that were doing well to continue doing so.'

Bergström had a similar explanation of why the report only mentions the good researchers by name:

> *'From the point of view of the research, the most important thing about evaluation is to prepare for the future. Though the authorities are concerned to get to know whether there are some things they can cut down, that is no concern of mine. ... That is their business. Contrary to this, I am very concerned that the research we evaluate and find to be good[90] gets help to develop.'*

Eckard emphasised a related aspect in his explanation of why the panel was careful with mentioning individuals:

> *'That [careless naming of individuals] might have produced later difficulties within the groups, or might have led to disadvantages for individuals, which we didn't intend to do.'*

Summing up these various explanations of why the report only mentioned the best researchers by name, it seems that Carvin was in favour of a somewhat 'harder' strategy than Bowman, Bergström and Eckard. Carvin thought that the panel at least should point out the good researchers because it would be meaningless to give an overall assessment of a department consisting of researchers doing work of very different quality. His reasons for not distributing negative criticism to individuals, seem country-dependent. Scandinavians are used to 'soft' treatment, so one would have to be a bit 'softer' there than elsewhere. Bowman, Bergström and Eckard seemed more concerned about the possible consequences of distributing negative criticism to individuals, than about whether negative criticism was 'impolite' in Scandinavia or not. They were concerned that negative criticism could lead to worse working conditions for the researcher. To serve the evaluees one should avoid saying anything that could endanger their funding, or as Eckard put it, the evaluation report should at least not cause problems for individuals.

4.1.7 The decision making of the panel

The report is unanimous and most of the panel members claimed that there were no major disagreements in the overall assessments. There were some substantial changes in the first draft however. There was some back and forth as to what the recommendations for future research should be and the general assessments of one of the fields were rewritten. There were also some conflicting views about the

quality of the work of one of the departments within the A-B field and some of the C-D work as well.

In the files of the Research Council there is no trace of disagreement within the evaluation panel, and most of the panel members did not recall any disagreement or did not want to talk to me about them. My presentation is therefore mainly based on information from the two panel members who did seem to speak freely on this subject, Bergström and Carvin, and from the secretary. The information from these informants offers a more substantial interpretation of the data available in the files than would have been possible without such oral sources. The written material alone, that is the preliminary drafts and the correspondence of the panel members during the editing of the report, tells the story about the changes in the drafts without indicating much about the reasons for the changes.

One point of disagreement was on the value of some of the research within the C and D sub-fields. Carvin, and to some degree Bergström, did not agree to all the assessments given by the C and D experts of the panel. Carvin put it this way:

> *'I get very tired of groups and professors who, you know, are kind of making their students do all of this [C work], and I think it is something that is about a hundred years out of date. So, we would argue about this, and I would state my radical view, and back down a bit, because, you know, in a sense we did have a gentlemen's agreement with the [C-D experts]. I mean, whoever was sort of in charge, so to speak, would kind of have the most say and the others could scream and yell at them, and suggest modifications, but generally did not sort of veto their description of the place.'*

Comparing the discussions of this panel to other evaluation panels he had been on, Carvin – noting that this panel had a much more heterogeneous composition than the others he had served on – said:

> *'[T]he views were so different, and you had in fact less discussions. Because you would basically have the kind of say: Oh, I think this and this, and I think that. But you.. Discrepancies were so great that you just sort of say: Oh, well I will bow to your judgements, in this case, because you know more about it than I do. So, I think.. Yes there would be an interesting kind of relationship there; when the committee is closer together, there is a whole lot of things you argue about – that are much more detailed, if you like, techniques whatever they are. Whereas when people get further apart, you have differences of opinion that you can't.. I think, budge people from their view very much. I don't know what the real reason is. Or you just feel you don't know enough about it, you are uncomfortable with it, you think it is pure science or whatever, but you are not willing to kind of stand and fight a lot for it, because you are just uncomfortable, you don't know enough. So, you know, I think the probably, the greatest fights in science*

come between people who are doing exactly the same thing, or almost the same thing. Because only they really know the details that can be argued and discussed. At some point.. At least if you get further apart, you get less and less to disagree with.'

Bergström said that A-B experts and the C-D experts of the panel had different views and also different approaches to the evaluation. The C-D experts' assessments were quite descriptive, while the A-B experts were more evaluative, making firmer judgements, he said. He had the feeling that the C-D experts found the A-B experts' assessments a bit too demanding. He did not say explicitly that he thought the C and D sub-fields received assessments in the report that were too kind, however. All of the C-D experts on the panel stressed that there were no major disagreements on the assessments, and said that they could not remember any specific comments on their drafts.

In addition to divergent opinions between the A-B experts and the C-D experts on some of the C-D research, the A and B experts also disagreed on the assessments and recommendations regarding some research within the A sub-field. Bowman wanted a more positive assessment and more favourable recommendations on this sub-field than Bergström and Carvin. The informants' accounts on the disagreements diverge. When interviewed, Bowman did not recall any substantial disagreements.[91] Asked whether he had any comments on his draft on sub-field A, he did recall comments from other panel members and some revision of the draft. From his point of view these were comments helping him to recall details and write a 'more complete catalogue of what the Review Committee saw and heard' and he did not get the impression that anybody disagreed with his general judgements:

'I am notorious in letting details slip away. There were numerous things that embarrassingly were called to my attention that I then put into my report. But my general appraisal was pretty well accepted.'

Asked who were the most active participants in the group, Bowman mentioned Bergström, Carvin and Eckard. He especially mentioned that Bergström had a dominating influence as he had far more inside information about the Norwegian research than the rest of the panel members. Carvin was a central panel member as he was 'very widely recognised' in his field in Norway.

Each of the panel members had his field of responsibility (i.e. his own research field), and Bowman got the last say on the assessments

and recommendations regarding his field, the 'disputed' A sub-field. Bergström's and Carvin's accounts of the disagreement both give a partial explanation of the outcome. Bergström said:

> *'I sat down to write that paragraph [on this A field][92] because the kind [Bowman], he had difficulties finishing it, and didn't know quite what to write. It [my version] became maybe a bit categorical. I simply pointed to the flaws ... But [Bowman] he thought, well of course much can be improved in Norway, but if you compare it to other countries, it is not that bad at all. And as we did not see [this sub-discipline] as that essential – you know, no one was willing to invest in [it] in Norway at that time ... – so, [Bowman] he rewrote it, and I might have given some small comments on it, but it wasn't much to bother about.'*

Carvin could not remember anything about this paragraph, but rereading the report during the interview, he said that the description of especially one of the institutions doing research in this particular sub-discipline, was 'certainly very watered down' and written 'very sympathetically'. He remembered that Bowman was rather less critical of people than Bergström and himself, and also that Bowman had one particular colleague at this site whose work he highly admired, so Carvin could understand Bergström's attempt to rewrite Bowman's draft in that context. He could not remember that he had participated in this himself, however, and trying to explain why the assessments in the final report were 'not at all congruent with my memory of the place', he offered some context-dependent excuses. Recalling that this particular site was the last one they visited and that due to a special dinner[93] for the whole university that evening – to which the panel members were invited – he concluded that the panel didn't have their usual evening meeting discussing the day's site visits. This fact was particularly crucial for Carvin's participation in writing the draft, as he left afterwards for three months' fieldwork without any means of communicating with the rest of the world. When he came back and found the preliminary report on his desk, the evaluation work was no longer fresh in his memory, and he did not bother to ask for any changes.

With regard to the games discussed in Section 3.2.4, the Chicken Game (Figure 3.1 C) seems to best explain the outcome on this question. Both parties seemed to think that the best solution would be that their own perceptions and standards dominated the report, but both parties still seemed willing to yield to avoid dissension. In addition, the winner (Bowman) seems not to have perceived the controversy (until reading my draft, see note 91). If one party does not

perceive that there is disagreement he is likely to be the winner of a tacit game with the structure of Figure 3.1 C.

4.1.8 Reactions to the report

Some research groups have been provided with resources by referring to the evaluation report. The research group obtaining the best assessments and most positive recommendations for future priority within the mentioned 'disputed' A sub-field was not offered any additional resources. This provoked complaints from this group towards the Research Council.

4.2 Mixed panel evaluation of an engineering research institute

4.2.1 Background

This evaluation had some specific reasons at the same time as it was part of the more or less regular evaluation activity of NTNF, the Norwegian Council for Scientific and Industrial Research. NTNF had a plan for 'regular ad hoc evaluations' of the sector institutes. Normally these evaluations were conducted by consultancy firms, but for the institute in question, which I here will name NIE, NTNF appointed a panel of experts in addition to a consultancy report. Documents from the board of NTNF say that consultancy reports had been useful for the institutes but that it is unclear how useful this kind of evaluation has been to the Council. NTNF wanted to try a 'peer review' approach to obtain experience with a broader spectrum of methods for evaluation. The arguments against the consultancy reports were, firstly, that they were too general in their analyses of the relations between the Council and the institutes, and secondly, that the assessments of the quality of the activities of the institutes had not been at the level required by the Council.

The specific reason for evaluating this specific institute at the given time was that NTNF had reasons to believe that the research area where the institute was operating needed special attention. A previous evaluation of the field had raised questions about whether the research projects in this field were too small and pointed out the

lack of a total national strategy in the field. The evaluation was therefore meant to have consequences for the Council's research policy in the field.

NIE was a 40-year-old institute, established by NTNF, but four years prior to the evaluation it had become detached from NTNF and was now owned by the two central industrial 'branch' organisations in its area, in addition to SINTEF, the Foundation for Scientific and Industrial Research at the Norwegian Institute of Technology. NTNF-sponsored projects and general funds amounted to about 20 percent of NIE's income. NIE's main source of income was commissions from industry. As a user-owned research institute, NIE had various tasks to fulfil. In addition to executing research commissions in a broad range of disciplines, the Institute should solve current problems for industry and provide continued training for the engineers in the field.

As mentioned above, NTNF ordered both a consultancy report and appointed an expert panel for the evaluation of NIE. An initial plan seems to have been to have two parallel and integrated evaluations, but eventually the consultancy report was completed prior to the first meeting of the expert panel. The consultancy report offered both an analysis of the internal organisation of NIE and gave the market's views on the Institute (on the basis of surveys and interviews). It was pointed out that the Institute might have future problems in attracting research contractors. This might force the Institute to do more consultancy and less research, the consultancy report said.

The consultancy report advised NTNF to work out a clearer strategy for support to the sector, both by a better integration of its own activities towards this sector and by specifying the roles of the different actors operating within the sector. NIE itself was given various advice, like undertaking more technologically-oriented research, working out a strategy for operating in the international market, and educating all project leaders in marketing and communication.

4.2.2 The selection of panel members and the organisation of the work

To complement the consultancy report, a mixed panel was appointed consisting both of peers and non-peers. The Research Council first selected panel members based on suggestions from sources independent of NIE. After discussing this selection with NIE, NIE was

asked to propose members and the Research Council appointed some of NIE's proposed candidates. The aim with input from NIE was to ensure that the evaluees regarded the panel as competent. According to the Research Council official serving as secretary for the panel only a minority of the final five panel members were proposed by NIE. According to the director of NIE all final panel members except one (Director Rainert) were proposed by NIE. Being interviewed for this study all panel members except Dir. Rainert said they knew NIE fairly well, and thought that NIE itself had proposed them as evaluators.

The panel appointed for the evaluation consisted of five experts – two Norwegians and three Swedes. None of the panel members had participated in a similar evaluation before:

Professor *Svenson* was a professor at a Swedish university department in the field. He had had good connections to NIE throughout a number of years. He knew the researchers there and their publications. He also had a good relationship to NIE's director.

Professor *Olson* was a professor at a Swedish university department in the field and also had a background from industry and therefore had a twofold perspective on NIE. Due to a tight personal time schedule he only participated in the first panel meeting and wrote about one page of the evaluation report.[94] He had prior knowledge about NIE through his background from industry, where he had commissioned research from NIE. He said he thought that they did a good job then, and that he also had a good prior impression of their publications and their international reputation.

Director *Rainert* was one of the two Norwegians on the panel. He held several positions that qualified him as an evaluator on the user side of the Institute. He had a leading position in one of the largest industrial firms in the sector NIE operates in, and he was the chairman of the research committee of a major interest group organisation. Rainert was the only one on the panel who had not had any prior direct contact with NIE.

Director *Gundersen* was the technical director of one the Norwegian companies that were NIE's regular customers. He had been on various committees at NIE and said he knew both the projects and the researchers at NIE. He also had a broad interface on the user side, saying he knew the field and Norwegian industry in it very well.

Director *Carlson* was the technical director of a Swedish company that sometimes gave commissions to NIE. Except for this he did not have any special prior connection with NIE.

The *terms of reference* for the panel were twofold, asking for an assessment of the Institute's role in its sector in Norway and assessments of its scientific level ('faglige nivå') both in a Norwegian and in an international context. This twofold mandate corresponded to the two groups represented in the panel – the customers and the researchers.

The panel had no formal chairman, but as mentioned an official from the Research Council served as secretary. There were three meetings of which the two first were held at NIE. The agenda for the first meeting included presentation of NIE by its director, presentations of the conclusions of the consultancy report by its author, and discussions of the mandate, the need for data, and the division of tasks. There was also time for visiting the different departments at NIE.

A preliminary table of contents was set up at the first meeting and as no initiative was taken to divide the tasks within the group, each participant contributed where he could. Their contributions varied from one to ten pages. These drafts were discussed at the last meeting and then edited into a final report by the secretary.

4.2.3 The decision making of the panel

Being interviewed, Dir. Rainert gave a picture of the evaluation as unproblematic teamwork. There was an informal and natural division of tasks on the evaluation panel. Everybody was equally active in the work and there were no disagreements among the evaluators. In his opinion, the objective of the research was to contribute to the economic growth and the competitive power of industry, and he saw the market as the ultimate evaluator of the research.

Prof. Svenson told me that he had been worried that emphasis on industrial applicability should dominate the work of the panel, as some of the panel members had strong roots in the industry and might not understand the needs of academic research. He said he was concerned that NIE should not be strangled by industry, but enjoy some scientific autonomy. He made a special effort to get this point of view into the evaluation report. He thought he had succeeded in preventing the applicability point of view from dominating the evaluation work, and said that in the end the panel members all had a good understanding of each other's views and that the evaluation

report reflected a good balance of the different views of the parties to the research.

Dir. Gundersen said that Dir. Rainert was the natural leader of the panel. Dir. Gundersen himself had sympathy both for the arguments of scientific autonomy and the market arguments, and could not remember that there had been any disagreement about which kinds of arguments should be emphasised in the evaluation report. Neither Prof. Olson, nor Dir. Carlson, remembered any disagreements. The latter two were, like Dir. Rainert, definitely in favour of the market-arguments as criteria for evaluating NIE.

The conclusions of the report are a mixture of the different views of the panel members. It states that NIE is the leading research and competence centre in its field in Norway. The Institute is also at the international front in several areas, but to maintain its leading position it needs increased general funds for basic, long-term research – tasks that are not taken care of by contract research. At the same time it says that the main challenge for NIE is to keep its position in the market, securing continued contracts from industry. Confronted with these mixed conclusions of the report, Prof. Svenson commented:

> *'Yes, those are [Rainert's] and my opinions that appear there. So, I mean, we have not backed down.'*

To sum up, the accounts say that the panel work gave the panel members a better understanding of each other's point of view, but that they did not change their opinions.

4.2.4 The basis and the criteria for the assessments

In the letter of appointment to the experts it was emphasised by the Research Council that the consultancy report would provide the information on NIE's relations to its customers and a market analysis. The task of the panel would primarily be to assess the quality of NIE's activities on an international level.

However, few of the panel members seemed to be very interested in the data provided by the consultancy report, which consisted of questionnaires to 70 customers and phone interviews with 44 of them. The panel members dealing with the user side of the research, which turned out to be four of the five members, mainly used their prior knowledge of NIE and its sector as basis for their assessments. In addition some of them conducted informal interviews with those they

considered the central actors in the sector, or with their central contacts in the sector – 'people that they knew could speak freely' – about NIE. Annual reports, strategy documents and similar, were also useful background information.

The assessments of the scientific quality were mainly written by Prof. Svenson and based on publications, papers presented at international conferences and his prior knowledge of the research activity at NIE. As important criteria he emphasised intra-scientific novelty in terms of contributions to theory and the body of knowledge. Publishing in journals that would make the work known to the world was also important, he said.

As mentioned, three of the other evaluators were in favour of more market-oriented criteria. Prof. Olson said:

> *'What [NIE] does is more what we often call applied research. As far as I understand that is also the object of [NIE], that they are to do research that is directly applicable, not something that is speculative, that will be applicable in 20 years. ... Then extra-scientific relevance gets high priority: applicability, use, effects. When evaluating [NIE] that may be the thing to emphasise. On the other hand, if you were to evaluate the technological university, one will have to make other emphases. ...*
>
> *Q: 'What about pure intra-scientific criteria, do you ever look at that?'*
>
> *A: '... It is very difficult to make an objective evaluation of.. The best objective evaluation one can get of such research institutes, is to what extent they are able to compete on the international market. [In your own country] you always have a certain priority, you always have a certain privilege to get a commission in Norway, but if you get the same kind of commissions in France, England, Germany, the United States, then you have that [a good indication], because then one is leading. And to get such commissions it is important to publish, and that way prove that you are bright. ... and those publications say that you fulfil all those [scientific] requirements. And then of course those requirements are extremely important for doing a good job, that it really will become applicable and used. If you do not fulfil those requirements then the ... applicability will also be low. So that is very important.'*

Dir. Rainert seemed to agree with Prof. Olson. As mentioned, he clearly saw the market as the ultimate evaluator of research. In addition, he also seemed to think that not only extra-scientific requirements, but also intra-scientific requirements had some relevance: 'To get a whole, you need both' he said.

Dir. Carlson said that he was only concerned with extra-scientific relevance. He emphasised the need to separate the different audiences for research, and decide whether applicability should be assessed in

relation to the specific commissioner, industry in general or society at large. He also stressed that scientists and users/customers apply different criteria when evaluating research, as they have different time horizons. Both ways of looking upon research could be found in the evaluation report, but 'they were not particularly elaborated' in this evaluation, he said.

Dir. Gundersen was concerned both with intra-scientific and extra-scientific criteria. After saying that he would stress correctness, profundity, theoretical contributions and publication as properties of good research, he said that applicability was very difficult to assess:

> *'Who can tell whether research has applicability? Remember all the research which was not thought applicable before a long time had gone by. I am afraid that if you only emphasise applicability, you will not get the lead research that is needed to move a society forward. Because then you shut off the research to the current level of knowledge. ... [The scientific criteria], they are important, and I am very afraid of that applicability. Of course they have to do research aiming at applicability, but one must not limit the research to that. ... We would never have invented penicillin if the objective had been to invent penicillin ...'*
>
> *Q: 'What kind of criteria were emphasised in the evaluation?'*
>
> *A: 'I would like to stress the applicability of what [NIE] does towards [my kind of industry], there applicability was strongly emphasised. Of course that is more goal-oriented research and development, so there applicability plays quite another role. ... Something else that was strongly emphasised, was [NIE's] international reputation. That was an important criterion. ... At least I was very concerned that the Swedes should offer opinions on that.'*

4.2.5 Evaluation strategies

As most of the panel members were not researchers themselves, they were not concerned with the 'politics of evaluation' – what kind of evaluation strategy would best serve the evaluees – a question that concerned all the evaluation panels that consisted of researchers. Prof. Svenson was the only one on the panel that evaluated NIE who expressed concerns about adopting a 'soft' evaluation strategy. He did so indirectly, emphasising that one ought to avoid giving money only to those who are best, because they might be best precisely because

they already have good working conditions. One should be careful not to strangle the important small details of the structure, he said.

In general, Dir. Carlson was critical to these kinds of evaluation, and also seemed to favour 'harder' strategies. Talking of cronyism, he said that because of personal connections evaluations tended to be too nice to the evaluees, which meant that it is hard to interpret the meaning of a positive evaluation. If some research got a negative evaluation, however, one could be sure that it was really bad. He also mentioned that Prof. Svenson had been less critical to NIE than the other panel members.

Also Dir. Gundersen was concerned about the weaknesses of evaluations. He said that evaluations tend to be biased because those who conduct them always have some vested interest in the subject of evaluation – in one way or the other. Knowing that what you write may entail consequences for those you evaluate, can very easily affect what you write, he said.

Prof. Olson, like Dir. Carlson, seemed to be in favour of 'harder' strategies than the one chosen for the evaluation of NIE. He thought that evaluating an institute alone gave no real measure for assessment. To know how good something is, you need to see it in relation to a comparable unit, he said, advocating comparisons as the basis for evaluation.

Dir. Rainert, on the other hand, did not seem to be concerned with the weaknesses of evaluation, nor that one should adopt 'harder' evaluation strategies. He said that the evaluation report might not have told the people at NIE something they did not already know, but that putting it into print would help them to take action to improve. In other words, evaluations serve their functions by giving incentives, and not necessarily new information.

4.2.6 Reactions to the report

The director of NIE was initially not in favour of having his institute subjected to an evaluation. He seems to have been satisfied with the result however. The report has been used in 'marketing' by NIE towards its customers. The minutes from the board meeting handling the evaluation report say that the report 'tas til orientering'[95] and that the Council should continue its work to clarify NIE's strategic research role and strengthen and clarify its position towards the Ministry. The Council also asks for continuation of the work towards a more active Council policy regarding the basic funding for the institute (which

were channelled through the Research Council from a sector foundation).

4.3 Mixed panel evaluation of three social research institutes

4.3.1 Background

The Research Council was given the responsibility of channelling basic funding to a specific kind of social research institute. As part of their policy towards these institutes the Council decided that these institutes should be evaluated. There was no particular issue that called for an evaluation, but the Council perceived a general external demand for evaluating the institutes.

The stated purpose of the evaluation was to assess how the institutes functioned in their regional and national 'network of knowledge', how they fulfilled the needs of their public and private customers, and how research quality was taken care of and promoted, i.e. the Research Council wanted an evaluation both of structures, research quality and the use and applicability of the research.

4.3.2 The selection of evaluators

In addition to the usual expert panel, the Research Council appointed a seven member reference group for the evaluation. The Council selected the members of the evaluation panel without consulting the institutes to be evaluated, but the institutes were allowed to appoint their representatives to the reference group. The elected reference group included three Research Council representatives, the directors of the three institutes to be evaluated and a representative for the users of the institutes. The expert panel comprised three Scandinavians:

Professor *Hubbard* was an assistant professor at a social science department at a Norwegian university. She had experience from so-called evaluation research[96] and had also been involved with science policy studies. She was known to central staff within the Research Council who considered her particularly suited to the task. She had only limited prior information about two of the institutes to be evaluated (gained through personal contacts).

Professor *Overton* was a social science professor at a Swedish university. He was considered by the other panel members to be the one on the panel most competent to evaluate the quality of the research conducted by the institutes. He had extensive evaluation experience and a good Scandinavian network. He had no prior information about the institutes to be evaluated, but knew a couple of the people working there from other contexts.

Mr. *Ostman* was a Swedish civil servant who had been engaged in research administration. He had undertaken some policy evaluations and was considered the most competent member to evaluate the applicability of the research. He had good prior connections to the Norwegian research community, but he only knew two of the researchers in the institutes to be evaluated. Ostman was known to the Council from previous work, whereas Overton seems to have been proposed by Nordic contacts of the Council.

4.3.3 The terms of reference

The mandate given to the panel was quite comprehensive. As mentioned, structures, research quality and the use and the applicability of research should be assessed. In the terms of reference, the panel was asked to assess four topics:

- *'Whether the organisation, administration and way of financing are appropriate according to the objectives of the institute.*
- *The institutes' portfolio of projects over time, in relation to their objectives and the possibilities in the market.*
- *How research quality is taken care of.*
  The assessments of quality should encompass both the products of research and the research communities themselves. This implies assessing the relationships between the building of competence and short-term assignments, staff possibilities for professional development and their regional and national research networks.
- *How the institutes take care of mediation and contact with users, and how the contractors benefit from contact with the institutes.'*

These topics were given no ranking, and should therefore be considered as equally important. However, as the present study is concerned with how research quality is assessed, the presentation will be clearly biased, primarily dealing with assessments of research quality.

4.3.4 The evaluation work

The panel met nine times, including three site visits and three meetings with the reference group. At their first meeting, which was mainly administrative, the panel decided that Professor Hubbard should serve as chair of the panel. They selected one of the institutes for an 'exploratory site visit' and discussed what kind of secretarial services they would need.

The panel first met the reference group at their fourth meeting, just prior to the visits to the institutions. The reference group primarily had an advisory and informative mission, but its members, especially the directors of the institutes, also seem to have had some concrete impact on the emphasises of the evaluation report. The minutes of the meetings show that the directors of the institutes were very eager to provide information that was likely to moderate the statements of the report.

In addition to the material provided by the reference group, the evaluation panel applied a broad spectrum of information sources for their work: questionnaires to the institutes, site visits including interviews with administrators, researchers and customers, research publications selected by the institutes, their annual reports and other documents, government reports and general statistics.

The panel members modified the terms of reference they were given. In the preface to their report they state that they have put more emphasis on contributing to the general debate about these kinds of institutes than to give final assessments of the three institutes to be evaluated. As stated in the preface, the intention has been to evaluate this kind of institute 'as an idea, as it is expressed by' the three institutes to be evaluated.

Only four pages of the report concern quality assessments. Even here there are no concrete assessments of the research conducted by the various institutes. All statements are general and more concerned with explaining the situation of the institutes and give advice on how the research quality could be improved than assessing the research.

The closest one gets to an assessment is a conclusion saying that 'the quality criteria are promoted and taken care of relatively poorly'. The indicators of this, the report says, are that the researchers seldom publish in peer-reviewed journals and that quality promotion is sometimes put aside to acquire research contracts. This statement is clearly ambiguous. It may be interpreted as saying that the panel has

reviewed publications from the institutes and found that they do not meet the standards of good research, but do not want to say this directly; or it can be read as though the panel has no opinion of the actual quality of the research, but simply want to point out that the structures underlying the production and publications of the research are not likely to promote high quality research.

4.3.5 Evaluation strategies

Hubbard, who wrote the report, was clearly in favour of a 'soft' evaluation strategy. She said that the panel members were quite conscious that they did not want to compare the institutes. She was personally very interested in looking at working conditions and resources, she said, emphasising that it is meaningless to look at results without taking resources into consideration. When asked whether it would not also be more difficult to evaluate results than resources she said:

> *'It is also more difficult to assess research quality than to interview people about how they are getting on. I think some accused us of shirking the most difficult task. ... But first you have to map resources, then you can look at quality. ... It is not very useful to say that a product is so and so bad if you can not also say something about how it can be improved.'*

Hubbard was also concerned that evaluation reports should not be 'misused'. It would be absurd, she said, if a report written after a two-day visit to an institute should determine that institute's future. Her own soft approach can be interpreted as a strategy to prevent such use.

Overton expressed mixed views on evaluation. On the one hand, he held the view that the function of evaluation is to stimulate competition that would entail increased activity and also to build up a long-term mutual understanding of what is good and bad research. On the other hand, he seemed to agree fully on the approach chosen for the evaluation in question. Comparisons are hard to make, he said, and make no one happy, except for those getting the best scores. He also stressed that it is not easy to evaluate research quality. He compared it to evaluating a violinist:

> *'It is correct, every note, but it lacks talent, and you can hear that. So, there are some aspects that are difficult to assess – in which assessments are very subjective in research just as in art.'*

It seems that Overton agreed to the same evaluation strategy as Hubbard, but for other reasons. A soft approach was preferable, not primarily to prevent misuse of the report, but because it was the easiest way to handle the task.

Ostman did not express any specific opinions on evaluation strategies. Questioned about why the panel emphasised the working conditions and resources of the institutes without really evaluating their research output, he answered that they did initially talk about how to evaluate research quality, but soon found an interest in understanding the institutes' situation. At the end of the evaluation process they were satisfied with the approach they had adopted, and they also had no time to go into evaluating output.

4.3.6 The basis and criteria for the assessments

The evaluation report does not mention any reviews of publications. Nevertheless, Hubbard and Overton looked at some of the publications, and it also appears that they had some discussions on how to evaluate publications in applied social science. At least they both answered the same when asked about what they emphasised when reviewing the publications. They said that they checked whether a researcher was up to date on his/her topic or not.[97]

Hubbard said that most criteria for evaluating research quality are interrelated, so that it would be difficult to say which are the most essential criteria for good research. When asked what she looked at when reading the few publications from the institutes that she picked for review in her own area of competence, however, she answered that she looked specially at the reference lists to see whether an appropriate literature review had been done.[98] The main concern of the evaluation was the research environments, she said, whether the working conditions of the institutes were appropriate for fulfilling their tasks. Their main sources of information were informal interviews with the evaluees, annual reports and other documents from the institutes.

Overton emphasised that the quality of the research was not the most important output from the institutes. It was the relevance and applicability of the research to its region that was important. An important criterion to him was the researchers' attentiveness to their customers. When reviewing publications from the three institutes, he paid attention to methods and the authors' knowledge of related research in the area. He said that the publications were probably the

most important documentation for the evaluation, but an evaluation with such an encompassing mandate could not be done without interviewing the evaluees. Overton also arranged for some close colleagues to review some of the publications. These reviews were informal and are not mentioned in the report.

Ostman evaluated the applicability of the institutes' work and did not take part in the assessments of research quality. As indicators of applicability he used the institutes' ability to get assignments and earn money, in addition to the information from his interviews with selected customers.

4.3.7 The decision making of the panel

Overton stressed that even though Hubbard did nearly all the writing, he was sure that they had talked about every aspect of it and agreed in advance. According to his memory they had long discussions before they decided to look more at resources than output. He said:

> *'I can't remember any major opposing views at all. We found solutions that all of us could accept. We are all quite talkative and have our views, so the group functioned thanks to our common opinions. None of us are people who back down easily if we think others are wrong.'*

Hubbard said that it would be impossible to tell who was the most active on the evaluation panel, adding that the report was very much teamwork resulting from very inspiring 'brain storming' meetings. Asked about how they decided not to compare the institutes, she said:

> *'That was not a major topic of discussion. We talked about it and then we said that... or we were rather in agreement that we should not look too closely at each of the institutes. We very quickly agreed that we should not have separate chapters on the various institutes, as one very easily could imagine this report to have. But that was not the way we wanted it.'*
>
> *Q: 'We?'*
>
> *A: 'I think I mean the three of us in the group. In any case I did not notice that the others disagreed.'*

When asked about contact with the reference group, Hubbard said that the role of the reference group had been more to confirm the panel's choices of approach than to influence the direction of its work. The need to emphasise resources (i.e. input) more than output, was for instance confirmed by the reference group, she said.

Ostman also confirmed the picture of teamwork and agreement in the evaluation panel. The meetings were very intensive and it was hard to remember who said what as they all 'followed each other's thoughts', he said.

There is no reason to settle for this very general description of an idyllic teamwork with no discrepancy in opinions, however. The minutes from the meetings of the panel can, together with some interpretation of the interviewees' statements, tell a more nuanced story.

According to the minutes, how to evaluate research quality was not a topic until the second meeting – a meeting at which Overton was not present.[99] The minutes from the second meeting state that it was decided that research quality should be evaluated indirectly by looking at publication channels, co-publications and to what degree publications were peer reviewed or not. Direct review of publications is too big a task, the minutes state.

The minutes of the third meeting, however – a meeting at which Overton was present – say that a selection of research reports should be assessed in relation to the national and international research frontier in the area. These reviews should be made by experts in the various areas.

The interviewees gave partly conflicting information on the question of the degree to which such reviews were actually conducted. Overton said that this was his task, and that he asked close colleagues, colleagues whose opinion he trusted, to review those publications which were outside his own field of competence. Hubbard said that she reviewed a few publications in her own field of competence, and that Overton chose some people to look at some publications, but that this was done to a much lower degree than initially planned. Ostman said that assessing research quality was a task they all renounced really, and that they did not fulfil the mandate at this point.

The report says that the institutes themselves chose the publications to be reviewed, but says nothing about who actually reviewed them, or whether all the chosen publications were reviewed, or only publications in certain areas.

All in all, it seems that we can conclude that assessments of publications were done on a smaller and less systematic scale than

planned at the third meeting. Anyhow, the results of reviews are included in the report in a very vague and indirect manner.

Various explanations can be provided for this lack of explicit assessment. Firstly, the fact that Overton, the panel member said to be responsible for the quality reviews, did not attend several of the panel meetings must have had some effect on what was done, and not done, in this area.[100] We have seen that Overton expressed views on evaluation that would be consistent with a 'harder' evaluation strategy than the one chosen in the report. He said that the function of evaluation was to stimulate competition that would entail increased activity and also to build up a long-term mutual understanding of what is good and bad research. We have also seen that the indirect approach for assessing research quality chosen at the second meeting, at which Overton was not present, was changed to a direct approach at the third meeting where Overton was present. The fact that Overton was not present at the last two meetings might have been a significant factor behind the indirect approach to evaluating results in the final report – however the interviewees themselves denied the importance of Overton's absence.

A second, and more comprehensive, explanation for the lack of explicit assessments, is more in line with those of Overton's views on evaluation that calls for a 'soft' strategy, and not the 'hard' view just referred. Overton stressed three problems to which indirect quality assessments and no comparisons would be the solution: It is difficult to evaluate research quality, it is difficult to compare it, and in addition comparing makes no one happy, except for the one given the best scores. Bearing this in mind, Overton certainly had no reason to object to the approach chosen by the panel.

A third explanation, complementary to the two others, is the explanation given by Hubbard and Ostman, which emphasises that they simply adopted the approach that they found the most interesting – trying to understand the conditions under which the institutes were working. Hubbard was especially interested in this aspect, and as she was both chairing the evaluation panel and writing the report her interests in 'evaluation research' were the natural basis for the work.

There is also a fourth complementary explanation, which has already been mentioned above as an interpretation of Hubbard's views on misuse of evaluation reports. A 'soft' approach explaining the situation of the 'evaluees', instead of assessing them, is probably the best strategy if one wants to assure that an evaluation is used by

the funding authorities, not to cut the resources of the evaluees, but to improve their situation.

In addition, there might be an element of *collective shirking* (the term is introduced in Section 3.2.1) behind the lack of explicit assessments on research quality. Shared responsibility[101] for assessments might have resulted in less effort devoted this seemingly unattractive task. There is however no evidence for such a group effect. The *individual* shirking of Overton from several of the panel meetings seems, on the other hand, an important factor for explaining the lack of explicit assessments.

4.3.8 Reactions to the report

The institutes which had followed the evaluation process from the reference group seemed satisfied with the evaluation report. Except for a general focus on their problems and being asked by the Research Council to co-operate more closely with their neighbouring higher education institutions, the report had no direct consequences for their situation (according to Research Council staff responsible for the evaluation).

The Council, which also had broad representation on the reference group, found the report valuable, but not exactly what they had asked for or expected. Partly as a response to its experience with this evaluation, the Council set up a committee to clarify the criteria appropriate for evaluating applied social science.

4.4 Peer evaluation of three humanities sub-fields

4.4.1 Background

The question of evaluating studies in the humanities came up about six years after the first evaluation of fields within the natural sciences in Norway. In the meantime, the research councils' obligations to carry out evaluations had been expressly stated by the Government. The Council for Research in the Humanities started to initiate field evaluations partly as a response to this demand and partly because the idea that this kind of evaluation was an important task for any research council, had gained general acceptance within the Council.

As a model for their new kind of evaluations the Council for Research in the Humanities looked to the previous Norwegian field evaluations and adopted the peer review panel approach used for the evaluations of the science disciplines.

Policy considerations influenced the Council's selection of fields to be evaluated. The Council had objections to commence reviewing areas of research where they expected to receive a mainly negative evaluation. That would entail much negative publicity and produce resistance towards evaluations in the research community, they thought. At the same time the Council wanted to evaluate areas they thought needed evaluation, i.e. they wanted evaluations that would be constructive in the way that the evaluation report could be used to improve the standards of research. In this way they ended up evaluating fields in which they thought there was something to be done, but in which the situation was not too bad. These policy considerations illustrate a point made more or less explicitly by several of the interviewees (but not particularly in this case) that 'they just told the Research Council what they already knew'.

4.4.2 The selection of evaluators

The selection of panel members was made by Professor Hummel, a member of The Council for Research in the Humanities. He informally consulted two of the researchers in the discipline that he knew well and trusted, for their views on candidates for the panel. The conditions for the panel composition set by the council were that the panel should have a highly competent, non-Norwegian member from each of the three sub-fields to be evaluated, and both male and female members. The council also wanted a Norwegian co-ordinator for the panel. This ought to be a professor from another area in the humanities who had some research policy experience.

One foreign professor and several Norwegian professors who were asked to be on the panel, declined, and the Council ended up with the following four panel members:

Professor *Hummel* was a well-known researcher with extensive knowledge of the Norwegian research communities. He said that as he had written the terms of reference and selected the panel members himself, it was an unsatisfactory solution that he served as a panel member, but as all the other pertinent candidates for a co-ordinator refused had declined, he felt it was the best possible solution. He

stressed that the desired distance between the panel and the Council would be difficult to obtain.

Professor *Bargel* had his main competency in sub-field A. In addition to some Norwegian acquaintances, he had some prior information on the Norwegian scene through his work on committees for appointing professors in Norway. He also had prior experience from an OECD panel evaluating higher education systems.

Professor *Lawrence* had his main competency in sub-field B. He had contacts with the two major Norwegian departments in the discipline and had visited Norway several times. He had been subjected to evaluation himself, but had never been on the evaluator's side in this kind of evaluation before.

Dr. *Miller* had her main competency in sub-field C. She described herself as the youngest and least eminent of the panel members. She had little prior knowledge of the Norwegian research community and thought the evaluation would be an interesting experience and might prove useful for future connections in Norway.

4.4.3 The terms of reference

'- *To assess within an international perspective the scientific quality of the activity of Norwegian groups in [these] studies during the past decade.*
- *To assess the general ability, past and present, of the research groups to absorb new ideas and methods, and to participate in international collaboration and in the development of important new areas of research. The assessments shall take into due regard the research groups' extensive obligations to train future teachers ... and to provide them with the necessary ... skills.*
- *To assess the groups' resources, in terms of personnel and funding.*
- *If inadequacies are found, to make proposals as to what existing efforts should be strengthened or abandoned, what new areas should be embarked upon, and what resources are required to do so, taking into consideration the general need for fundamental knowledge on the one hand, and national requirements on the other hand.*
- *To give advice as to what ought to be the minimum size of the total Norwegian effort in the field ... at the university level, and in what areas or groups this effort should eventually be primarily concentrated.*
- *To review the role of the Research Council in the development of the field up to now.*
- *To make recommendations for the future.'*

Being interviewed, Hummel, Miller and Bargel said that the terms of reference had been 'stretched' by the panel. They had not only evaluated research activities, but also focused a lot on teaching and

curricula. This seemed to be mainly an effect of Professor Bargel's and Dr. Miller's concerns, and Bargel's previous OECD evaluation of education systems.

4.4.4 The evaluation work

The site visits lasted one week and included departments at Norway's four universities. In addition to the site visits the evaluation panel

gathered for a planning meeting ten months before the site visits, and an editing meeting three months after the site visits.

The Council for Research in the Humanities provided their research co-ordinator for the area as secretary to the panel. He had the job of gathering information and statistics requested by the panel. The panel was very concerned with gathering information, and the list of required material that was one of the results of the panel's first meeting, include age, geographic mobility and degrees of the faculty members, statistics on students and their careers, curricula, reading lists and teaching methods, resources for research, teaching and administration, interdisciplinary and international contacts, comparative information from other Scandinavian countries, in addition to important publications within the last five years, information on present research projects, reports from chair appointing committees, reviews of publications, *et cetera*. Some of this information was obtained through a questionnaire to the involved units and some from already available survey data.

Tasks were divided between the panel members according to their field of research. The assessments written by individual panel members (that is, the assessments of the quality of the research on various topics) encompass only six of the report's total 47 pages. The rest of the report covers descriptions of the gathered material – mainly drafted by the secretary and the co-ordinator, and revised by the rest of the panel – and conclusions and recommendations that were written more as teamwork by the whole panel.

The part of the report that dealt with the research assessments mentions the most eminent researchers by name. The explicitness and the level of details of the assessments vary from topic to topic. Professor Lawrence offered the clearest evaluative statements. His paragraphs on area B are written in a way so that one understands which departments are best and which are not so good.[102] He also says that achievements and publications are very unevenly distributed among faculty members.

This is also stressed in Dr. Miller's account of area C. In addition, she mentions the two departments where this problem is most serious. Otherwise her text mainly describes gaps in area C and gives advice on how to establish the kind of research which is lacking. Being interviewed, Miller said she compensated for her position on the panel as the most 'junior' and least experienced evaluator by working

'terribly hard'. 'I took care to be ultra-conscientious', she said, and added that she 'read everything they had published'.[103]

Professor Bargel's account of area A is the least detailed. He is mostly concerned with describing general problems and none of the departments subjected to evaluation are mentioned in his text. Half of his text deals with the differences between major approaches to A-studies. In this connection the only Norwegian scholar in the field apparently interested in the approach that Bargel thinks to be the most fruitful, is mentioned.

4.4.5 Evaluation strategies

The various views on evaluation and evaluation strategies expressed by the panel members while being interviewed correspond with these differences in evaluation styles.

Professor *Bargel* was concerned that the purpose of the evaluation should be to improve the field subjected to evaluation. In his opinion, the recommendations were the essential part of the report. He paid careful attention to what Professor Hummel (i.e. the co-ordinator of the panel and a member of the Council for which the evaluation was undertaken) said about what would happen to the report. To ensure the policy relevance of the report and that the recommendations could be acted upon without resistance from the departments, he tried to be updated on the Government's policies and to write the report in such a way that it would not provoke the researchers. Commenting on the last point he said:

> *'You don't get very far in any evaluation without comparing ... [but] the best way of dealing with comparisons ... is not to make too much fuss about it. I think it is better that a good deal is implicit rather than explicit. Because I think it alienates people if they are compared too brutally with each other. And moreover we realised that – this is a fundamental point about the whole report – that we wanted to have a report that would generally influence what people want to do, and wouldn't put them off or provoke them into feeling that it was a waste of time. So from that point of view, we could have been a good deal tougher, I think, in some of the things that we said. But I don't think that would necessarily be a very helpful way of doing it.'*

Dr. *Miller* also argued against explicit comparisons, stressing the negative effects of ranking:

> *'Initially we thought that we would have a general comment and then comments on specific universities: 1,2,3,4. And then we decided that that might not make sufficiently clear which problems were shared and which ones were exclusive. ... [And also] because*

then that is like ranking. Of course [this department] is better than [that department]. But we didn't like [the first department] to say: OK, well we don't have any problems. Let them solve their problems. ... We were not asked by [Hummel] to specifically decrease a ranking element, but what we did decide was to variate more, so that we were saying, strengths here, weaknesses there by topic. ... I suppose I thought that we might end up with what appeared to be the implicit ranking that everyone had ... And we didn't feel that that was that productive, because there were some really quite strong and interesting eyes at [the lowest ranking department], for example.'

Professor *Lawrence* who, as mentioned, offered the clearest evaluative statements on the various departments in his section of the evaluation report, was the only one on the panel who didn't stress the need for a 'soft' evaluation approach to produce a report that could be acted upon with the co-operation of the departments. On the whole Lawrence wasn't over-enthusiastic about these new kinds of evaluations either. He said that more and more evaluations go on, and they take a lot of time which people might spend doing more worthwhile things. He made an exception for cases like Norway however, saying that one probably needed an outside view once in a while to overcome the kind of cronyism that is inevitable in a little country like Norway.

4.4.6 The basis and criteria for the assessments

Like most of the other cases for my study, the evaluation panel on the humanities areas did not have any discussion about what criteria to adopt when evaluating the research. Professor *Lawrence* put it this way:

'We tried to be more practical than theoretical about it. You can ask what is research quality and how do you judge it, but we didn't get involved in those things, because the reason why we were chosen by them [the Research Council], presumably was that they trusted our judgement in these matters.'

He said that scientific criteria, like correctness and profundity, may be difficult to assess. When assessing one needs some kind of evidence and pure scientific criteria is not a matter of evidence. The evidence at hand for the evaluation was number of publications in important journals, number of books, and then one could judge these publications in terms of useful contributions to the field. His main concern was with results, originality, novelty and to a certain extent the extra-scientific relevance of the research. The properties of the surroundings and the properties of the researchers he found irrelevant for quality

judgements, but said that lacking other evidence one might look to the reputations of the researchers, that is, the opinions of other peers.

He said the main sources of information were the questionnaire material and the publications. The site visits were important to the degree that they covered the gaps in the questionnaire material. With a more precise questionnaire they might not have needed the site visits, he said.

Professor *Bargel* said that all the different criteria on my list were inter-related, but that one couldn't do without the 'purely scientific criteria'. As the main consideration of the evaluation, in addition to scientific criteria, he mentioned capacity to communicate with scholars and students and 'produce people who have got out of their university work a kind of education which will fit them to do a variety of jobs in society and make their lives richer.' He was also concerned with the relation between the potential for publications and the actual productivity, talking of productivity in terms of the quality of the publications, not the amount. He stressed that he was not concerned with citation impacts and the like. The various sources of information were all important, he said; the statistics, the departments' answers to the questionnaire, the site visits and the publications.

Dr. *Miller* also found all the categories of criteria and all the various sources of information important. She mentioned in particular dissemination of results, originality and influence on the direction of the discipline. In addition she emphasised tenacity – a goal-oriented motivation to finish work and publish.

4.4.7 The decision making of the panel

The panel seems to have had no problems reaching agreement. Commenting on this, *Lawrence* gave a general explanation of why most academic evaluation panels run smoothly:

> *'In practice there is not very much confrontation in these [kinds of] meetings. If there had been some confrontation then the whole committee would have been [unable] to work. So we had to trust one another's judgements. And when we disagreed with one another we disagreed quietly and mildly, and whoever was in the minority would give way, I think. I think that of course people vary in their strength of their judgements, the strength of their opinions. And so I think there were certain things that [Professor Bargel] had very strong views about, and we didn't really try to persuade him out of that, and then in other cases, [Dr. Miller] had very strong opinions about something and we don't [disagree] with her. We never had a strong area of disagreement. We never had somebody taking a very strong view on one side, somebody else taking a very*

strong view on another side. That would have led to some discord, but I don't think it ever arose. ... I have spent so much of my life in this sort of situation. About 20 or 30 committees for appointing people to jobs or chairs. ... All I can say is people decide to work together and the situations where people decide to oppose one another firmly almost never arise. Does that mean that we are all being too nice to one another? Maybe, I don't know. We have to get through the job, we can't argue for hours. We have to say: 'OK, you have a good opinion on that, I'll accept your opinion.' It is a system that works.'

Professor *Bargel* said that they all made comments on each other's drafts and that nobody's part of the report was written solely by one person. They had no confrontations on the panel, but their emphases were a bit different. He said he himself put more emphasis on the 'interpenetration' of the various studies:

'... how to get the whole thing working in one integrated piece. So that was my own bias, would I say, in relation to the way in which I viewed people and structures. But I didn't think we disagreed really basically about people at all. I don't think we would have made any different verdicts on relative standing of the different institutions. I think we agreed basically on that.'

Dr. *Miller* stressed that they had all been equally active in the group:

'We were almost religiously equally divided. I noticed myself in fact, suddenly looking at my watch, to be sure that we were each saying about the same amount. And I noticed that other people were speaking in a kind of measuring way, not too much, not too little.'

Commenting on the different emphases of the panel members, she said that she had a sense of Hummel and Lawrence being more concerned with scientific criteria than Bargel and herself.

On the whole, the discussions in the group seem to have had a decisive influence on how the report presented assessments, and the emphases of the recommendations. Here Bargel and Hummel seem to have been the leading persons. The text from each of the panel members was revised so that comparisons became less explicit. There are still some differences between the sections written by the various panel members in the final report, however – the sections still vary in the explicitness of comparisons. Concerning the assessments of research, each panel member used his own basis and criteria without much intervention from other members, and as they all reached very similar conclusions, basis and criteria became no object of discussion.

4.4.8 Reactions to the report

After the evaluation report had been published, there was a hearing where representatives from the various departments as well as Hummel and Bargel were present. Bargel said that at this meeting he encountered considerably more scepticism of the evaluating process than they had come across on their site visits. Hummel, on the other hand, who had seen the immediate reactions after the release of the report, said that it had calmed down before the hearing. He had been quite surprised by the first reactions, because he thought that before publishing the report they had removed everything that could make people 'jump'. He thought the report was quite meek and tame. His explanation of the strong reactions to the conclusions of the report was that the Norwegian researchers lacked a broad international perspective on their work. They lived with their own local measures and thought that their work was excellent and that their way of running a department was the only acceptable way of doing it. (The report had recommended that the chair of a department should be responsible for fostering a culture of productivity in the department.)

Asked about the follow-up of the recommendations of the report, Hummel said that he thought it had had some effects, but he did not know any details.[104] He thought I would need to pose the question to each of the units subjected to evaluation. Such a survey would exceed the scope of this study. Of more importance in this context is the fact that – at least to some degree – it was up to the evaluees to initiate and monitor the implementation of the evaluation.

4.5 Peer evaluation of a natural science research programme

4.5.1 Background

The programme in question had been given special priority by the Government. Large public funds had been invested, and the Government required that the Research Council conducted an evaluation of the results of the programme. Five other programmes of the same type were evaluated simultaneously. All these programmes were very broad and included various programme areas. They were special research

programmes – they were the research priority areas of the Norwegian Government.

The natural science programme in question covered an applied field, and evaluations of the relevance and applications of the research, as well as its scientific quality, were organised. A peer review panel was appointed to assess the scientific quality of those areas of the field where the major part of the research effort had been canalised. A consultancy firm and an applied social science institute got the task of evaluating the relevance and applications of the research in the whole field. The research policy process and the development of the programme were evaluated by another applied social science institute.

Four different evaluation reports were written: one peer review report, two different reports on relevance and applicability,[105] and one report on the development of the programme and the research policy process. The peer review report was finished some months before the three other reports. A fifth report that summarised the peer review report and the applicability/relevance reports was also produced by the consultancy firm.

The fifth concluding report describes the development of the sector, its problems and the role of research. The report concludes that the programme had not yet succeeded in realising some of the main objectives, like transferring competence to industry, increasing the profitability of industry and improving the ability to compete with other countries. It states that it seems that the priorities have lacked an overall strategy and that they have not been in sufficient accordance with the needs of the market. All in all however, the general priorities seem adequate, and one is well on the way to reach the objectives of the programme, the report states. The report also points out which programme areas get the best, and which get the worst, scores from industry, and it says that four of the eight programme areas have not been given sufficient priority.

A summary of the peer review report (written by the peer panel itself) is also included in the fifth report, but the conclusions of the peer panel – which are not in accordance with the consultancy firm's findings when interviewing people from industry – are not discussed in relation to the other findings of the report. However, the findings of the different reports are brought together on the last page of the concluding report where recommendations for future efforts are given.

4.5.2 The selection of peer panel members

The peer evaluation was organised by the Research Council's advisory committee for research in the field. This committee appointed a panel of five foreign researchers and a Norwegian co-ordinator. Candidates for the panel were mainly proposed by members of the advisory committee, but they consulted members of the relevant research communities more informally. The central criteria for appointing panel members were seniority and reputation – people whose judgements would be trusted.

The Norwegian co-ordinator, Professor *Evensen*, was familiar to the Council as the co-ordinator of a similar evaluation task. He did basic research partly relevant to the field, but had never been involved in the programme. A letter from the advisory committee to the peer panel says that Professor Evensen is to review important papers written in Norwegian. Eventually he did not review any papers. His tasks by his own choice were mostly administrative.

The expertise of the five foreign peers covered different parts of the programme. All of them had prior knowledge of the research going on in Norway and some of them had collaborated with Norwegians:

Professor Brown had extensive relations to Norway. He had, for instance, worked on a joint project with Norwegian colleagues for a couple of months in Bergen (about 30 years before the evaluation). At the time of the evaluation he had retired and was that person on the panel who had most time to spare for the evaluation work. He stayed on in Norway a week after the site visits to edit the report.

Dr. Bernard was engaged in mainly applied research. He had not been to Norway before, but knew some of the Norwegian researchers from conferences and meetings. Two of the Norwegian researchers had been on review panels that had evaluated Dr. Bernard's programme.

Professor Porter had had some relation with Norway. He had trained several of those who were now senior in the field. He had also published some papers jointly with Norwegian colleagues.

Professor Cage belonged to a different discipline than the rest of the panel. Cage had some prior knowledge of the research going on in Norway. He had also had a prior visit to one of the institutions subject to evaluation.

Professor Smith did not participate in the site visits in Norway as he was ill. His contribution to the evaluation report was based on publications and written with the assistance of Professor Brown. (Smith did not find it convenient to be interviewed by me, and as he was peripheral in the process, I did not insist.)

4.5.3 The organisation of the peer evaluation

The *terms of reference* of the panel comprised three main questions:

> '1. *How does [the] Norwegian [research in the field] compare to that of other nations in terms of scientific quality (from a limited number of representative fields, aspects, etc.)?*
> 2. *Is Norway a leader in certain areas of [the field]? More importantly, are there areas where Norway has a special advantage?*
> 3. *Are there important areas where [the] Norwegian ... research has a weak international standing and, if so, why?'*[106]

As the programme consisted of about 600 projects and eight different programme areas, a complete peer review of all the research would be an immense task. Only selected topics from five of the eight programme areas were subjected to peer review. The selection was done by the Research Council's advisory committee for the field which selected some central topics from the programme areas to which the major part of the finance had been given. The 'most appropriate institutes to represent this research' were then identified and included in the evaluation. This meant the institutes to which the majority of the money had been channelled within these topics.

Each of the five peer evaluators were chosen because of his competence within one of the five topics (also called subjects or subject areas) to be evaluated. Prior to the site visits they got a list of published articles – and a parcel of selected papers for review – within their subject area from each of the relevant institutions. Each institute involved had been asked to send 10 to 20 of their best publications within each area under review. The total number of reviewed articles varied from about 30 to more than 60 for each reviewer. In a letter to the panel members, the chairman of the advisory committee says:

'A sample of 10-20 publications/reports will be selected within each subject from a reference list complied by [the project group consisting of personnel from the consultancy firm and the social science institute]. Selection will be made primarily as a cooperative effort between the research institutions and the project group, and the publications will be forwarded to the members of the review group according to their speciality. The members may request additional publications from the reference list. The members responsible should then compose an individual report on the quality of the research in accordance with item [1)] in the mandate.'

This text does not specify the criteria for selecting publications, but all the interviewed reviewers meant that the institutions to be evaluated had selected the papers to be reviewed, and that they had selected their best publications. This crucial selection criterion is not mentioned in the evaluation report, nor are the selected papers listed. Professor Cage wanted to include a list of the reviewed papers to his chapter of the report, but he was not allowed to do this. He was not given any reason for this. Being interviewed, the Norwegian co-ordinator, Professor Evensen, and the chairman of the advisory committee could not remember this, but expected the reason for this decision to be mostly editorial, that the various chapters should have about the same sort of content – either all or none of them should include a list of reviewed papers.

The panel had a ten-day trip to the various Norwegian research units. The group split up at different sites much of the time – each member visiting the units doing research within his area of responsibility. The panel members wrote separate chapters on their own subject areas, with no, or very little, involvement from the other panel members.

The interaction between the panel members in producing the report was concentrated around composing the first four pages of the evaluation report, containing the general conclusions of the peer evaluation. These conclusions included half a page with very positive answers to the three questions set out in the mandate, general assessments (of the organisation of the research, the research facilities and the staff, careers and publishing trends), a discussion of the balance between basic, strategic and applied research, and recommendations for future research priorities.

The report's main criticism of the programme is that there is not balance between the efforts in basic and applied research. It states that future success must depend upon a stronger foundation of basic knowledge. '[G]ood results that may have been obtained by luck or

intuition, can lead to disaster without a knowledge of the underlying mechanisms.' The paragraph on the importance of basic research ends by saying that the imbalance between basic and applied research is the most serious criticism of the report. The importance of this message is further stressed by the fact that this clause is the only one in the whole report that is underlined.

In addition, the Research Council is advised to prolong the contract periods, establish a central steering committee for the whole programme and secure a better accordance between the hiring of staff and the investments in new space and equipment.[107] As mentioned, the answers to the three questions set out in the mandate were very positive. The report states that Norwegian research in the field compares favourably with other nations, Norway leads in many aspects of the field, has several potential advantages, and there are no areas of the field where Norway has a week international standing.[108]

4.5.4 The basis and criteria for evaluating the research

As there were clear-cut divisions of task in the panel and little interaction, it was up to each of the evaluators to decide the basis and criteria for evaluation. When being interviewed all the evaluators said that the site visits were essential to the evaluation.

Professor *Brown*, who edited the report, considered himself to be the least applied researcher of the panel members. He said that the criteria for evaluation would depend on what kind of research one is evaluating – pure, strategic or applied – and that applicability would be an important criterion when evaluating this applied research programme. But one would always depend on basic research to understand causes and secure success, he said. He stressed novelty and originality as important criteria, but seemed to think that all the criteria on my list would be relevant for evaluating either applied or basic research.

Commenting on the emphasis of the report Professor Brown said:

> *'You take a particular institute and look at a particular programme, probably there is one paragraph on it [in our report]. ... There wasn't time to go into great depth. And that is one of the reasons why, to some extent, that we tended to spend more time on these overall problems of space and pure and applied research: the more general things which came over to us. I think it would have been quite unfair to criticise a particular programme ... when [one] had only spent 20 minutes looking at it.'*

Dr. *Bernard* said that all the criteria on my list, except the properties of the researchers and the properties of the surroundings, were important. When asked what he thought the most important, applicability or originality, he said that he would probably put applicability on top. 'Originality always impresses me, but applicability is what drives me' he said. He thought that an article in a reviewed journal was sure to be original, correct and consistent. When evaluating, he did not 'second guess' the publications, he said. Evaluating the programme he looked at output in relation to resources (input). As an indicator of high quality, in addition to publication in reviewed journals, he mentioned attracting students from abroad. He emphasised the informal information from the site visits as important. He had no questionnaire for the site visits and took no notes. When stressing the importance of the site visits he added that the site visits would not make much sense without his prior knowledge of Norwegian research and having read the publications.

Professor *Porter* said that all the listed criteria were important in some way or the other, and he was not willing to rank them. When evaluating he put emphasis on the quality of the people and the quality of their research. As an indicator of quality he mentioned the journal in which the results were published. The best research was the innovating and exiting one, he said – the research solving a real puzzle. The site visits were the most useful source of information for his assessments. But he said that in most cases the publications had given a correct picture of what was good and what was not so good.

Professor *Cage* was the most systematic of the evaluators. He had his own list of criteria and a questionnaire for the site visits. These lists were printed in his chapter of the evaluation report. His list of criteria was extensive. The emphases were on novelty and originality, relevance to the sector, and scientific criteria like accurate descriptions of methods and logical links between cause and effects. Where the work was published was also on the list. With regard to the list of criteria I confronted him with during the interview, Professor Cage said he would put correctness and applicability on top. He said that properties of the researchers were the least important. Reputation is irrelevant as a criterion, he said. It would only back established reputation. That would be dangerous with regard to young people, he thought.

Professor Cage said that the evaluation would have been difficult without prior knowledge of what was going on in Norway. He held a

different view from Professor Porter on the relation between publications and site visits as sources of information. When arriving Norway for the site visits Professor Cage had already written his review of the publications. This report was very different from the final report, he said. He therefore assumed that the site visits were more useful than the publications. (The content of his text that deals with the publications is actually not changed very much in the final report, but emphases are different and a few points are added, and one comment on the lack of international collaboration is omitted.)

4.5.5 Evaluation strategies

There are two sources of information on the strategies adopted by the reviewers. Firstly, the texts each of the panel members composed for the report, and secondly, the views on evaluation strategies expressed in the interviews. Reading the text one finds that the evaluation style differs from chapter to chapter (each panel member wrote and signed a chapter on his area). Some chapters are mainly descriptive, while other chapters contain more evaluative statements. All the chapters give rather general assessments and, for instance, do not mention the names of any of the researchers subjected to evaluation. When asked why this general approach was adopted, the reviewers gave various reasons.

Professor *Brown* thought that the chairman of the advisory committee and Professor Evensen had been concerned that people should not be mentioned by name. The signal reaching the panel was that this was a new venture for the evaluees and that one should not make it too difficult for them. Comparing this approach with his experience from other countries, Professor Brown said: 'I think this was our Norwegian way of doing things. Softly, softly, softly sort of approach'. He also said that there wasn't time to go into great depth (quoted above).

Dr. *Bernard* felt that it was unnatural to mention names, and could not remember this as a topic for the panel discussions. Doing evaluations he would always try to be constructive, he said. He also said that he – on the basis of the limited time available for this kind of review – would not like to write anything that could have an impact on careers. He saw the purpose of the report as something quite beyond criticism or praise of the people active in the programme. Such evaluations are conducted because governments are under

pressure to prove that they are spending their money wisely. As foreign expertise usually carries more weight than a statement from the scientific communities having obtained the money, one risks what happened in this case, he said; the foreign reviewers learnt far more from the review process than the Norwegians that read the report.

Professor *Porter* seemed to share Bernard's views on the purpose of the evaluation – it was a question of accountability towards the Norwegian tax-payers. He perceived the evaluation as a sector review, and said this kind of review was very different from a review of an institution. Sector reviews are to be general and not name anybody, whereas reviews of institutions are more specific and can name the researchers being evaluated.[109] As to the approach of the evaluation report in question, he said that they were very gentle, but that they spoke their minds.

Professor *Cage* thought that it had been clear from the outset that the assessments should be general and not of individuals. It was clear from the papers they had received from Norway before the evaluation, he said. He also thought that one would need much more time for the evaluation if one were to make a just assessment on individuals. One should always be careful when putting things in writing, he said, as it could cause tremendous damage to somebody's reputation. He also said that they didn't have the time to find out anything new. This supports what Bernard said, that the reviewers learnt more than the Norwegians from the evaluation.

The variations in the text of the evaluation report produced by the different reviewers are generally in accordance with the interview data. Professor Brown – who said that the panel adopted a soft approach not to make it difficult for the researchers and that there also wasn't time to go into details, so they tended to concentrate on general issues like space and the balance between pure and applied research – also produced the most policy-oriented text. After descriptions of the various research projects and topics, some general comments on their usefulness and potential and their good quality, he expressed concern about the lack of funds for basic research and a general lack of project management.

Dr. Bernard and Professor Porter, who expressed very similar views on the more 'symbolic' purpose of the evaluation, both wrote short, general and 'soft' assessments that were in accordance with this view – that the purpose of the evaluation was to see whether it was worth spending money on the programme. However, their approaches

were different. Professor Porter's chapter is mainly descriptive. He gives accounts of the different research groups, their facilities and where the work is published. His comments contain both direct and indirect assessments. An example of the first is: '[the reviewed papers of this group] were all of good international standard.' Examples of the latter are: 'the ... group has an international reputation' and '[the articles of this group] were published in journals of the highest calibre'. All the four units evaluated by Professor Porter get good assessments and it is difficult to read any rankings of the units from his comments. Dr. Bernard, on the other hand, makes it quite clear which of the two institutions he evaluated to be the best in his subject area. This conclusion does not seem to be unexpected, as this subject had been given very low priority at one of the institutions.

The most systematic of the evaluators, Professor Cage, who set up his own list of criteria and had a questionnaire for the site visits, wrote one of the most descriptive chapters. His chapter is by far the longest of the report. It contains descriptions of the facilities, listings of the topics of the site visit presentations, descriptions of the topics of the reviewed papers, and also statistics on the language and security classification of the research reports. Comments and assessments can scarcely be found in the sections presenting the different research units, but are gathered in a discussion section at the end of the chapter. This makes the evaluative statements in the chapter particularly general.

4.5.6 The decision making of the panel

As mentioned, there was little interaction between the panel members when composing the report, except for writing the four pages containing the general conclusions and recommendations. When asked about how they reached agreements on these general conclusions and who was the most active panel member, the interviewees offered various replies. I will begin with Dr. Bernard's account which was by far the most outspoken:

> *'I think in the committee there was a definite division of power.... I think that most people deferred to [Porter] or to [Evensen] as the senior members, or the members with probably the biggest international or national or whatever reputation in their field in science. I think another member of the committee, [Brown], would like to think he would rank very highly. So there was kind of jockeying for influence.... In the case of [Brown] – he very early on in the evaluation came to the conclusion, that .. to a single*

conclusion: that there was not nearly enough basic research going on in Norway. So it seems to me that the rest of the tour he spent his time trying to prove his theory. ... I think most people were in general agreement, I think.. The one that I remember as the one being most forceful, is [Brown]. I think I was much softer, in the opinion on basic research, but that is probably because of a personal bias for applied research. I mean, I am realistic enough to realise that applied research depends upon some basic research.'

Q: 'And [Porter], in what way was he active in the group?'

A: (Laughs) 'Well, to be honest with you, there seemed to be some friction between myself and [Brown]. [Porter] seemed to be sitting back smiling, watching this and instigating things every now and then.... I think everybody agreed that there should probably be a little bit more basic research. It was a matter of degree, some people thought there should be much more.'

On the whole, this account is compatible with what Professor Brown said, but Brown stressed that it was very easy to agree on the recommendations:

'[Cage] might well think that applicability was the most important thing, and I suspect [Porter] would be somewhere in between. [Porter] would be interested in some more pure work on [certain subjects]. But his money is in [an area] which is essentially in applied. I would say I was the least applied of the group. So I would tend to be more interested in innovative, original science. The ... research [in our field] do tend to divide into two groups of people.'

Q: 'Was it your idea to write about the balance between pure and applied science?'

A: 'I think I felt strongest. But I think the others.. If you talk to [Porter] he would agreed 100% with that....'

Q: '[How did you put together] the recommendations?'

A: 'That actually went very smoothly. Just occasionally I think.. I think there was a point where [Cage] were.. ... What happened was that we each put forward what we thought would be good for future research in our own field. If you take number seven ... [Cage] would have put that in. And it wouldn't have been for us to argue that it shouldn't go in. So we each had our input into that. But of course these are incredibly general, very very non-specific goals. ...

Q: 'Which are yours?'

A: '[Number] one would have been mine, two would have been mine, and then I would have a say in eight. [Cage] would have had ten and eleven and seven. And [Bernard]

would have done three, [Porter] four and five and six. You could have put in twenty things. That was a very easy thing to agree on.'

None of the other interviewees contradicted Brown's information on who were the instigators of the different recommendations. It appears that not only the individual chapters, but also most of the recommendations were the sum of the contributions from the various reviewers – and not ideas or conclusions arrived at through discussions or negotiations on the panel. The only point where there seems to have been divergent views and discussions bringing about (tacit) compromises was the question of what emphasis should be put on the recommendation for more basic research.

A 'reconciling' reconstruction of the process based on the divergent versions offered by the informants, would be that the panel members held partly conflicting views about the need for more basic research to underpin applied research. They had no open confrontations on this issue, however, so it was not that clear that they were two against two on this issue: Brown and Porter were more strongly in favour of recommending more basic research than were Bernard and Cage.[110] As nobody made any direct objections to a strong recommendation for reallocations of funds to basic research there were no difficulties in reaching agreement – however there might have been some indirect argument and tacit negotiation on the issue, which moderated the statements in the report somewhat. The final conclusion was generally framed, and did not claim that basic research was more crucial in some areas (e.g. the areas of Brown and Porter) than in others – which might have been the outcome in case of more open discussions about the issue.

When it came to writing the various area evaluations, there was hardly any interaction between the panel members at all. The coordinator and the organiser of the evaluation offered their opinions on the drafts, but the evaluators read each other's drafts without commenting on them. Porter was concerned with making use of his broad area of competence and talked with the other panel members about issues concerning their field of competence, but he did not offer direct comments on their drafts.

4.5.7 Reactions to the report

There was no formal hearing or forum where the evaluated units could give their reactions to the evaluation report. As the report was rather general and did not capture great attention or interests in the research community, there does not seem to have been any need for such commenting either.

The evaluation was used as basis for a policy report from the advisory committee of the field. The advisory committee recommended a restructuring of the programme to improve the co-ordination of the programme, and pointed out central research questions based on the evaluation report and other sources. The panel's recommendation for more basic research has been central in later policy processes. It has still been difficult to trace concrete effects of the report on distribution of funds.

4.6 Peer evaluation of a multidisciplinary research programme

4.6.1 Background

The background of this programme and its evaluation were similar to that of Case 5. The programme had special priority from the Government and it had asked the Research Council to evaluate its results. However, it was not the same research council that conducted the two evaluations, and they were organised in different ways. The evaluation of this multidisciplinary programme was divided in two: one report on the development of the programme and the research policy process – produced by an applied social science research institute – and one peer review report. The Research Council's secretary for the programme served as secretary for the peer panel, and as a link between the two different evaluations. As for Case 5, this study only deals with the evaluation conducted by the peer panel.

4.6.2 The selection of evaluators

The research institutions involved in the programme were not allowed to propose members for the panel. The Research Council asked the other Norwegian research councils involved in the programme and a

Swedish research council to propose candidates for the peer panel. The panel was finally appointed by the chairman of the Council organising the evaluation. The five members appointed to the panel were all Nordic professors in the disciplines concerned with the multidisciplinary programme to be evaluated.

One of the five, Professor *Davidsen,* was a Norwegian scientist holding a chair at a foreign university. As the only Norwegian of the group Professor Davidsen also served as a co-ordinator for the panel. The panel members had separate specialities within the topics covered by the multidisciplinary programme, but some of their topics were overlapping. Davidsen's area of specialities within the programme to be evaluated for instance overlapped with the specialities of Professor *Jensen,* while Professor *Ellis'* area was more clearly limited, as she had her training from another discipline than the rest of the panel and had very few common research interests with the rest of the group.

The two remaining panel members, Professor *Carlin* and Professor *Ulberg,* had their training in the same discipline as Jensen and Davidsen, but they both had their special competence within specific areas of the programme to be evaluated which did not overlap with the competence of any of the other members of the panel. In addition, Carlin had a broad interest and competence in most of the topics of the programme. He and Jensen were the only ones who felt a bit 'at home' in Ellis' area.

All the panel members had prior knowledge and relations to the research communities in Norway, as they were all part of Nordic research networks. Ulberg was the one with the least prior knowledge and fewest connections to Norway.

4.6.3 The terms of reference

'... give an assessment of the preliminary scientific results of the research and communication, including effects on research training and recruitment, and characteristics of the results regarding productivity, quality, novelty and (potential) applicability ...' The panel was also asked to evaluate the programme's success as a policy tool and to give advice concerning future priorities.

4.6.4 The evaluation work

The panel met five times and in total was assembled for eight days. At their first meeting the panel members divided the subject areas of the programme between them, and set up a questionnaire to be sent to the research groups active in the programme. They were very concerned that they should only include research that had been sponsored through the programme.[111]

During the visits to the various institutions active in the programme, all the panel members took part in the meetings – regardless of whether they concerned 'their' area or not. Including the Norwegian co-ordinator, they all wrote their separate chapters of the report on the topics which they had been assigned at the first meeting.

The various drafts were put together and edited by the co-ordinator with assistance from the secretary and a draft for the report was distributed to the panel members for comments. This draft was then discussed and revised at a final meeting. Just after this meeting the panel presented their preliminary report (orally) at a public conference that was arranged to publicise the programme. The final editing was then carried out by the co-ordinator, Davidsen, and the secretary, with some assistance from Carlin who worked at the same place as Davidsen.

4.6.5 The basis and criteria for the assessments

Professor *Davidsen* emphasised intra-scientific relevance as the most important criterion for the evaluation. In addition, results and extra-scientific relevance had importance. Properties of the researchers and the surroundings were not relevant in such a context, he said. He added that though intra-scientific relevance was the main criterion for the assessments, reading the evaluation report one might get the impression that extra-scientific relevance was the main criterion.

As a basis for the assessments, Davidsen emphasised the publications from the programme and the interviews with the research co-ordinators. He also made a few informal contacts with people in the research community to gain information. He did not seem to think that the site visits were crucial to the report. He said that the rankings of the various research groups would have been the same without the site visits. However, he thought that some groups probably had received a more favourable assessment than they would have without the site visits.

Professor *Carlin* said that there were three different types of research in the programme: research to map the diffusion of problems, research analysing the causal mechanisms of problems, and research to reduce the problems. Different kinds of criteria for evaluation would have differing degrees of importance for these various kinds of research. He thought that the various panel members had adopted criteria appropriate to the different areas of the programme.

'Pure scientific criteria', such as methods, would be of importance for all the research in the programme. He seemed to mean that when evaluating the extra-scientific relevance of a programme that had not yet had the time to have extra-scientific impacts, more scientific criteria – like connections to previous research, adequate methods, good analysis and presentation of results, and possibility of replication – could be used as indicators of future applicability of the research. When evaluating the programme he was also concerned about the properties of the researchers and organisational criteria, like recruitment, reinforcement of networks and the quality of the co-ordinators.

As a basis for the assessments, he mentioned his prior knowledge of the research areas, the publications of the programme and information gained during the site visits. He also had separate conversations (i.e. without the presence of the rest of the panel) with some researchers in one of the areas in order to gather specific information.

Professor *Ulberg* strongly emphasised publication in international journals, stringency and theoretical contributions as important criteria for evaluating research. He said Davidsen and Carlin had made him modify the emphasis on international publications in the evaluation of a specific research topic. International publication was not adequate for this topic, they had said.

In addition to publications, Ulberg stressed the interviews with research co-ordinators as important for the assessments. He also asked for the opinions of colleagues in his own country on the status of the Norwegian research.

Professor *Jensen* said that different criteria applied for the various kinds of evaluations. Each type of research and research institution had to be assessed according to its specific purpose. For the programme in question, extra-scientific relevance was primary, he said. However, research institutions with different kinds of tasks were involved, and they had to be evaluated accordingly. The university departments could be assessed on purely scientific criteria, but the

applied institutes had to be evaluated according to applicability for their customers.

As a basis for his assessments, Jensen said that the site visits were essential. He also used opportunities for informal information. He did not put much emphasis on reading the publications of the programme. He said it was not realistic to do a systematic evaluation of all the written material, and only read thoroughly those items he thought especially interesting.

Professor *Ellis* said that results – in terms of theoretical contributions, productivity and publications – were the main concern when assessing research programmes. That the results had to fulfil scientific criteria – like consistency and profundity – was self-evident. Regarding the importance of intra-scientific and extra-scientific relevance, she was concerned that small countries should concentrate their research[112] on topics that were special to their country and not do all kinds of analyses that are 'fashionable' internationally at the moment (but she emphasised that it was very important to be internationally oriented in the way that one was up-to-date on the research from other countries and participated in the international research community). In that way one could become really good at particular areas and 'export' high quality research to the international research community. This would both yield good research and relevance for society at large. She mentioned originality, continuity and cumulativity as features of good research. Publications were the most important basis for the evaluation report, but the overall impression provided through the site visits was also very useful for the report, she said.

4.6.6 Evaluation strategies

Davidsen seemed to be in favour of more outspoken evaluations. He was concerned about the pressures placed upon evaluators to adopt 'soft' evaluation strategies.

> *'People place quite strong restrictions upon themselves when writing [evaluations] ... I don't know if one should. Research communities are small ... I don't think one really wants to express one's opinion if a project is bad. One doesn't get further than vague statements. Then it is up to the reader. Reading vague statements makes one understand that maybe this is not so good. ... Due to various ... considerations – for instance you don't want to scorn the adviser or the whole group – you wrap up your critique, for instance use the structural context to explain why some institutes are better than others.'*

Carlin said that the approach of the report was a subject of discussion for the panel. He and Ellis tended to favour an approach that was more detailed on the project level, while the others made more overall statements, he said. When editing the report these differences in approaches were moderated in favour of the overall style. Carlin thought that the final report was neither a good policy evaluation, nor a good micro-evaluation, but that this was probably the most useful approach seen from an outsider's point of view.

Jensen said that it was not reasonable for a visiting panel to produce an evaluation of the results of research. It should be on a more general level, on the development of the field and so on. He said that he had adopted a much softer evaluation strategy than Davidsen, whom he thought had adopted an unduly harsh approach on his drafts. He thought it more constructive to point out what was good than to criticise specific researchers or groups. He added that he thought the responsibly of the co-ordinator of a panel (i.e. Davidsen) should be to co-ordinate and edit, not to write on one area of the evaluation.

Ulberg said that peer panels tended to write kind evaluations, as they normally were asked to evaluate their friends and international colleagues. He said that if the politicians asking for the evaluations would choose to interview the peer panel in an informal setting, instead of reading an evaluation report, they might get to hear some 'truths' that would not be put into print.

Ellis seemed to be in favour of a 'soft' evaluation strategy, not causing unnecessary harm to people. At the same time she said that it was very annoying when colleagues expected her to write favourable assessments of their research so that they could obtain a position they had applied for. She said it was easier to do evaluations in foreign countries in the sense that she did not have to worry so much about reactions to the report.

4.6.7 The decision making of the panel

The evaluation report produced by this panel is the least vague among the six cases studied. There is no explicit ranking of the concerned institutes and research projects, but it is not difficult to read out of the text which units were better. This panel was also the one with the most explicit disagreements among its members. These disagreements seem mainly to include two of the panel members, Professor Davidsen

and Professor Jensen, and concerned the assessments particularly of one of the institutes active in the programme. That both Davidsen and Jensen were quite willing to talk about the matter, and also provided me with copies of their correspondence on their controversy, also demonstrates that the controversy was open and explicit.

Being interviewed, Professor Davidsen was quick to say that the starting points for the discussions on the panel were not reflected in the final report. He said that from the discussions during the site visits he had got the impression that there were no major disagreements between the panel members. Disagreements didn't come to the surface until the panel members had read each other's drafts, he said. Particularly one of the panel members had been critical to his first draft, saying it was too harsh in its assessments, Professor Davidsen told me, and continued:

> *'The others didn't take a clear stand. ... So it was an internal discussion between two persons in the group. It was solved of course.'*
>
> *Q: 'How?'*
>
> *A: 'I moderated my critique. There weren't very many other possibilities. There were some rather clear accusations that personal conflicts had influenced the [my] draft. I thought that that was a totally unreasonable interpretation of course.'*

Later on I got the story from the point of view of the other party to this disagreement, Professor Jensen. He started giving me his views on evaluation, research policy and the importance of applied research. He had an extensive background in academia, research administration and sector research (applied research), he told me, continuing:

> *'I was the only person on the panel with that kind of orientation. The others were classic, pure university people who weren't concerned about applicability when assessing the research. That was what brought about some very strong tensions. I, for instance, from my point of view, could not make sense of the assessments [Davidsen] gave the Norwegian research. It was nonsense*[113] *because there is a pure research orientation and then there is an applied research orientation. ... The others used traditional criteria – number of publications in prestigious international journals. I revolted against that in the group.'*

Later on, talking about writing the report he said:

> *'I thought [Davidsen] was too close to the Norwegian scene and it was very clear to me that he had old enemies in Norway who should be punished through this evaluation. ... There I intervened and said that if that was to be into print I should be allowed to*

write naughty things about some people ... within the [B area].[114] *I had chosen the strategy of pointing out what was good and not so bad. ... General critical comments on individual researchers serve no purpose.'*

The data leave no doubts about the nature and outcome of the decision making. This is an example of group decisions resulting in moderated evaluation statements. The nature of the disagreements between Jensen and Davidsen are rather complex. Professor Jensen gave three reasons why he could not accept Professor Davidsen's assessments. He disagreed with the underlying criteria used by Davidsen (applied research should not be evaluated by the criteria of basic research), he suspected that Davidsen's draft had a personal bias (that he wanted revenge on old enemies), and he disagreed with Davidsen's evaluation strategy (harsh criticism will harm people and serves no good purpose). The last was also a consideration of fairness. The report would not be fair unless one used the same evaluation strategy for all the concerned units.[115]

Davidsen said that because of the accusation against him he had no other choice than to moderate his assessments. These accusations were made in writing in a letter from Professor Jensen, dated a week before the final meeting. The letter was distributed to all the panel members. In addition to criticising the draft report, Jensen in this letter said that he could not accept it and threatened both to write a dissenting minority statement to the report and to rewrite his own chapter of the report, giving the 'B area' harsher criticism.[116]

4.6.8 Reactions to the report

The Research Council made a plan for the continuation of the program on the basis of the evaluation report. The main ideas of the report were influential: further support especially to some 'young underdeveloped areas', and giving priority to the most qualified groups. The program structure was also simplified, as suggested, but in another way than the report recommended.

5 Case comparisons: analysis with emphasis on common features

The structure of this chapter follows the scheme of analysis (Table 3.1) starting with the last column then going 'backwards' to see in what way the process or the organisational constraints may explain the outcome. First, the outcome – in terms of the assessments and criteria of the six evaluation reports – is analysed (5.1). Then, the effects of approaches and processes are analysed. Lastly, the influence of the organisational constraints on the process and the outcome is discussed. Overviews of the data are given in Tables 5.2 to 5.7 – the scheme of analysis with data from each of the cases.

5.1 The evaluations reports

5.1.1 Assessments 'between the lines' and focus on various units

All the reports have some *descriptive* parts. They describe research topics, resources, composition of staff, and similar, of the units under review. One might also expect all research evaluation to contain explicit *assessments* of the research, but as mentioned, Case 3 is an exception and avoids this.[117] Half of the cases have substantial *explaining* elements. In Case 2, the evaluation of an engineering institute, the report explains why certain structures are essential for the institute. In Case 3, the evaluation of social science institutes, the report places particular emphasis on explaining why the framework conditions of the institutes are unsatisfactory. In Case 6, the evaluation of a multidisciplinary programme, the report to some degree explains why some projects and units involved in the programme have been more successful than others.

A central common characteristic is that the reports *emphasise what is positive* and are very careful with negative assessments – except indirectly, for example, by stating that more resources are needed or

simply by omitting evaluative statements on the not-so-good units or projects. They all avoid explicit grading or ranking of the evaluated units. This is done to varying degrees. The report of Case 3 gives the reader no idea of which institute is the better or the worse. In the other cases such information might be read between the lines, Case 6 being the most explicit.

Regarding the *'unit of analysis'* of the various reports, the two most similar are the two field evaluations. They both evaluate sub-fields by looking at the research groups at the various institutions in the field and are the only evaluations mentioning individuals, i.e., the best researchers are mentioned by name in the report. Their common approach may be explained by the fact that both panels were provided with copies of previous evaluations of research fields.

The two institute evaluations assess different levels. The panel evaluating the engineering institute assesses the institute's expertise in various areas pointing out which are the best and which need to be strengthened. The evaluation of the social science institutes evaluates 'the kind of institutes' without assessing the various units or their research areas.

The two programme evaluations also differ. In Case 6 the programme areas are evaluated by looking at the projects sponsored by the programme at the various units/institutes involved, while in Case 5 the field and the sub-fields as such – not the programme/programme areas – are the primary subject of evaluation. The emphasis on assessments on the level of research units or projects varies between the reviewers in Case 5 (and also in other cases, but not as clearly as in Case 5 where there was no editing of the individual drafts).

The conclusion to draw from this lack of correlation between the scope of evaluation (discipline/field, programme or institution) and the units focused upon in the evaluations, is that the scope of the evaluation does not necessarily influence what levels or units the panels focus on when assessing the research (see Section 5.3). However, when the given scope is related to a given model or institutionalised 'unit of analysis' (for instance, from previous reports), the scope might be important for the units focused on in the reports (as for the field/discipline evaluations that were provided with copies of previous reports, see Section 5.3.1).

5.1.2 Tacit criteria and 'evaluee-supporting' conclusions

When it comes to the question of whether the assessments are *based on* the evaluators' own opinions and direct assessments or opinions of other informants/indirect indicators, there are more clear differences between the various kinds of evaluations. The two mixed panel evaluations of applied institutes (Cases 2 and 3) both depend explicitly on information from external experts or users in addition to the panel members' own judgements. The two field evaluations, on the other hand, both appear to be based only on the panel members' own (direct) assessments. No external experts were explicitly consulted. The programme evaluations give a more mixed picture. The evaluators in Case 6 and most of the evaluators in Case 5 said that they based their assessments on their own direct judgements. The exception was one applied scientist evaluating applied science in Case 5 who emphasised indirect indicators.

In general, the applied-oriented evaluators found it important to consult other experts or take the general viewpoint (i.e. the 'reputation') into account, while academics seemed to think that their own assessments were sufficient. Some academics also said that letting others influence the assessments they had been asked to conduct would be wrong; one of them also referred to a kind of 'Matthew effect'. Evaluations based on reputation reinforce present differences in status and possibilities of obtaining funds for the researchers/units under review, and might do so without any basis in the actual standard of their research. Yet, if academics have an impression of breaking some tacit rule when basing their assessments on the opinions of others they might hesitate to be frank about their evaluation practices. The data *might* therefore say more about differing *rules* for academic and non-academic research evaluation than about substantial differences in actual *practice*.

Few of the reports explicitly state which *criteria* underlie the assessments (as mentioned, Case 3 and one of the areas in Case 5 are the exceptions). Central parts of the interviews with the evaluators dealt with the question of the bases of the assessments. A list with the following concepts was given to the interviewees as a starting point for talking about evaluation criteria:

Table 5.1 *List of criteria given interviewees*

Pure scientific criteria:	consistency, correctness, stringency, profundity, etc.
Results:	theoretical contributions, productivity, publications (how much, where)
Intra-scientific relevance:	relevance of subject, novelty, originality, cumulativity, citations
Extra-scientific relevance:	applicability, use, effects
Properties of the researchers:	achievement, motivation, ambitions, reputation, international position
Properties of the surroundings:	equipment, freedom, group size, financing, organisation

Confronted with this list the evaluators reacted differently. Some talked about the bases of evaluation in terms of evidence and clear criteria for good research, and gave the impression that evaluation work is a straightforward and easy task for them. Others said evaluation work is difficult and/or that they base their judgements more on feelings or impressions of the quality of the research, and were not willing to set up a list of criteria. The majority answered somewhere in between these two positions.

Evaluators within the social sciences and the humanities (Cases 3 and 4) were generally more reluctant to rank or point out central criteria than evaluators from other areas. The reactions of evaluators from other areas varied more. Some evaluators in Case 1 (natural science fields) seemed quite certain in their choice of criteria, not hesitating to point out a few criteria as the most important, while others on the same panel emphasised that the criteria were interrelated and that they were all important in some respect.

The evaluators of the engineering research institute (Case 2) were less concerned about pointing out specific criteria, but they had firm (and divergent) opinions about the *general* bases and objectives of evaluating research. Some emphasised the market as the final evaluator while others emphasised the need of quality assessments by peers.

Most of the evaluators of the natural science programme (Case 5) had less firm opinions. One panel member emphasised that criteria depend on the kind of research under review, another said that scientific quality is difficult to assess, a third said the criteria were

difficult to rank, a forth had no difficulty pointing out his central criteria.

The reactions of the evaluators of the multidisciplinary research programme (Case 6) were also mixed. Two of them said that criteria depend on context. Others on this panel had no problem doing a general ranking of the criteria/pointing out the central ones.

Trying to conclude something about the criteria underlying the various evaluations despite such ambiguities, we see that output (i.e. results), and intra- or extra-scientific relevance seem to be central dimensions in most of the evaluations (see Tables 5.2–5.7). Some kind of output and intra-scientific relevance was central to all evaluations, except Case 3. The members of this panel did not think that intra-scientific relevance was central for the evaluation of applied social science, neither did they assess the research output explicitly. Extra-scientific relevance, on the other hand, was, as expected, a dimension in all evaluations dealing with applied research (Cases 2, 3, 5 and 6).

Getting down to *the criteria for assessing these dimensions* the picture is more obscure. As mentioned, there are some differences in clarity between the cases. Some of the panel members evaluating natural sciences emphasised criteria like novelty and profundity (see especially Case 1), while panels dealing with humanities/social science were more often unwilling to pick out specific criteria. This may indicate a somewhat clearer basis of assessments within the natural sciences studied than within the social sciences and the humanities (Becher 1989; Whitley 1984). Another possible explanation is that natural scientists (in the studied research fields) think that assessments *should* be criteria-based, and therefore speak in terms of perceived common criteria, like novelty and profundity, even though the nature of their assessments might not be substantially more based on explicit criteria than assessments in social science or humanities. The fact that panels of natural scientists did not discuss the basis or criteria of assessments while the social science panel did, gives little help in answering the question. That the natural scientists were experienced in this sort of evaluation task and found no reason to discuss the basis of this, is compatible both with obscure, tacit and personally based assessments, and with clear and standardised assessments. However, the variations in answers and the way answers were framed, indicate tacit and personal assessments also in natural science fields studied.

As the criteria to use were not explicitly discussed (except in Case 3) on the evaluation panels, the panel members were often unaware

of the different opinions among the panel members about the basis of assessments. Yet, some informants were good at guessing other panel members' emphases. This indicates some reliable 'prejudices' about colleagues' opinions or some kind of tacit communication between the panel members about the basis of assessments. In Case 3, which ended up not assessing quality, the panel seems to have had the most profound debate about how to assess the quality of the research under review, indicating that criteria are discussed when assessments are conceived as difficult and also that such discussions may stress the problems and blur explicit assessments rather than promoting them.[118]

In accordance with the 'gentle style' and overall positive assessments, the conclusions of the reports are positive and can generally be said to praise the research under review. The degree of praise varies as mentioned. The report in Case 3 is too vague to be said to praise (or criticise) the research or the units under review, but still gives a positive impression of the units. The report in Case 5 is by far the most positive – distributing close to unstinted praise. All reports contain some policy recommendations, mostly in terms of more resources to particular areas/efforts. Four of the reports also emphasise the importance of basic research. The reasons for such positive and 'evaluee-supporting' conclusions are analysed below.

5.2 The evaluation approaches and processes

5.2.1 Methods: site visits and reviewing publications

The *primary* sources of information varied for the six evaluations. All the evaluation panels conducted site visits and reviewed research publication. Case 3 might be said to be an exception when it comes to reviewing publications. The panel only made some sporadic reviews and the evaluation report does not directly refer to the evaluators' opinions on the research/publications. All panels also received various kinds of written information from the research groups, either at the site visits or mailed in advance (overview of publications, staff, resources, organisation, strategic documents, etc.). Other methods adopted were questionnaires to the evaluees (Cases 3, 4 and 6), interviews with external informants (Cases 2 and 3), and statistics on input and output indicators (Cases 3 and 4). A reference group

consisting both of external and internal[119] informants was appointed in one case (Case 3) and served as an information and discussion forum for the evaluation panel.

What does the information sources say about the focus of the evaluation? In the four cases using questionnaires to evaluees and/or information from external informants, site visits and review of research publications must in some sense be less dominating information sources than in the two cases not using such information (the two natural science cases, Cases 1 and 5, in which publication review and site visits were the only major sources). However, the implications of such additional information sources vary. In Cases 4 and 6 (one field and one programme evaluation) the questionnaire information largely complemented the peer review part based on publication review and site visits, while in the two institute evaluations (Case 2 and 3) the use of a variety of information sources meant less emphasis on the review of scholarly quality.

5.2.2 'Peer-supporting' evaluators

As mentioned, the evaluation panels treated the evaluees 'gently'. Positive aspects are emphasised and one has to read 'between the lines' to discover negative aspects. Four of the evaluations (Cases 1, 2, 3 and 5) also defend the research they evaluate against heavy external pressure by recommending more resources for basic research. Generally the evaluators were concerned about the potential use of their evaluation report, i.e., they considered the political context of the evaluation while writing their reports.

Most peer evaluators saw their role as a helper primarily for the research under review, not the Research Council itself. The evaluees were their primary allies, not the body commissioning the evaluation. Not all evaluators held this point of view, and their emphasis on the need of protecting the interests of the evaluees varied. But as experts promoting a gentle approach were represented on all panels, no report contained any kind of harsh criticism. *Evaluators seeing their role as helpers of the evaluees were one reason for the vague reports. When putting together the reports the peers avoided details that would not serve the kind of policy they thought appropriate.*

The peer evaluators were in this respect 'allies' of the evaluees. They wanted to help the 'needy' units, as well as securing support to good units.[120] In such a situation vagueness serves a *political* end, i.e. giving the evaluators some control over the potential allocative use of

the evaluation report – at least 'preventing' funding authorities from using the report to cut down funds – by emphasising the evaluees' needs of more resources.

Assessments on a departmental/group level may also bring up the problem of avoiding to rank good researchers in a 'bad' department/group according to its low average. One solution to this problem was to avoid explicit ranking or comparisons of groups/departments, and in this way avoid an evaluation report that would appear as unfair for some of the evaluees (and also avoiding comparisons/ranking of groups departments that might be used for reallocations that would punish good individual researchers in departments/groups that in sum were below average). Another way to handle the problem was to mention the good individual researchers by name in the report, and thereby direct assessments more to the individual level than the departmental level.

5.2.3 Interaction/discussion: Controversies handled through compromises

Disagreements on the panels were an additional reason for vague reports. *The resolving of disagreements invariably resulted in more vague reports.* Dissension in a public evaluation was seen as undesirable by the panel members. Dissension might be seen as detrimental both to the *authority* of the report and to outsiders general *confidence* in the research area under review,[121] i.e. the peer reviewer's research area. One way to reach agreement on the final draft was to avoid all critical comments that not all panel members could agree on, resulting in (more) positive reports, as in Cases 1 and 6. In these two cases, the party promoting the most negative evaluation yielded, and a consensus report was obtained despite the lack of agreement among the panel members. In Case 6, one panel member won partly by threatening to write a dissension. In Case 1, the negotiations were far more tacit and dissension was never an alternative. Those least willing to invest time and effort on the evaluation process 'yielded'. The panel member most willing to invest time and effort was the one working in the sub-field subjected to controversy, and he got the final say – writing a positive assessment and apparently without perceiving that two other panel members had opinions clearly diverging from his. In Chapter 4 it was suggested that this situation can be understood as a Chicken Game where the winner has not perceived the controversy

(discussed in Section 3.2.4). With regard to the controversy in Case 6, it was suggested that part of the explanation of the outcome was that the parties had asymmetrical preferences (as illustrated in Figure 3.2).

In both cases (1 and 6), the yielding of the most negative panel members resulted in less criticisms in the evaluation reports. The panel members held divergent views on what, if anything, might be criticised. They all had a possibility to dissent to other's assessments, but a wish to avoid dissension. The result was more positive but somewhat vaguer reports, i.e. reports with less distinction between the various units under review in terms of good or bad research. In both these situations there was some overlap of competencies that allowed some intervention from other panel members. In situations with less overlap the pull toward kind conclusions may be even stronger. Because scholars are likely to defend their own kind of research, a panel member defending a unit or area under review is likely to be closer to this research than the potential opponents are likely to be, most likely giving the 'defender' the last word, as the potential opponents do not know the area under controversy well enough to legitimately contest the defender's assessments (e.g. the tacit disagreement between the A-B experts and the C-D experts in Case 1). In general it may be supposed that more is at stake for the defenders than for the (potential) opponents.

Another way to handle disagreements was to include the various divergent opinions in the final reports, without presenting them as divergent or saying that they represented the views of particular panel members (see especially Case 2). This corresponds to the tacit dissension outcome of the Prisoner's Dilemma situation that was described in Section 3.2.4 – an outcome where everybody gets his/her opinions included in the report (and no one yields or attacks). This kind of 'agreement' resulted in more pluralistic reports containing all opinions represented on the panel, but not necessarily any clearer conclusions, as the various opinions were not seen in relation to each other.

What group effects on the panels can be identified? Four kinds of group effects were discussed in Chapter 3:

(1) The interaction has qualities that enhance the review work, for example that more ideas/information are considered by each member, or that the group members gain new insights through dialogue.

(2) The group members try to impress each other and therefore work harder (or appear tougher) than when working alone.

(3) Shared responsibility results in collective shirking.

(4) The group situation leads to uniformity/groupthink, including impairment of critical thinking, less rigorous review and suppression of minority opinions/false consensus.

If we limit the analysis to group effects regarding the assessments of the research under review, there is little evidence for the first kind of effect. As the processes on the evaluation panels were characterised by division of tasks, little interaction on assessments and mostly tacit compromises in case of disagreements, the opportunities for *dialogues leading to new insights* were limited. In Case 1, one panel member said that he always learnt a lot in these kinds of panel meetings with so many different views represented. However, this was the only panel member mentioning such learning. (Other panel members put emphasis on what they had learned from the research under review.) This does not mean that dialogue on the panels did not affect the assessments. In Case 6, for example, panel interaction lead to modified emphasis on publication in international journals and more overall and less detailed assessments. In general, there was much more dialogue on how to solve the evaluation task and on the policy conclusions, than on the assessments of research as such. Especially in Case 3 the data give the impression of fruitful dialogue on the bases for evaluation and the content of the evaluation report. In Case 2, on the other hand, it was stated that the panel members obtained a better understanding of each others' points of view through the dialogue, but they did not change their opinions.

In the case with some overlap of competence and open confrontation on the panel (Case 6), one panel member *worked harder* to get his point of view into the report. The same was the case with the peer representative on the only panel with a majority of non-peers (Case 2) who made a special effort to get his point of view on the importance of scientific autonomy into the report. In both cases the special efforts affected the outcome. This phenomenon is not equivalent to the group effect (2) above. More weight was attached to good argumentation, but not necessarily to more thorough review work. The latter phenomenon, a more thorough review to make a

good impression, seems to have appeared for a panel member in Case 4 – a 'junior' panel member with no prior experience with this kind of evaluation.

The material contains two cases of individual *shirking* (Cases 2 and 3), but nothing pointing towards collective shirking, except maybe for Case 3 where review of research quality was reduced to a minimum.

The phenomenon of *groupthink,* which by its very nature is hard to detect, has not been identified in any of the cases. It can be concluded, however, that the kind of setting and processes found – division of tasks and little interaction on assessment – tell us that groupthink was not very likely to dominate the evaluations.

5.2.4 Contextual norms

The evaluators' accounts say a great deal about what is seen as the correct way of conducting research evaluation in various contexts. Some of the contextual factors are obvious. The restricted *time* and large *scope* of the evaluation set limits as to what could be done, and reduced the evaluators' ambitions to do a rigorous review of the evaluation material. Compared to micro-level evaluations (like grant review and manuscript review) the reviews were rather superficial. One evaluator of a programme evaluation (Case 6) said, for instance, that he looked through the research publications and read the abstracts but had no ambition to go through it all. In another case, the scope of the evaluation was seen as far too encompassing and only a little part of the programme, consisting of more than 600 projects, was selected for review (Case 5). The detail of review also vary between members of the same panel, indicating that in addition to scope and other contextual factors, various individual factors influence the rigour of the reviews (group effects may also influence the rigour of review, as explained above).

The combination of *peer* evaluators and *public* reports also seems to foster some contextual rules. As the evaluations might be acted upon by authorities outside the scientific community, to consider the possible effects of the evaluation report was seen as part of the task, and most of the peer evaluators were careful not to write anything that might harm the resource situation of the evaluees. As mentioned, several interviewees emphasised that they tried to offer an evaluation that would serve the institutions – to point out the positive sides and be very careful with negative assessments. In this way the interests of

the field/institute/programme or specific researchers were taken care of by strategic emphases in the evaluation report.

The strength of such contextual norms may vary between disciplines and between countries. Some non-Nordic evaluators interviewed said that they perceived Nordic evaluations as less outspoken than evaluations in their own countries. As panels consisting of Nordic experts were not 'softer' than the international panels, this would indicate that non-Nordic experts with a more outspoken evaluation tradition complied to 'Nordic norms' – adopting a less outspoken style. If so, the site of the evaluation was more important than the nationality of the experts. However, the data does not allow any conclusions on national differences. Several non-Nordic experts emphasised general needs to be careful in making negative assessments – arguments having nothing to do with complying to local norms. Some of these experts were from the same country as experts saying Nordic evaluations were less outspoken.

In Section 3.1 several central contextual considerations for peer evaluators were discussed. Most of these can be found in the accounts of the interviewed peer evaluators and help explain both the 'gentle' approach and the 'evaluee-supporting' conclusions of the evaluation reports. Section 3.1 points to various reasons for scholars for being loyal to their scientific community and/or paradigm. The strongest factor is perhaps the identity with a field, and/or a 'school' or paradigm – internalised through training and social interaction. There might also be more concrete (but maybe not much more conscious) concerns linked to protecting your field and/or 'school', regarding career etc. – guarding the value of your invested talent. Such loyalty and identity may explain a broad range of peer review behaviour. Firstly, scholarly viewpoint and identity matter, and might be decisive, for judgements on the adequacy and value of research. Secondly, loyalty to a field is likely to make the evaluator consider the effects of assessments and recommendations, as these will regard his or her field and colleagues (the evaluees). If a peer evaluator has to choose between loyalty to the evaluees or the body commissioning the evaluation, loyalty to the evaluees is the likely choice as they represent his or her field, at least if the evaluees share the evaluator's scholarly viewpoint.

Many of the interviewed peer evaluators seemed to understand their gentle approach as a result of obvious, legitimate norms of

government-commissioned research evaluations: the peer evaluator's role is to help promote the research in the field under review. However, in the interviewees' accounts of the motives of their antagonist panel members, we have seen that gentle and/or positive evaluations may also be understood as biased and subjective – i.e. not an indisputable, legitimate approach. This was notably not anyone's understanding of his/her own assessments, but his/her understanding of another panel member's assessments.[122]

Whereas none of the evaluation processes studied resulted in harsh reports, other kinds of contexts might have led to such results. The context that is the most likely to lead to an evaluation report distributing *harsh criticism* is an evaluation within a research field containing various 'schools' that define good research in clearly conflicting ways – and there is no co-operation or loyalty between the parties – where the evaluation panel belong to another 'school' than the evaluees and consider the evaluees as non-constructive antagonists competing for the same resources as their own 'school'. In such cases there is no particular reason why an evaluator should not put his/her opinions into print without polishing them (se also Section 6.5.2). In the cases I have studied, however, these kind of 'unpolished' reports were avoided because the panels contained persons covering all relevant 'schools' concerned. As mentioned, compromises between the evaluators (overt or tacit) removed harsh comments from the final reports.

Cases 1 and 6 would have contained rather clear criticism if the more 'negative experts' on the panel had written the assessments alone. 'Negative experts' include those who did not mind pointing out what they did not think worth supporting. 'Tough' peer evaluations do occasionally appear in Norway. Two examples are an evaluation of Norwegian work research (NORAS 1992) by a one-man 'panel', and an evaluation of computer science (NAVF 1992), both provoking critical response from the evaluees (see also Section 7.4). One might end up in a tautological argument claiming that these evaluations were harsh because the 'panels' were homogenous, if the indicator of a homogenous panel is that the panel is able to agree on harsh criticism. It is a fact however, at least according to the hearings/reactions to the reports, that these evaluation 'panels' did not contain representatives from the traditions that were criticised.

5.3 Organisational constraints

Having studied the content of the six evaluation reports and the various evaluation approaches and processes, the question remains in what way organisational constraints affected the evaluation processes and products.

5.3.1 The scope and subject of the evaluation

The first factor in the scheme of analysis (Table 3.1) is *the scope of the evaluation task*. The size and number of institutes, fields or programmes selected for evaluation might influence both the evaluation process and the content of the evaluation report. As differences in scope are not evident in the evaluations studied, case comparisons regarding this issue are not easy. Selecting my cases I expected the two field evaluations to be larger and more encompassing than at least the single-institute evaluation. Measured in number of faculty or projects under review, there is no evidence that the work-load for each evaluator varied substantially with the various kinds of evaluation (evaluations of institutes, programmes or fields). The number of researchers or 'R & D person-years' at the three institutes in Case 3 (85 person-years) are close both to the number of R & D person-years at the one institute evaluated in Case 2 (about 100 person-years) and to the number of permanently employed faculty in the fields evaluated in Cases 1 and 4 (71–100). The size of the two programmes (Cases 5 and 6) are more difficult to measure in number of researchers. One programme contained 349 projects, the other 602 projects. For the largest programme (Case 5) only some subject areas were selected for review and as the evaluators reviewed 30-60 articles each, the scope of the evaluation did not differ considerably from the other cases.[123] However, judged by the responses from interviewees, it is somewhat more difficult to get an overview of large national programmes than a field limited by a set number of people in a set number of departments. With this exception, the size of the evaluation task should be considered a constant in the present study, i.e. a variable that is not likely to have affected any variation in processes or outcomes of the six cases. Therefore, we cannot conclude anything about how the size of the evaluation task for instance affects what unit/level of assessments the panel adopt, except for the obvious conclusion that the

assessments of these evaluations are all more macro-level, and not as detailed when it comes to the assessments of research outcomes, as for instance peer review for scientific journals.

There are variations in the object of the six evaluations that would naturally relate to the unit/level of assessments – whether *institutes, programmes or fields* are being assessed. However, no clear relation between the scope/object of the evaluation and the unit/level of assessments can be found in the material. The scope/object of the evaluation in terms of institutes, programmes or fields, does not necessarily influence what levels or units the panels focus on when assessing the research. As mentioned, the two field evaluations use the same units/levels, while both the institute evaluations and the programme evaluations differ. The research councils might still have some influence over such a basic feature of the panels' work. The two field evaluations were commissioned by the same research council and they had a common 'model of reference'. Both panels were given access to copies of previous evaluation reports and were in that way *informed on what kind of report the Research Council expected.* These two evaluations both evaluate departments/research groups, they mention individual researchers and have some degree of implicit comparisons. It is clear that *the standard set by previous evaluations* by the Research Council in these two cases more or less directly *influenced the approach adopted by the evaluation panel.*

The two institute evaluations and the two programme evaluations, on the other hand – differing in their choice of units of assessments – were commissioned by different research councils, meaning that the councils in these cases have given different signals about what they wanted, given no such signals, or that such signals had no influence on the work (see the discussion on the effects of the terms of reference in the following section). A very plausible alternative in all four cases is that as there was no tradition for the kind of evaluation in question, the commissioning body did not know precisely what it wanted and therefore gave no signal on what the units of assessment should be.

When it comes to effects of the kind of *research discipline* under review there is no unambiguous evidence that evaluations are more standardised in any of the areas. As mentioned, the natural scientists gave somewhat clearer statements on criteria and basis of evaluation than the evaluators from the humanities/social sciences, but this might just mean that the latter were more inclined to talk about the tacitness

and ambiguities of evaluating research, not necessarily that the natural scientists' evaluations were more standardised.

There are some obvious differences in the standards used assessing basic or applied research. Extra-scientific relevance is considered only when assessing research to which funding is given for applied purposes. In these cases the terms of reference also ask for such assessments. Yet, some of those reviewing applied research used the same 'standards' as for basic research, and one evaluator who represented the user-side thought that the standards of basic research were fundamental also for applied research. In Cases 2, 5 and 6 basic–applied was the main dimension of disagreement between panel members.

In conclusion, the commissioning bodies seem able to influence the aspects to be considered and assessed by the panel (i.e. *topics* such as extra-scientific relevance, but not necessarily the '*standards*' used for assessing it) as well as the unit of analysis of the evaluation report (e.g. research groups, research areas, programmes, institutions) provided that clear signals or instructions are given on these matters. It should be noted that influence on the unit of analysis was obtained when providing the evaluation panel with copies of previous evaluation reports that provided a 'model of reference'.

In the following section the role of the commissioning body is further investigated in terms of the effect of the terms of reference/mandate of the panel, the information/signals given on the purpose of the evaluation and its planned used, and the working conditions given the panel in terms of time and resources.

5.3.2 Terms of reference, planned use of the report, and the time and resources for the panel work

According to the evaluators the *terms of reference* had no or little effect on the work in the groups. However, those not 'following' the mandate were very concerned to justify their approach. In one case (no. 3) a central question of the mandate was obviously not addressed. In another case (no. 4) the evaluators deliberately exceeded the terms of reference, extending its tasks to an evaluation of teaching (in addition to evaluating research). Studying the reports we see that most questions of the mandate are directly addressed, and the evaluators interviewed seemed concerned that the Research Council should be satisfied with the result. I conclude then, that the mandates did steer

the work of the panels, but that the panel members in retrospect were only concerned about the mandate if there had been problems following it. On the other hand, the terms of reference generally gave no particular guidelines on the evaluation process, nor the standards or criteria for evaluation, except stating that the research should be evaluated in an international perspective or that extra-scientific relevance was to be assessed (see also the exception of Case 6, below). *The mandates' potentials for steering the focus of the assessments were therefore limited.*

The major difference in signals given the panels concerning the *planned use of the report* seems to be between programme evaluations and non-programme evaluations. In the case of the two programme evaluations it was clear to the panels that their evaluation was likely to influence future funding of the programme. Realising the possible consequences of their reports affected the two panels' approach and reports differently. In Case 6, the panel was particularly concerned with answering effect questions: whether the various projects would have been realised without the programme, which units were most able to use grants effectively, *et cetera.* In Case 5, the panel members conceived the mandate as formulated to make them answer general questions relevant to the future funding of the programme, and not requiring a study of effects of grants or comparisons of research units. They found the answers to the mandate obvious and answered them briefly, generally and positively. *Here differences in the terms of reference seem vital to the divergent approaches chosen by the panels.* In Case 6 the panel was asked to evaluate the programme's success as a policy tool, whereas the terms of reference in Case 5 contained no such question (see Sections 4.5.3 and 4.6.3).

In both field evaluations on the other hand, the panel members had divergent impressions of the commissioners' purpose with the evaluations. The general impression was that the fields under review should get help to improve, and/or that the Research Council had a general need for information about the activities in the research fields. There were divergent opinions on what effects a field evaluation report might or should have. Some panel members took it for granted that the report would have funding implications, others were much more cynical about the Research Council's ability or willingness to implement their recommendations. One of the latter emphasised that he tried to avoid the 'need for more money' part of the recommendations to make it more acceptable to the commissioning body.

Also the members of the panels of the institute evaluations had divergent views on the purpose of the evaluations. Most said they did not know what kind of evaluation report the commissioning body expected. There was a clear difference between the two panels in their perceptions of possible consequences. Most evaluators of Case 2 said the evaluees were initially concerned that the evaluation might have negative consequences for the institute (a fear at least one of the evaluators shared), while the panel of Case 3 seemed to conceive the commissioning body as wanting to help the institutes to improve – at least they saw the evaluation as a general and undramatic event.

Looking at the importance of *time and resources*[124] available for the work, few seemed concerned about the time limitations, but emphasised that they could, of course, have reviewed the material more thoroughly if they had had more time. In Case 6 for example, one panel member said that it was not realistic to do a systematic review of all the written material, and he only read thoroughly those items he thought especially interesting. In Case 5 the time limits as such had concrete consequences for the process and the report. Here the panel was expected to finish the report during the site visits in Norway, and had no time for discussions on the individual drafts or for co-ordination and editing of the drafts into a common report. The only interaction was on the general conclusions and recommendations. Lack of concluding discussions (due to a dinner invitation) combined with later absence of one central person in a controversy also influenced the assessments of one area in Case 1 (see Section 4.1.7). Individual time restrictions also had consequences in Case 3. Part of the reason for the lack of assessments of the research in this case was the absence of one of the panel members from several of the panel meetings. In general, it should be noted that limitations in time and resources may partly explain the vague assessments and absence of explicit negative criticism. The thoroughness of the review may be understood as too limited to allow well founded negative statements on the research under review.

Four of the six panels had a Norwegian co-ordinator.[125] In the cases with mainly non-Scandinavian panels (Cases 1, 4 and 5), panel members emphasised the co-ordinator as a valuable source for inside information and general information about the Norwegian system. All panels had one or more secretaries to their disposal. The panel members seemed satisfied with the secretarial resources they were

offered (except for the case where the secretary did not come from the Research Council.) The extent to which the secretaries were set to gather information varied. The secretary's role in editing the panel's report seems most important in the case where there were no formal chair or co-ordinator of the panel (Case 2).

5.3.3 The composition of the panels

The composition of the panels seems to have been of vital importance for the review process and for the outcome of review. The heterogeneity of the panels, for instance, inhibited the evaluators in commenting on each others' drafts, and in most cases there was no real potential for positive group effects (mutual insights gained through discussions, more information and a larger spectrum of ideas being considered by each panel member, or panel members acting as mutual motivators/supervisors, see Section 3.2 and 5.2.3). Only two cases (Cases 1 and 6, see below) had any group interaction on the research assessments as such, in neither case did interaction result in any changed opinions on assessments.

The discussion in Chapter 3 implies that the degree of *overlapping competence* between the panel members influences both co-operation and conflict on a peer panel. As professional platform is one of the main bases of peer review and peers have restricted fields of competence, overlapping competence is needed both for co-operation and (open) conflict between peers. Experts without any overlap in competence will have little to gain from co-operating on assessments, and poor possibilities of (authoritatively) questioning each other's assessments.

The data from the six cases support the suggested link between overlap in competence and co-operation/conflicts on assessments. The two cases with some overlap of competence, Cases 1 and 6, were the only cases with *interaction* on the research assessments, and the only cases where panel members clearly *disagreed* on assessments of the research.[126] In the other cases there were no apparent overlap of academic competence and therefore less possibility to disagree or co-operate on such judgements. It should also be noted that the *division* of areas under review (task division) between panel members seems to have been evident in all cases except Case 6, which turned out to be the only case with *open* controversy between panel members.

Another issue is *prior relations between evaluators and evaluees*. There seem to have been obvious differences in the scope of the evaluees'

international network of the various cases.[127] However, those with the broadest international network were those getting the broadest international evaluation panel. There is no case of a panel with no connection to any of the evaluees. In all cases some evaluees were known, others were unknown to the panel. The closeness of connections varied, and depended on the way the panel was selected. Some cases where the evaluees were given the possibility to propose panel members ended up appointing some evaluators with loyalty relations to evaluees (Cases 1 and 2 in particular). Two of those using only underhand or 'administrative' contacts for proposals for panel members ended up with evaluators having less close connections to the evaluees (Cases 3 and 4). In Case 6, it seems to have been especially difficult to find evaluators with 'due distance' to the evaluees. Here evaluees were given no influence on the selection, but the council still ended up with some evaluators with close prior knowledge of, and connections to, the evaluees. The reason may be the close Scandinavian relations in the area and that evaluators had to be able to read Norwegian. The 'administratively selected' non-Nordic panel of Case 5 also had close prior knowledge of the evaluees, but here the interviewees did not conceive prior connections to be a problem.

The selection of panel members may explain the *potential for controversy* on the panels. In Case 1, one of the sub-disciplines had a well-known controversy. The council accepted the parties' own proposals for panel members, and all parties were therefore represented on the panel, giving an obvious potential for controversy. We might say that such an outcome was predictable to the Research Council as they were informed on the divergent views and approaches of the field and let the evaluees themselves propose the panel members. However, the Council did not know the scholarly viewpoints of all the panel members they appointed in advance, and the disagreements on the panel were therefore only partly predictable to them.

In Case 2, researchers and users with very different views on the issues to be studied were put on the same panel. Here the selection of a mixed panel was a central feature of the design of the evaluation, and a broad representation of opinions is likely to have been the aim of the commissioning body. As the views of the various parties were

presented (mixed) in the evaluation report, the commissioning body in this respect might be said to have got what it wanted.

In both Case 1 and Case 2 disagreements were handled by tacit compromises. In Case 6, there were panel members with divergent evaluation standards and opposing opinions about the work under review, leading to confrontation and open controversy in the group. A conclusion was reached through straightforward bargaining. In this case it is hard to relate the potential for controversy directly to the selection of the panel members. The representation of divergent scholarly viewpoints on the panel does not seem planned from the part of the Research Council, neither was it an effect of evaluees' (formal) influence on appointments, as the evaluees in this case were not asked to propose their evaluators. There is no evidence that the commissioning body had any advance information on the divergent opinions of the panel members.

The same goes for Case 5 where there were partly divergent views on one central topic (basic versus applied research). The Research Council had full control over the selection of panel members but seemingly no information on the various scholarly viewpoints within the panel on this topic, neither prior to the evaluation, nor in retrospect – as the various viewpoints in this case were not explicitly stated. What Cases 5 and 6 have in common, and which was the conscious choice of the commissioners, was that the evaluation should cover all fields of a broad research programme, requiring a broad spectrum of evaluator expertise and heterogeneous panels. Thus, another way to obtain broad representation on an evaluation panel – rather than letting the evaluees propose members – is to review a broad heterogeneous area of research and ('administratively') appoint a relatively large evaluation panel.

This is contrasted by the two panels with least variation in opinions (Cases 3 and 4) which were the panels with fewest members. The research under review was not perceived by the commissioner to require a particularly broad or large evaluation panel (three experts on each panel, whereas Cases 5 and 6 had five experts each). As in Cases 5 and 6 the panels were put together without input from the evaluees. However, Cases 3 and 4 differ from Cases 5 and 6 when in comes to advance knowledge about the experts. Most of the selected experts were already known to the commissioning bodies, who do not seem to have had broad representation of opinions as a notable criteria for selecting the panel members. (This does not imply that a limited

representation of opinions was a conscious choice, only that a broader representation might have been the result if this was seen as important by the commissioning body – provided that experts with divergent opinions were known to the commissioning body.)

We conclude then, that the commissioning bodies to some degree 'designed' panels which were characterised either by consensus or by divergent opinions. Panels unanimous in their opinions (allowing consistent and clear conclusions[128]), were obtained by appointing a small group of experts already known to the body organising the evaluation, or its trusted advisers (Cases 3 and 4). Broad representation of opinions on the other hand (resulting in nuanced, but non-critical, reports with vague conclusions), was obtained by letting all parties involved getting their candidate onto the panel (Cases 1 and 2). A larger group of experts appointed administratively by a commissioning body without substantial background knowledge about the experts also resulted in broad representation of opinions (Cases 5 and 6). In generalising such conclusion, the diversity of the field/subject of evaluation must of course be considered. The broad representation of opinions in Cases 5 and 6 resulted foremost from a decision to cover a broad heterogeneous field, and not from the administrative appointments without background knowledge. Nevertheless, more background knowledge on potential experts might have given other results. Combined with a commissioner thinking that 'the best experts are those complying to standard traditional peer review criteria of basic research', this would certainly have given a more narrow representation of expert opinions on the panel in Case 6.

5.4 Summary

The analysis has sought insight and understanding of the constraints, processes and outcomes of a particular kind of evaluation: research council-commissioned expert evaluations of research units, fields and programmes. The central features found in the evaluations reports studied are vague assessments ('between the lines'), tacit criteria and conclusions and recommendations supporting the researchers under review. The processes on the evaluation panels were generally characterised by division of tasks, little interaction on assessments and

mostly tacit compromises in case of disagreements among panel members.

The time and resources available for the panels set important limits both for collaborative/overlapping assessments and for the thoroughness of review. We have seen that focus on the composition of the panels is important, especially for understanding the (lack of) interaction on assessments and also often for vague assessments in the evaluation reports. Overlapping competence is found to be vital both for co-operation and for open conflict between academic experts.

The commissioning bodies to some degree (more or less unconsciously) 'designed' for a consensus panel or for a panel with divergent opinions. The appointment of a small group of experts already known to the body organising the evaluation, or its trusted advisers, gave unanimous panels. Broad representation of opinions on the other hand, was obtained by letting all parties involved getting their candidate on the panel.

In general, the mandates' potentials for determining the focus of the assessments were limited. However, when clear (enough) mandates, 'models' or instructions were given, the commissioners seemed able to influence the aspects being considered and assessed by the panels, and also the unit of analysis of the evaluation reports. We have seen that the different approaches of the two panels evaluating research programmes can be explained by differing mandates. In two other cases the commissioners provided the panels with copies of previous evaluation reports, and the access to these previous evaluation reports clearly influenced the approach adopted by the panels. On the other hand, one panel did not address a central question of its terms of reference and another panel deliberately exceeded the terms of reference. In the two later cases the panel did not perceive the terms of reference adequate to the situation or task, and the commissioning body accepted the approach adopted by the panel.

An important factor for understanding the evaluations is the 'social norms' to which the experts comply. The combination of *peer* evaluators and *public* reports seems to foster some distinct rules for such evaluations. The evaluators were concerned about the potential use of their evaluation report, i.e., they considered the political context of the evaluation while writing their reports. Most peer evaluators saw their role as an aid to the research under review and this was one reason for the vague reports. When putting together the

reports the peers avoided details that would not serve the kind of policy they thought appropriate. Disagreements on the panels were an additional reason for vague reports. The resolving of disagreements invariably resulted in more vague reports. Scholars are likely to defend their own kind of research, which is also the kind of research in which they have the most expertise. On heterogeneous panels the other party to the controversy seldom know the area under controversy well enough to legitimately contest the 'defender's' assessments.

An overall conclusion is that the composition of the panel is of great importance when designing an expert evaluation of research units, fields or programmes. The composition set the potential for interaction, divergent opinions and conflict – and partly for the vagueness of the assessments – in addition to the fact that the selection of panel members decide what scholarly positions and opinions are allowed access to the evaluation report.

Table 5.2 *Case 1: Peer evaluation of research fields within the natural sciences*

Organisational constraints directly or indirectly set by the Research Council	**The evaluation panel's way of approaching their task**	**The contents of the report**
Scope of the evaluation: Four sub-fields Kind of research: Basic (and some applied) Natural sciences Mandate: Focused and detailed questions, but no criteria specified. Time and resources: Deadline exceeded. 13 months from appointment to report. Norwegian co-ordinator from nabouring field. Secretary from the Research Council. Selection of reviewers: Peer reviewers (recommended by the evaluees) covering major areas and paradigms. No Norwegian peers. Heterogeneous group, but some overlap of competence/research fields. The reviewers' constraints and dispositions: Most reviewers had friends/colleagues among the evaluees. Divergent research interests and competing 'schools'. No concrete signals given concerning the planned use of the report.	Methods: Review of publications. Site visits/interviews. Written information from the research groups (plans, resources, publ.lists). (Separate bibliometric study) Group discussion/ interaction: Division of research fields under review between the panel members mentioned in the letter appointing the panel. Common group discussions in the evenings. Separate 'sub-group' meetings. Final editing by mail. Heterogeneous group without open discussion. Disagreements in the report avoided by tacit compromises. Majority yielding. Group effects: 'Peer supervision'/one member rewrote his assessments.	The report's stylistic emphasis: Describing Assessing Unit of analysis/assessments: Sub-fields. The research groups. The institutions/ departments. Individuals mentioned. Implicit comparisons. Direct/indirect assessments Based on the evaluators' own views/assessments Central criteria: - intra-scientific relevance - quality (novelty, profundity) - input and output (quantity in bibliometric appendix) Conclusions: Positive, praising. - Good/valuable research that deserves further support. - Particular areas should be strengthened. - Lack of large-scale projects. - Need of travel funds for small groups. - Obtain a better balance between basic and applied research.

Table 5.3 *Case 2: Mixed panel evaluation of an engineering research institute*

Organisational constraints directly or indirectly set by the Research Council	**The evaluation panel's way of approaching their task**	**The contents of the report**
Scope of the evaluation: One institute Kind of research: Applied Science/engineering Mandate: Twofold (market role/relevance and research quality) No criteria specified. Time and resources: 14 months from appointment to report. One panel member not time to participate fully. Secretary from the Research Council. Selection of reviewers: Mixed panel with majority of non-peers. The evaluees had substantial influence on the selection. All panel members from Nordic countries. No formal chair or co-ordinator. More concern to cover the various user groups than the research areas. No apparent overlap of peer competence. The reviewers' constraints and dispositions: One peer reviewer had friends/colleagues among the evaluees. Heterogeneous group: clearly divergent dispositions for the task. No concrete signals given concerning the planned use of the report.	Methods: Site visit/information from the institute and department managers. Interviews with central external informants. Reviewing publications (done by one evaluator). No guidelines or common planning of methods for the evaluation. Group decisions/interaction: No clear division of tasks between panel members. Disagreements in the report avoided by tacit compromises. The different points of view edited into one 'coherent' report by the secretary. Group effects: One peer working harder. One peer shirking. Fruitful dialogue? Obtained understanding of each others' points of view, but no transformation of opinions.	The report's stylistic emphasis: Describing Explaining Assessing Unit of analysis/assessments: The institute and its different areas of competence. No individuals mentioned. Implicit comparisons of areas of competence. Direct/ indirect assessments: Based both on the evaluators' own views and assessments made by other informants/ colleagues. Central criteria: - output - applicability - relevance for specific user groups - quality (scholarly reputation, novelty/ contribution to theory and body of knowledge) Conclusions: Positive, praising. - Make the institute more attractive for contractors. - More basic long-term funding.

Table 5.4 *Case 3: Mixed panel evaluation of three social research institutes*

Organisational constraints directly or indirectly set by the Research Council	**The evaluation panel's way of approaching their task**	**The contents of the report**
Scope of the evaluation: Three institutes Kind of research: Applied Social sciences Mandate: Comprehensive (organisation, research quality, market relevance and user contact). Some specifications of criteria. Time and resources: 10 months from appointment to report. One panel member not time to participate fully. Secretary from a non-involved institute. Selection of reviewers: Peers from university departments and (one) non-peer. Evaluees no direct influence on choice. All panel members from Nordic countries. Norwegian chair. All research areas and paradigms not covered. No overlap of peer competence. The reviewers' dispositions and constraints: More interested in and qualified in 'evaluation research' than in quality assessments. Signals given concerning the planned use of the report: Input to Research Council's policy generally.	Methods: 'Evaluation research' Questionnaires to the evaluees. Statistics on input and output indicators. Site visits/interviews. Interviews with customers. Reference group. Reviewing publications. (Superficial quality review.) Group discussion/interaction: Common group discussions and some division of tasks. No apparent controversies on the panel. Group effects: 'Fruitful dialogue' (One panel member shirking.)	The report's stylistic emphasis: Describing Explaining Unit of analysis/assess-ments: The 'kind of institute'. No individuals mentioned. No comparisons/ranking. Direct/indirect assessments: Based on the evaluators' own views and assess-ments made by other informants. Central criteria: - input/framework conditions - extra-scientific relevance - applicability - (quality indicators: methods, reference lists/knowledge of previous research) Conclusions: Vague assessments. Problems and explanations emphasised. Two alternative solutions: - More basic funds/time for basic research and/or co-operation with other institutions.

Table 5.5 *Case 4: Peer evaluation of three humanities sub-fields*

Organisational constraints directly or indirectly set by the Research Council	**The evaluation panel's way of approaching their task**	**The contents of the report**
Scope of the evaluation: Three sub-fields Kind of research: Basic Humanities Mandate: Focused and detailed. No quality criteria specified. Time and resources: Seem sufficient. 19 months from appointment to report. Norwegian co-ordinator from nabouring field. Secretary from the Research Council. Selection of reviewers: Evaluees had no formal input on the selection. All peers from non-Nordic countries. One peer per sub-field/no overlap of peer competence. The reviewers' dispositions: Competence and interests exceeding the mandate. Oral/informal signals given concerning the planned use of the report (influencing panel's concern for policy relevance).	Methods: Reviewing publications. Questionnaires to the evaluees. Statistics on input and output indicators. Site visits/interviews. Group discussion/interaction: Division of tasks and common group discussions. Interaction on conclusions and presentation of assessments: one panel member moderated the explicitness of his assessments. Transforming opinions/tacit compromises. Group effects: One member working harder.	The report's stylistic emphasis: Describing Assessing Unit of analysis/assess-ments: Sub-fields. The research 'groups'. The institutions/depart-ments. Individuals mentioned. Degree of comparisons varies from one reviewer to the other. Direct/indirect assessments: Mainly based on the evaluators' own views/assessments. Central criteria: - input and output - quality - intra-scientific relevance Conclusions: Praising and criticising. - Recommending new positions, means encouraging productivity, interdisciplinarity, scholarships for graduates, improved teaching/curricula.

Table 5.6 *Case 5: Peer evaluation of a natural science research program/priority area*

Organisational constraints directly or indirectly set by the Research Council	**The evaluation panel's way of approaching their task**	**The contents of the report**
Scope of the evaluation: A large programme encompassing several fields and containing many sub-programmes. Kind of research: Applied, strategic and some basic. Natural sciences. Two disciplines involved. Mandate: Concrete, easy answered questions. No criteria specified. Time and resources: Final report prepared during the 10 days of site visits in Norway. (5 months from appointment to report) Norwegian co-ordinator from neighbouring field. Secretary. Selection of reviewers: Evaluees no formal influence on the selection. All peers from non-Nordic countries. One peer per selected sub-programme to be reviewed. No apparent overlap of peer competence. The reviewers' constraints and dispositions: Colleagues among the evaluees. Divergent competencies and research interests. Evaluation supposed to affect the continuation of the programme/future efforts in the field.	Methods: Reviewing publications. Site visits/interviews. Rigorous/superficial reviewing: varies between evaluators Group discussion/interaction: Firm division of tasks between panel members. Common group discussions about the general conclusions. Tacit compromises on the recommendations. Report edited by the panel member with most time/interest. Easy agreement due to firm task division Group effects: 'Jockeying for influence' Division of tasks resulting in no group interaction on writing assessments.	The report's stylistic emphasis: Describing Assessing Unit of analysis/assess-ments: (Emphases vary from reviewer to reviewer.) The field/sub-fields. The research groups/institutions. Projects. No individuals mentioned. Some implicit comparisons. Direct/indirect assessments: Most evaluators made direct assessments, one mainly used indirect indicators. Central criteria: (Vary between reviewers) Output, quality, reputation, intra- and extra-scientific relevance. Conclusions: Very positive/praising. - Pointing out areas for future efforts. - Underlining the need for basic research.

Table 5.7 *Case 6: Peer evaluation of a multidisciplinary research program/priority area*

Organisational constraints directly or indirectly set by the Research Council	**The evaluation panel's way of approaching their task**	**The contents of the report**
Scope of the evaluation: A large programme encompassing several fields and containing many sub-programmes. Kind of research: Applied, strategic, some basic. Multidisciplinary. Mandate: Comprehensive: Assess results and effects according to general quality requirements, the objectives of the program, applicability and productivity. The programme as a policy tool and recommendations on future priorities. Time and resources: 10 months from appointment to report. Pressed for time in the final phase. Secretary from the Research Council. Selection of reviewers: Evaluees no formal influence on selection. Nordic researchers/peers broadly covering the various fields. Norwegian chair. Some overlap of competence. The reviewers' constraints and dispositions: Divergent research interests. Divergent 'allies' among the evaluees. Evaluation supposed to affect the continuation of the programme/future efforts in the field.	Methods: Questionnaires to the evaluees. Site visits/interviews. Reviewing publications. (Publications were reviewed with varying rigor.) Group discussion/interaction: Division of fields between evaluators agreed on in first meeting. Evaluators wrote separate chapters but commented on each others' draft. Disagreements resolved by bargaining and logrolling. Open confrontation. Group effects: Worked harder (for ones point of view). No changed opinions.	The report's stylistic emphasis: Describing Explaining Assessing Unit of analysis/assess-ments: Program areas/projects. The research groups/ institutions. No individuals mentioned. Some comparisons and implicit rankings. Direct/indirect assessments: Based on the evaluators' own views. Central criteria: - output - quality (various specifications) - intra- or extra-scientific relevance (varies between reviewers) Conclusions: The least vague of the 6 cases. (Implicit rankings) Positive, praising, some criticism. - Certain 'young' areas should be further developed/get further support. - Give priority to the most qualified groups.

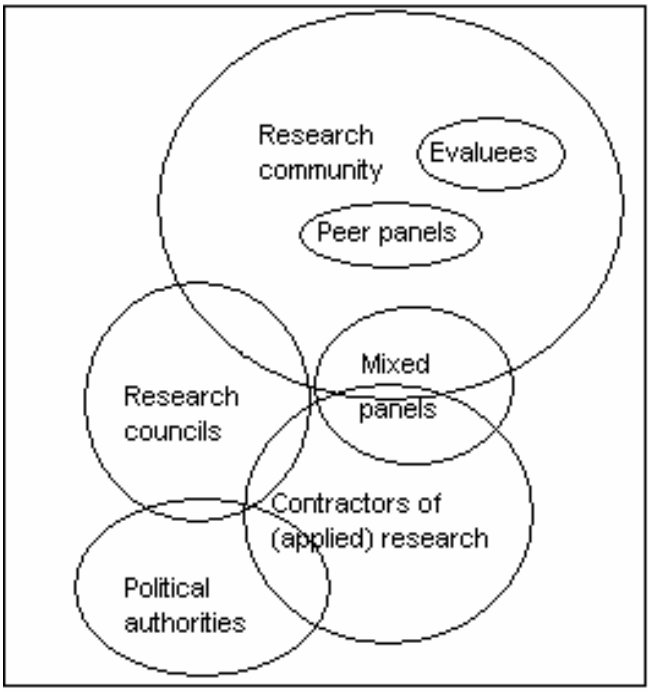
6 Implications to be drawn for the understanding of expert panel evaluation

What implications for the understanding of expert panel evaluation of research may be drawn from this study? This chapter focuses on the purposes and motives connected to research evaluation, the concept of bias, the nature of assessments and decisions and the way the evaluations *define* good research. In the last part of the chapter ideal types are used to illustrate central features and mechanisms of decision-making in expert panel evaluation.

6.1 The purposes of, and motives for, evaluating research

There are many parties to expert panel evaluations of research institutions, programmes and fields: the granting authorities demanding evaluations,[129] the research council allocating grants and commissioning evaluations, the expert panels executing the evaluations and, not to forget, the evaluees. Figure 6.1 illustrates the parties involved in expert panel evaluations of research and how they relate to the research community and contractors of research.

Figure 6.1 *Actors involved in research evaluation*



The parties often have various objectives for their participation, and they may also have divergent perceptions of the purpose for which the evaluation is being done. The official purpose may also be rather vague, not saying much more than that research institute A, research programme B or research field C and D 'are to be evaluated with regard to its framework conditions and scientific quality'. This is a kind of purpose each party may specify one way or the other, for example that the purpose is to see *how the research conditions may be improved*, that the purpose is to see *whether it is worth using money on this research* in the future, or simply a general need for *control of how public money is spent*, i.e. an evaluation that will not be acted upon unless serious flaws are found.

There may be various goals connected to the official purpose of the evaluation, e.g. implications or variants of the official purpose (topic-connected goals[130]). One possible goal connected to the kind of research evaluation dealt with in this study, is to *gather information as basis for policy decisions*. This seems an evident implication of evaluation. Why

would a research council ask for an evaluation if not to get information, and why would it ask for this kind of information if not as basis for policy-making?

Many objections may be raised to such a seemingly self-evident implication of evaluation. The (non-official) purpose of an evaluation may be predominantly symbolic, e.g. to promote a picture of the commissioning body as a dynamic and responsible organisation, or an authority supervising the research community. There might not be many plans, intentions or possibilities (in terms of reserved resources) to any follow up – neither to learn from the evaluation nor to act on it (Brofoss 1997).

Evaluation may also be a ritual activity, adopted from other organisations or sectors without much consideration of any purpose other than those attached to ceremonies. A well-known idea in organisation theory is the combination of the two, i.e. a ritual used for symbolic purposes, like a ceremonial outfit put on to be fashionable (Meyer & Rowan 1977). Research evaluation may very well be something governments and research councils think they must do, because if they do not evaluate, they will be perceived as out of date, still committed to a 'mediaeval' understanding of government.

Nevertheless, evaluation as a mere symbolic activity, and to some degree as a mere ritual activity, is 'parasitic' on the purpose and meaning of evaluation. If no one cares about the results of an evaluation, and everybody knows that nobody cares and that the evaluations will have no implications whatsoever, the evaluation has no symbolic meaning either, and it is even hard to see what kind of ritual meaning it may have.[131]

The data regarding the evaluations dealt with in this study, suggest that *gathering information for policy-making offers elements of the official purpose* that are acknowledged by all parties.[132] We will take this as a premise when discussing the 'goal structure' of the parties: the instrumental and ulterior goals of the various parties, and their possible non-topical goals (see note 130). Setting gathering information for policy-making as part of the official purpose does not exclude symbolic purposes or ritual functions from the analysis. As symbols and rituals do not make much sense as *official* purposes for the kind of evaluations dealt with here, the motives and purposes will not be discussed from such a point of view.

There may be several goals *ulterior* to the official purpose, i.e. goals for which the official purpose is instrumental. These ulterior goals may

vary between evaluations and also between parties to the evaluations, for example depending on their interpretation of the official purpose. Ulterior goals of evaluating research may be better research, better research conditions, better resource allocation, *et cetera*, all of which may be specified in various ways. Ulterior goals may also regard the aim of the research itself (e.g. increased knowledge of a phenomenon, economic growth, welfare or solving environmental problems).

Whereas ulterior goals are final purposes, *instrumental* goals are means to attain the official purpose. To appoint non-biased evaluators would be an example of a goal instrumental for evaluating research and gathering information for policy-making. Openly or obviously biased judgements will not do as a legitimate basis for policy-making (this is a premise for the acceptability of evaluations, see Chapter 2). Non-biased judgement was a norm the interviewed evaluators took for granted, but not always thought their fellow panel members had followed. Non-biased judgements and fairness/equity may be said to be topic-connected goals or goals that are an integral part of the official purpose.

As far as the official purpose implies non-biased judgements, any motive or purpose for evaluation that implies judgements or outcomes with a particular bias may be said to be *anti-topical*, i.e. logically irreconcilable with the official purpose. Opinions about what biased judgements mean, may vary considerably between parties. Political authorities may have a plain concept of bias as wrong conclusions and partial evaluators. Evaluators themselves may have a more nuanced concept of bias, e.g. judgements not consistent with prevailing peer opinions.[133]

There may also be goals of research evaluation that are neither topic-connected nor anti-topical. These may be called *a-topical* goals and include the kinds of goals that the panel members mentioned as motivating them to participate in the evaluations. Establish networks, learning about the research going on in Norway and travel to Norway and meet friends are examples of such a-topical goals. Goals that are (first) considered a-topical may prove to be *counter-topical*, i.e. to work to the disadvantage of the official purpose if furthered. Counter-topical goals of research evaluation include goals that *lead to* biased judgements. As biased judgement in itself is a rather meaningless and unlikely goal; goals leading to bias is a more interesting category. Goals that may bias outcomes include goals that demand a certain

conclusion of the evaluation, e.g. better conditions or more resources for certain research units.

What do the goal structures of the various parties look like? As mentioned, we may see gathering information as a basis for policy-making, as the official goal, and non-biased and fair assessments as topic-connected goals, acknowledged by all parties. Implied instrumental goals, shared by peer panels, mixed panels and commissioning research councils, are unbiased evaluators and a report suited for policy-making. Further instrumental goals may be found in the terms of reference (the mandate) of the panels. These ask for information on the scientific quality of the research, and sometimes also the usefulness and payoffs. The terms of reference may be understood in different ways by the parties and depend on how the ulterior goal of the commissioning body is understood. If the evaluators and evaluees suspect the ulterior goal to be related to a more or less symbolic or ritual 'need' for evaluations, the terms of reference will be interpreted very differently from a situation in which the ulterior goal of the research council is believed to be better research conditions or better allocations of grants. As ulterior goals are not openly stated, the evaluators have the possibility to approach their task as if the ulterior goal is what they think it should be – they may, for example, suspect it is to cut research budgets, but act as if it was to improve the conditions for doing good research.

The most interesting point where the goal structure of the various parties may vary is the possibly counter-topical goals. As we have seen, this particularly includes goals that may lead to biased outcomes. All parties may have such goals. An aim of panel members to influence research policy and help the evaluees, e.g. maximise research budgets may, as we have seen, influence the content of the evaluation report. Similarly, of course, the evaluees want to maximise research budgets. They also want to improve, or at least maintain, their reputation, and will be eager to provide the evaluation panel with all the good arguments. It is normally part of the game that evaluees shall have good opportunities to talk their cause, though not put blinkers on the evaluators. The commissioning research council may also want to maximise research budgets and may frame the task of the evaluation panel in a way which will contribute to this. Political authorities, on the other hand, may want to evaluate research to get a basis for cutting research budgets. As far as the authorities have no possibilities of influencing the panel's conclusions, this will not be a

counter-topical goal. However, a pronounced aim of cutting the budget, may develop into a counter-topical goal in the sense that it makes the evaluation panel bias the conclusions in favour of the evaluees in order to prevent budget cuts. On the other hand, if it is clearly stated and understood, a goal of cutting research budgets may also produce more thorough assessments and comparisons of units and analysis of the consequences of cuts in the relevant budgets.

What possible counter-topical goals did the evaluators have? We have seen that objectives of the peer evaluators that influenced the content of their reports included the goal of influencing research policy (more basic research and better research conditions, including better research budgets), and that of helping evaluees, at least to avoid negative effects of the evaluation. User-side evaluators, on the other hand, also often wanted to influence research policy in terms of better research conditions (which may imply increased budgets and better research or lower costs for the contractors). In addition to fulfilling a (perceived) demand for documentation of use of public funds, better research conditions also seemed to be a central objective for the research councils.

This does not mean that all categories of evaluators, as well as the commissioners of the evaluations, had non-topical goals that proved counter-topical, i.e. detrimental to the official purpose of the evaluations. Recommendations on how to better research conditions certainly contain information well suited as basis for decision-making, and may be seen as *part* of the official goal. However, in some cases actors were said to have opinions or stakes that may have led to biased conclusions in terms of *who* should get increased research budgets or who should *not* profit from the evaluation.[134]

On a general level, we may say that *the goals of the different parties to the evaluations seem to converge – improving research conditions was a central goal of all parties directly involved.* This means that the evaluators' search for a way of approaching their task which was compatible both with their obligations to the body commissioning the report and with 'decent loyalties' to the evaluees, need not be that hard.[135] Evaluees are concerned about their reputation and budgets, while the research councils are concerned about their legitimacy and documentation that public money is spent wisely, but they have common interest in arguments for better research conditions and increased budgets.

Evaluators might find it problematic to sympathise with the evaluees as well as producing the kind of report demanded by the research council (i.e. they perceive a 'multi-principal' problem), for instance because they suspect the research council to have reallocating purposes. The data indicate that this need not be a problem. The approach chosen by the panels studied is characterised by general praise, no (harsh) criticism and vague assessments on 'not-so-good' units, and arguments for the need of better research conditions. This approach seems to fit the interests of all directly involved parties.

The goal conflicts that did arise in the six cases, were internal to the evaluation panels and dealt primarily with the value attached to different kinds of research. Conflicts on the evaluations of basic contra applied research, and the evaluation of different research areas in

relation to each other, might be seen as based in panel members' divergent policy aims.

6.2 Peers and bias: a revised view

As demonstrated above, defining bias in expert panel evaluation reports is a tricky matter. We may easily identify actors' aims that might have led to biased assessments, but it is very difficult to say whether a particular assessment is biased or not. The mild assessments and the lack of ranking of the units under review that at first glance might be seen as a result of evaluators unduly biased in favour of the evaluees, may also be explained by contextual factors rather than individual bias of the evaluators. The number of comparable R&D-units in a small country like Norway is rather limited. Each department, institute or group has its particular research areas and approaches, and in many cases a recommendation saying that a particular unit should be given lower budgetary priority would be understood as if the kind of research done by the unit should not be done in Norway. As a peer panel will often include someone doing that kind of research, such a radical conclusion is rather unlikely. To oppose such conclusions may hardly be said to represent unacceptable bias; rather it is a predictable consequence of peers evaluating research units in a small country. Peer panels may of course criticise approaches *not* represented on the panel (and this they do, see Section 1.4.1 (footnote), the end of Section 5.2.4 and Section 7.4), but in a heterogeneous panel harsh assessments by one panel member are likely to be substantially moderated by other panel members before an agreement on the content of the evaluation report is reached.

When scholarly viewpoint is decisive, the *composition* of the panel (for which the commissioning body is responsible) is deemed to influence the outcome – either 'endanger' units under review by excluding representation from certain research areas, approaches or 'schools', or protecting units by including their proper peers on the panel. If panel members have no scholarly or personal loyalties to the evaluees, harsh criticism is much more likely. In the present study, the *scholarly* loyalties, as opposed to personal loyalties, definitely seem an important basis of potential bias. Scholarly viewpoint and interests, of course, include personal engagement, and personal and scholarly bias

(see Table 2.5) may be hard to separate. In fact, purely personal bias in peer assessments may be very difficult to identify if it does not counter one's scholarly bias. And even if assessments are inconsistent with the evaluators' scholarly viewpoint and scholarly interests, there might not be any kind of personal bias involved. If someone praises his 'scholarly enemies', it might be out of an ideal of generosity and pluralism – and not any personal interest or non-scholarly cognitive constraints.

In Chapter 2, it was suggested that *the process* might be a better indicator of bias than the outcome. A process of tacit negotiations and tacit compromises may mean a more narrow (perceived) representation of divergent opinions on the panel and less thorough review, than a process of open confrontation where a broader scope of views are explicitly discussed and seen in relation to each other.[136] Such an argument, of course, gives no *evidence* that the studied processes of tacit negotiation and tacit compromises resulted in 'biased' evaluations. It is nevertheless thought-provoking that the only case found with open confrontation on the content of the judgements and an explicit compromise between the conflicting judgements of the panel members, is also that case with the most explicit assessments in the evaluation report. The conflicting assessments of two panel members caused open confrontation, which gave a more thorough discussion and more nuanced conclusions, but the evaluation report still contains clearly sharper assessments than in the other cases studied.

Neither tacit compromises nor vague conclusions need, of course, be biased. Tacit compromises (as such) are 'biased' in the sense and to the degree to which they represent a false consensus (e.g. Case 1 where the majority yielded). There are various reasons for vague assessments, not all of which may be said to represent any bias. The assessments of an evaluation panel may be vague because:

(1) there are no clear differences in the research under review when judged by the relevant criteria
(2) the review has not been thorough enough, or the panel members do not have the competence needed to draw conclusions as to such differences
(3) the panel could not agree on differentiated conclusions, or
(4) policy considerations make the panel avoid such conclusions.

The first kind of reason is an unbiased one. Reason (2) represents bias due to sub-optimal thoroughness and information seeking or lack of

competence (category C bias). The vague assessments in Case 3 may be explained this way (provided that there *were* differences). Reason (3), lack of agreement on differentiated conclusions, may be a result of ' bias' due to scholarly viewpoint or scholarly interests of the members (Category A or B bias), and the composition of the panel. The lack of differentiated conclusions may here be a result of tacit co-ordination between the panel members (see Ideal Type I below, Figure 6.2), or it may be a result of panel discussions that *reduce* individual 'bias' (e.g. the moderation of assessments in Case 6). The last kind of reason (4), that of taking the potential effects on budgets or status into consideration, may be bias due to scholarly interests (category B bias). As mentioned, this seems a relevant explanation of vagueness in several cases.

Several informants dealt with the questions of bias and partiality, and gave diverse accounts on such questions. These vary between scientific (or 'empiricist') accounts and contingent accounts (see Section 2.2 and Gilbert & Mulkey 1984). Some stressed that they were impartial and objective, some meant that they (mostly *other* panel members) were partial and subjective. These conflicting accounts cannot be taken as *evidence* for partial or impartial assessments. The interview accounts still provide some central insight, as they refer to norms of impartiality and thoroughness, and problems with complying to such norms. On the one hand, there are panel members stating that the conclusions were really objective, that they always tell their frank opinion, had no ties to anybody and were free to have any possible opinion, or that they were able to be very independent minded. On the other hand, some state that other panel members had personal knowledge of evaluees and were less critical in their judgements than the rest of the panel (in one case the opposite: trying to punish enemies), had certain personal prejudices they tried to prove during the review, or had particular interests in certain fields. The accounts also include statements on problems doing a thorough review: that such assessments are a kind of art that necessarily is subjective, based on personal impressions, or on reputation. Others said that there was no time to go into depth, or that they were unable to give proper assessments.[137] This double repertoire for accounts on impartiality and thoroughness inform us of an ambiguous situation where interpretation of the rules differ: some evaluators think *other* panel members do

not comply to central rules, while finding that they themselves act in accordance with such rules (or at least want to present it that way). As suggested in Chapter 2, and further displayed in Chapter 5, there seems to be 'informal' rules for peer evaluation allowing pragmatism and prescribing some 'social sensitivity' in the assessments: rules that moderate the rules of impartiality and thoroughness – without being explicitly formulated.

Seen in relation to previous research (Gulbrandsen & Langfeldt 1997; Langfeldt 1998), the data substantiate that peer judgements on research are tacit and subtle, and based on diverse and non-easily operationalised criteria. This may be said to imply that guarantees against biased judgements are impossible. On the other hand, peer judgements on research may be understood, by definition, as being based on tacit criteria and craft knowledge, and divergent judgements as a fruitful starting ground for a process that defines and consolidates the notion of good research – trying to clarify principles, while still accepting divergent assessments. This interpretation may legitimate controversies and foster a more open debate on criteria, as well as on the distinction between various kinds of bias and their (il)legitimacy.

6.3 Tacit decision-making and tacit bases of assessments

Various degrees of tacit decision-making on the evaluation panels were found. In all cases compromises were tacit in the sense that differences of opinions were not included in the evaluation reports. In only one case was it fully clear to all parties that there had been a controversy, and what it was about, as there had been explicit bargaining to reach a solution. In some cases it is clear that dissension in the report was avoided by tacit compromises; in other cases decisions were tacit, but views were not clearly divergent and the outcome did not have the character of a compromise. In most cases however, panel members gave divergent accounts on the extent and content of disagreements, and the existence of divergent views was not clear to all parties.

The typical decision-making process of these panels can be characterised as 'sounding', with heavy emphasises on reaching consensus, on 'systematic' use of vagueness and on avoiding the

definition of clear alternatives (see Section 3.2).[138] Sounding reduces the potential for conflicts, avoids the definition of winners and losers, and makes it hard to reconstruct and document how 'consensus' was reached, as the actual decision rules and compromises are undefined and, as wee have seen, might be understood differently by the various panel members.

An extensive tacitness in *rules, criteria and standards* of judgements was also found. These were not discussed on the panels nor described in the evaluation reports, and neither were the informants able to describe them fully when asked about them. An important scientific norm is that processes are reliable, meaning that they always give the same result when they are properly carried out. If peer judgements on research necessarily depend on a personally incorporated body of knowledge and some kind of 'personal taste' (of the individual evaluator), peer review and peer evaluation is not, and cannot be, reliable in this sense.

This does not mean that there are no criteria for peer evaluation. The panel members had no problem in pointing out criteria for judging scientific quality, but they had problems explaining how they *use* these criteria when they evaluate research. Operationalisation of such criteria seems to be based on profoundly tacit knowledge. Furthermore, peer evaluators do not necessarily think in terms of for instance cumulativity, profundity or contribution to theory when making judgements about research quality. Nevertheless, when given a list with such terms, most of them could point out what they saw as important or not important, meaning that these kinds of criteria might be made explicit even if tacit at the moment of making judgements based on them.

6.4 Realism or idealism?

What is the role of the kind of expert panels studied in judging 'good research'? The evaluations include limited reading of research publications, and the emphasis may be more on abstracts and publications lists when reaching conclusions on, for example, the output of a unit and the scholarly relevance of its research. This kind of evaluation is substantially less thorough than traditional peer review and based at least partly on the opinions of other experts ('indirect

peer review').[139] This more superficial way of defining good research might be questioned from the point of view of realism. As argued in Section 2.2, thorough review may be expected to further valid results from this point of view (as well as from the point of view of idealism). According to 'optimistic realism', on the other hand, the judgements on research are obvious to competent evaluators, and an expert's opinion can be seen as 'true' – regardless of the process preceding it. Contrary to this, we have seen that competent evaluators do not always agree. Adding all the studies finding divergent assessments in peer review, optimistic realism is a very problematic perspective to defend (e.g. Travis & Collins 1991; Cole et al. 1981; Mahoney 1977).

In terms of idealism, the question is whether the evaluations *define* what is 'good research'. They define 'good research' to the degree that their assessments are embraced by others.[140] The kind of general praise distributed in these reports, probably contributes both to the reputation of the evaluees and their self-image. The assessments has also been referred to and used by the research units when applying for grants. On the other hand, there is at least one case where assessments were questioned by the Research Council commissioning the evaluation, implying that the evaluations do not *necessarily* define 'good research' authoritatively.[141]

Is realism or idealism the most relevant perspective, given the reports' role in research policy? The evaluations studied had no defined purpose that demanded the kind of thoroughness aiming, for instance, at ranking or grading the units under review. For policy makers without expertise in the field, an important indicator of whether judgements are correct is whether they are stated by experts whose opinions are widely respected by other experts. To make widely accepted judgements an expert's professional authority has to be unique and unquestioned, or he/she has to include (to some degree) the expected judgement of other competent evaluators ('common knowledge') and to pay attention to what are socially acceptable judgements (the 'informal rules' Table 2.3). In this respect, idealism seems appropriate for understanding the policy role of expert panel evaluation of research. A central official purpose of the evaluations was to provide general information that may be referred to when discussing policy measures. Putting 'common knowledge' into print may be part of this (evaluators in Case 5 saw this as a major purpose). Further, the authority of the reports may be more important than the 'novelty' of the information they contain.

The means for obtaining acceptance of the reports differ. In some cases commissioning bodies appointed experts in whom they had confidence and whose opinions they wanted to hear, without asking for the evaluees' opinions. In other cases the evaluees' suggestions for panel members were taken into consideration by the commissioning body, and a broad representation of opinions was ensured on the panel. The former approach contributes to an evaluation outcome accepted by the commissioning body, while the latter contributes to acceptance by the evaluees.

6.5 Decision-making on expert panel evaluation of research – illustrated by ideal types

This section focuses on the central features and mechanisms of decision-making in expert panel evaluation. The insight gained from the study is illustrated in the form of ideal types. These ideal types are analytic constructs. They are based on the combined insight from the empirical and theoretical parts of the study. They do not represent actual cases, but are constructed by the author to pinpoint what is found to be the central features and mechanisms of decision-making in expert panel evaluation.

Four types are presented. They are chosen to illustrate the conditions for what is found to be central features of the processes and outcomes of expert panel evaluations of research. Types I and III are cases of *strict task division and minority decisions* (Type I on a peer panel, Type III on a mixed panel). Types II and IV are cases of *group interaction and unanimous decision* (Type II on a peer panel, Type IV on a mixed panel). The outset conditions, the process and the results for each type are first described separately. All four types are summarised in Table 6.1.

After the four ideal types, two 'mixed types' are presented. These do not have the ideal types' 'extreme' outset conditions, and are more likely to describe actual evaluation processes.

6.5.1 Type I: Heterogeneous peer panel and general praise

Context and composition of the expert panel

In Type I the policy setting of the evaluation is *not* clearly defined for the members of the evaluation panel. *The 'official purpose'* which can be read out of the terms of reference (mandate) simply says that the scholarly quality and framework conditions of X are to be evaluated. No ulterior purpose for the task is openly stated by the commissioning body (e.g. that the evaluation is to serve policy making for better research in the areas, better research conditions or reallocations of resources). The task of the evaluation panel is consequently unclear with regard to the demands to provide a basis for specific policy decisions and the need to compare the units under review.

The assumptions as to panel members are:

1) The peer panel includes members from all areas under review, there is no overlap of research areas between the panel members and each member has clearly defined areas of competence. In none of the areas are there conflicts on research directions relevant to the evaluation.
2) Each panel member's central interests/stakes in the evaluation are based on his/her scholarly outlook.
3) Each panel member wants to give credit to those who pursue interesting research questions in a way which is promising from his/her scholarly viewpoint.
4) Each panel member wants to influence research policy by elucidating flaws in research conditions and providing arguments for the need of better research conditions in his/her own area.

Decision-making and result

The heterogeneity of the panel reduces group interaction on assessments. With clearly defined non-overlapping areas of competence panel members are not in a position to interfere with each others' assessments and conclusions. Other panel members cannot provide authoritative arguments against assessments made by the panel's expert in that area. This implies 'minority decisions' for each area following an implicit rule saying: 'If you are not an expert in the area, bow to the judgements of anyone defined as an expert in that area'. Each panel member is thus in a monopoly situation when assessing his/her area, while committed to respecting the other panel members'

assessments in their respective areas. The panel's one expert in an area is given the responsibility of making and writing all assessments and conclusions concerning his/her area.

When assessing their own area they each have the choice between balanced and nuanced conclusions, or to put weight on what is good and promising. Panel members are not supposed to represent any particular scholarly interests, but it is commonly known among the panel members that they all take a special interest in promoting research in their own areas. Each evaluator's minimum demand to the outcome is that the research in his/her area should not be given any 'worse conclusions' or less praise than other areas under review. When assessments on the various areas are written simultaneously without any co-operation between the evaluators, such an outcome can only be reached if all evaluators choose the same strategy: either balanced and nuanced conclusions *or* weight on what is considered to be good and promising. Each panel member has reasons to expect that the other panel members wish to present their own area in a good light, and in case some of them do not choose that strategy, the better reasons for the rest to do it.

The constellation of interests, in the simple case of two panel members, may be sketched as below:

Figure 6.2 *Ideal Type I: Constellation of interests*

		Member A: Balanced and nuanced	Member A: Positive
Member B: Balanced and nuanced		2 2	3 1
Member B: Positive		1 3	2 2

All parties have the same preference structure. They would prefer an outcome where their area is given more positive assessment than other areas (3). The worst outcome is the case where their area is given less positive assessment than other areas (1). In between come the cases where all areas are given equally positive or balanced assessments (2), between which the parties (need) have no preferences. (If they preferred equally balanced to equally positive (i.e. a Prisoner's Dilemma game),[142] or equally positive to equally balanced, the outcome would still be the same). Regardless of the other party's strategy each party then best furthers its interests by giving its own area a positive assessment.

The likely outcome of this game is that they all write very positive assessments of their own area. This is done by giving special credits to the area's most eminent researchers, groups, *et cetera*.[143] All evaluators also take care to elucidate and stress the need for better research conditions in their areas.

After reading each other's drafts the panel members agree to some adjustments in wording so that each area is given 'equal treatment'.[144] The result is an evaluation report that gives equal credit to the most eminent researchers in each area/unit under review, while not distinguishing between the areas, neither in praise nor in arguments for better research conditions.[145] This means general praise to all areas/units, and no comparisons between them, but some implicit ranking *within* units/areas, as the 'not-so-good' are not mentioned in the report.

As the policy setting of the panel's task is unclear, all panel members find such a diplomatic report a very good solution. There is nothing in the panel's mandate that demands more outspoken conclusions, the chosen solution is in agreement with norms of

collegiality and not harming anyone unnecessarily, and all panel members have been equally allowed to forward the interests of their areas.

6.5.2 Type II: Homogenous peer panel and clear ranking of units

Context and composition of the expert panel

In Type II, the policy setting of the evaluation is understood as providing a basis for reallocation in favour of the best units, and the panel members perceive this as an important task.

The assumptions as to panel members are:

1) All panel members take an interest in all research fields under review, and each member is considered qualified to assess research in all fields. The peer panel does *not* include members from all research directions under review, and there are no conflicting research directions represented on the panel.
2) Each panel member's central interests/stakes in the evaluation are based on his/her scholarly outlook.
3) Each panel member wants to give credit to those who pursue interesting research questions in a way which is promising from his/her scholarly viewpoint.
4) Each panel member wants to influence research policy by elucidating flaws in research conditions and providing arguments for the need of better research conditions in his/her own area.

(Points 2–4 are the same as in Type I.)

Decision-making and result

All panel members want to give credit to those doing the most valuable research, and provide a basis for reallocation in favour of the best units and arguments for the need of better research conditions in these units, and they therefore see the need for differentiating assessments. As the panel is homogenous, with no conflicting interests or incompatible judgements, interaction, open discussions and reaching unanimous conclusions on assessments are unproblematic.

The panel therefore easily reaches agreement on assessments and conclusions that point out which units are the most qualified and promising, and also distributes some honest and outspoken criticism

to those not pursuing research 'worthwhile' from the panel's point of view. Overlap of competence on the panel means shared responsibility for the criticism, and increases panel members' willingness to include outspoken criticism in the report. Not to openly break norms of collegiality and not to harm anyone unnecessarily, criticism is framed in a constructive way, including potentials, in addition to weaknesses and problems.

6.5.3 Type III: Mixed panel and divergent criteria

Context and composition of the expert panel

In Type III the policy setting of the evaluation task is *not* clearly defined for the panel, and its members do not want to use more time on the task than strictly necessary.

The assumptions as to panel members are:

1) The panel consists of both researchers and user representatives. There is a clear division of tasks between peer evaluators and user-side evaluators, as well as clearly defined areas of competence between the peer evaluators, and between the user-side evaluators. There is no overlap of research areas between the peer evaluators, nor overlap of branches between the user-side evaluators. However, there is no way to divide the object of evaluation clearly between the two groups of evaluators, as all (or a substantial amount) of the units under review should be evaluated from both the point of view of scholarly criteria and from the point of view of user criteria.
2) Each *peer* evaluator's central interests/stakes in the evaluation are based on his/her scholarly outlook. Each *user-side* evaluator's central interest in the evaluation is to ensure good conditions for research seen as relevant for their own branch.
3) Each peer evaluator wants to give credit to those who pursue interesting research questions in a way which is promising from his/her scholarly viewpoint.
4) Each peer evaluator wants to influence research policy by elucidating flaws in research conditions and providing arguments for the need of better research conditions in his/her own area.

(Points 3–4 are the same as in Types I and II.)

Decision-making and result

Peer evaluators on the one hand, and user-side evaluators on the other, have conflicting perspectives on the purpose of the research under review. This is understood by all panel members, but discussions are conducted as if there are no such conflicting views between the two groups of evaluators. Decision-making among the peers and among the user-side evaluators is as the decision-making of Type I: The heterogeneity within each group of evaluators (peers and user-side), as well as the limited time they want, or can, use on the task, reduce group interaction on assessments, and the panel's one expert in an area/branch is given the responsibility of making and writing all assessments and conclusions concerning his/her area or branch. This implies minority decisions for each area and for each branch following an implicit rule saying: 'If you are not an expert in the area/branch, bow to the judgements of anyone defined as an expert in that area/branch.' No expert wants the research in his/her area/branch to be given any less praise or priority than other areas/branches. Each panel member therefore writes very positive assessments of their area/branch and especially credits its most eminent researchers or the most branch-friendly researchers (as this is what they expect the other panel members to do, similar to the situation in Figure 6.2).

Such a procedure, with separate and uncoordinated area and branch assessments from peers and user-side evaluators is likely to result in divergent assessments on at least some of the research under review. The panel members see no need, nor want to spend time on discussing the divergent assessments or the divergent criteria underlying them. The result is a report where divergent assessments (due to separate peer and user-side assessments with divergent views on the purpose of research) are all presented in the report, but not seen in relation to each other or discussed in any way. As the common strategy of the panel members is to praise the kind of research he/she thinks valuable, while being careful with negative criticism,[146] the report contains praise to all research deemed as good from one point of view or the other (as all relevant points of view are represented on the panel).

With regard to the assessments of the research in various branches, and in various areas, the area assessments are more coordinated than the branch assessments. User-side evaluators do not see any argument or need for adjustment in formulations securing

'equal treatment' of areas or branches regardless of whom is the evaluator. The peer evaluators see such a need, but have no separate forum for discussing such adjustments.[147] After reading each other's drafts the peer evaluators individually do some adjustments in wording so that their area is not given any worse assessments than the area evaluated by the panel member most generous with superlatives. The result is an evaluation report that gives equally (clearly) positive assessments to all areas under review, and does not distinguish or compare between the areas, neither in praise nor in arguments for better research conditions. Formulations of the assessments of *branches* vary more, as user-side evaluators do not adjust their assessments.

As in Type I, the policy setting of the panel's task is unclear, and all panel members find the balanced and diplomatic report a good outcome. There is no defined object of the task that seems to demand more outspoken conclusions, the chosen solution is in agreement with norms of collegiality and not hurting anyone unnecessarily, and all panel members have been equally allowed to forward the interests of their area/branch. In addition, they have avoided time-consuming group discussions and negotiations on the conclusions of the report.

6.5.4 Type IV: Mixed panel and unanimous criteria

Context and composition of the expert panel

In Type IV the policy setting of the evaluation is understood as providing a basis for reallocation or better use of resources, and this task is perceived as important by the panel members.

The assumptions as to panel members are:

1) The panel consists of both researchers and user representatives. There is no clear division of tasks between the two categories of evaluators, nor within them.
2) No panel member represents relevant scholarly or branch interests.
3) The evaluators have no clearly diverging perspectives on research or judgements of the research under review, and no pre-set ideas about how to evaluate the kind of (applied) research to be reviewed, neither in terms of criteria nor perspectives based on scholarly viewpoint or branch interests.
4) Each evaluator (peers and users) wants the evaluation to give credit to those who pursue interesting and promising research

questions from the point of view of the 'general objective' of the research under review.

Decision-making and result

The decision-making is similar to Type II, with the difference that the evaluators start out without pre-set ideas about criteria and perspectives for their work. All panel members want to give credit to those doing the most valuable research, and provide bases for reallocations in favour of the best units, and they therefore see the need for differentiating assessments. As the panel is quite 'homogenous', with no conflicting interests or incompatible judgements, interaction, open discussions and reaching unanimous conclusions on assessments are unproblematic.

In the panel discussions, suggestions about what best supports the 'general objective' of the research under review, are put forward – what research directions, areas, branch orientations and quality criteria that are most important. All suggestions are easy to combine and sum up to that direction D research within area A, for branch B, scoring high on quality criteria C and C', is the kind of research that best supports the 'general objective' of the activity under review. This view is accepted by all panel members. With this common perspective the panel easily reaches agreement on assessments and conclusions that point out which units are the most qualified and promising, and also distributes some clear criticism to those not pursuing research worthwhile from this unified point of view. Criticism is framed in a constructive way, advising on how to better fulfil the criteria.

Table 6.1 *Overview of ideal type expert panel evaluations*

	Context	**Decision-making**	**Result**
Type I	Heterogeneous peer panel. Policy setting not clear. All relevant research areas represented on the panel. No overlap of competence. No conflicting research directions.	Strict task division. 'Scholarly bias' for all areas under review. Minority decisions on each area.	General praise. No comparisons.
Type II	Homogenous peer panel. Important policy setting. Some relevant research directions not included. Overlapping competencies.	Group interaction on assessments. 'Scholarly bias' for some research directions. Unanimous decisions.	Praise and criticism. Comparisons (rankings) of units.
Type III	Mixed panel. Policy setting not clear. All relevant areas and branches represented on the panel. No overlap of competence.	Strict task division. 'Scholarly' or 'user' bias for all kinds of research under review. Minority decisions on each area and branch.	General praise. No comparisons. Divergent views presented, but not related to each other.
Type IV	Mixed panel. Important policy setting. No one represents relevant scholarly or branch interests. No clearly diverging perspectives or judgements.	No predefined bias or criteria. Group interaction resulting in agreements on common criteria for all the research under review. Unanimous decisions.	Praise to those fulfilling the criteria agreed on, advise to others on how to fulfil them.

6.5.5 In between the ideal types

The ideal types are extreme cases. They pinpoint the logic and mechanisms of research evaluation, but in their pure form they are unlikely to (perfectly) describe actual evaluation processes. A total lack of overlap of competence between panel members, as in Types I and III, as well as the situation with no divergent opinions in Types II and IV, is unrealistic. A more normal situation is likely to include both some overlap of competence and some disagreements, and consequently a process of discussions, negotiations and compromises before agreeing on the content of the evaluation report – not the ideal type's simple processes of either unanimous decisions or minority decision.

The ideal types are far from 'ideal' in the normal sense of the term. The reports of Types I and III give information about the research under review, an overview that might be useful to outsiders, but give no new information to those familiar with the area/branch. On the other hand, they are fair in the sense that there is no bias towards particular areas or branches. Type II is unfair in the sense that some directions and areas are evaluated on their own premises, while others are not. Whereas Type IV, with no pre-set opinions and criteria, makes any outcome possible, i.e. the *process* is decisive, and for instance 'groupthink' and other related group effects may result in an outcome based on one-sided/narrow criteria.

However, the ideal types offer ample possibilities for 'in-between-types', some of which are much closer to the kind of evaluation those commissioning the evaluations would hope for (see Section 3.3). One possible mixed type between Type I and Type II is a heterogeneous peer panel with nuanced assessments, and between III and IV we may find a mixed panel with nuanced assessments.

Heterogeneous peer panels with nuanced assessments

Context and composition of the expert panel: The peer panel set to do the evaluation represents a broad spectrum of scholarly viewpoints, at the same time as the evaluators have some overlapping areas of competence. A central policy interest of the panel is to provide a report that might be used for reallocation in favour of the best units.

As in Types I and II, the panel members' central interests/stakes in the evaluation are based on their own scholarly outlook. They want to give credit to those who pursue interesting, valuable and promising

research from their scholarly viewpoint. In addition, they want to influence research policy by elucidating flaws in research conditions and providing arguments for the need of better research conditions in their own area (point 2–4, Types I and II).

Decision-making and result: The combination of overlapping expertise and divergent scholarly views facilitates and increases the need for group interaction. Furthermore, the panel members see a need for differentiating judgements, as they want to write a report that can be used as bases for reallocation in favour of the best units. To reduce the risk of an outcome not taking proper care of their research area, the panel members choose a strategy of open confrontation of conflicting judgements. When this includes in-depth discussions and open negotiations on the criteria of assessments and the content of the report (for instance, due to deliberate 'chairing' of the decision-making), the panel may reach compromises (and possibly moderated opinions) that integrate the divergent scholarly viewpoints and interests. If successful, the result is an evaluation report with differentiating judgements pointing out which units are the most qualified, and containing nuanced assessments stressing both strengths, potentials, weaknesses and problems of units and areas – an approximately unified view of a heterogeneous group of experts.

There is also the possibility that the negotiations define some of the viewpoints represented on the panel as 'winners' and others as 'losers'. This may be because some are better negotiators, or more willing to enter into hard negotiations, or because some scholarly viewpoint is held by a majority of the panel. Another possibility is that when confronted with the difficulties of reaching agreement on differentiating assessments the panel members give up the aim of providing a report that can be used for future priorities and allocations in favour of the best units, and end up with the same kind of minority decisions as in Type I.

Mixed panels with nuanced assessments

Context and composition of the expert panel: The policy setting of the evaluation task is understood as providing bases for reallocation or better use of resources, and the panel see this as an important task. The panel consists of both researchers and user representatives, and represents a broad spectrum of perspectives on the research under review. There is

some overlap in competence and no clear division of tasks between the panel members.

Each panel member (peers and users) want to give credit to those who pursue interesting and promising research questions from the point of what they perceive as the objective of the research under review.

Decision-making and result: To make a report that may provide a basis for better use of resources/reallocation, the panel see the need for reaching agreement on differentiating assessments. Overlap in competence, both within and between the two categories of evaluators (peers and users), facilitates group interaction on assessments. Divergent views on the purpose of the research under review as well as divergent assessments are discussed openly, and the panel members (jointly) try to formulate conclusions and recommendations that integrate the divergent views, while at the same time provide differentiating assessments. Through in-depth discussions and partly by open negotiations, they may reach agreement on an integrated perspective on the various goals and criteria for assessing the research, as well as agreement on nuanced assessments of the units. If not, they may end up with a negotiation result giving priority to a few branches and scholarly viewpoints, or minority decisions as in Type III (see the previous 'mixed type').

6.6 Summary

The chosen approach of the studied evaluation reports is general praise, no harsh criticism and vague assessments on 'not-so-good' units, and arguments for the need of better research conditions. This approach seemed to satisfy all directly involved parties. This means that there need not be a problem for evaluation panels to sympathise with the evaluees as well as producing the kind of report demanded by the research council.

Scholarly 'loyalties' (much more than personal loyalties) seem an important basis of potential bias in the evaluation reports. Scholarly viewpoint is decisive for assessments, which implies that the composition of the panel influences the outcome.

The panel members gave diverse accounts on questions regarding bias and partiality. The data display an ambiguous situation where evaluators sometimes state that *other* panel members did not comply with central rules, while they themselves acted in accordance with such rules.

Mild assessments and lack of ranking of the units under review need not be an indicator of evaluators unduly biased in favour of the evaluees. There are a number of possible reasons for vague assessments, of which some do not include scholarly or personal bias. Further, in our case vague assessments may be explained by contextual factors of evaluating research units in a small country. The number of comparable R&D units in Norway is rather limited. In many cases, ranking units and recommending that 'bad' units should be given lower budgetary priority would imply a recommendation that the country should not have this kind of research. Academic interests play a role when such considerations underlie mild and vague assessments. Such interests may be said to be expected to be part of a peer panel's considerations, and not defined as undue bias. There seem to be 'informal' rules for peer evaluation allowing pragmatism and prescribing some 'social sensitivity' in the assessments: rules that moderate the rules of impartiality and thoroughness.

The data substantiate that peer judgements on research are tacit and subtle, and based on diverse and non-easily operationalised criteria. Rules, criteria and standards of judgements were seldom discussed on the panels or described in the evaluation reports, and the informants were not able to describe them fully when asked about them. An important scientific norm is that processes are reliable, meaning that they always give the same result when they are properly followed. As peer judgements depend on a personally incorporated body of knowledge and some kind of 'personal taste' of the individual evaluator, it cannot be reliable in this sense.

Idealism seems appropriate for understanding the *policy role* of expert panel evaluation of research. A central official purpose of the evaluations was to provide general information that may be referred to when discussing policy measures. Putting 'common knowledge' into print is sometimes part of this. The studied evaluations had no defined purpose that demanded the kind of thoroughness aiming, for instance, at ranking or grading the units under review. For policy makers without expertise in the field, an important indicator of whether judgements are correct is whether they are stated by experts whose

opinions are widely respected by other experts. To make widely accepted judgements, an expert's professional authority has to be unique and unquestioned, or he/she has (to some degree) to include the expected judgement of other competent evaluators ('common knowledge') and pay attention to what are socially acceptable judgements.

The typical decision-making process of the panels is found to be 'sounding', with heavy emphasises on reaching consensus, 'systematic' use of vagueness and avoiding the definition of clear alternatives. Sounding reduces the potential for conflicts, avoids the definition of winners and losers, and makes it hard to reconstruct and document how 'consensus' was reached.

Ideal types (analytic constructs) have been used to illustrate what the study finds to be central features of the decision-making in expert panel evaluation. A heterogeneous peer panel with representatives from all involved areas and no overlap of competence, operating in an unclear policy setting, gives each panel member a monopoly situation over his/her area and results in a report that gives equal credit to the most eminent researchers in each area under review, while not distinguishing between the areas, neither in praise nor in arguments for better research conditions (general praise and no comparisons, Type I). A homogenous peer panel, not including representatives from some of the relevant research directions and operating in an important policy setting, on the other hand, gives group interactions on assessments, unanimous decisions and differentiating assessments pointing out both the best and the worst research (outspoken criticism and ranking of units, Type II).

Another kind of expert panel evaluation is 'mixed panel reviews' including both peers and branch/user representatives. A mixed panel in the same kind of setting as Type I described above – representatives from all involved areas and branches, with no overlap of competence, operating in an unclear policy setting – will similar to Type I yield a strict division of tasks and minority decisions on each area/branch under review. In the case of a mixed panel, the research is evaluated from both an academic and a user/branch point of view. When each expert has a monopoly on assessments on his/her area or branch, the setting implies that divergent views, and possibly divergent assessment on the research are presented in the report (Type III).

The last presented ideal type (Type IV) is also a mixed panel evaluation. As in Type II the policy setting is perceived as important by the panel members. The expertise of the panel members is general and there is no clear division of tasks between the two categories of evaluators, nor within the categories, and all evaluators start out without pre-set ideas about criteria and perspectives for their work. There are no conflicting judgements nor conflicting interests on the panel and agreements on differentiating assessments are easily reached. The outcome heavily depends on the process, and assessments may be narrow-minded.

These ideal types are extreme cases and unlikely to describe actual evaluation processes. A total lack of overlap of competence (Types I and III), as well as no divergent opinions (Types II and IV), is unrealistic. Moreover, the ideal types are far from 'ideal' in the normal sense of the term. They either provide no new information, are unfair or very process-dependent.

Possible 'good' evaluation processes placed in between the ideal types' extreme outset conditions have also been sketched. Moreover, it has been suggested that open controversies and open debate on criteria should be seen as criteria of good evaluation processes, and divergent judgements as a fruitful starting ground for a process that defines and consolidates the notion of good research.

7 In conclusion

The central findings and conclusions of the study are summarised in Section 7.1. Section 7.2 offers a retrospect discussion of research design and analytical tools. Policy implications are discussed in Section 7.3. Section 7.4 deals with unanswered questions.

7.1 Central findings and conclusions

Expert panel evaluation of research institutions, programmes and fields is a central method for funding authorities to get information wanted for formulating research policy. This kind of evaluation is particularly interesting as a decision-making process on the borderline between politics and science.

Problems with identifying good research are studied, including the concept of bias and partiality in peer judgements. Furthermore, the focus is on the organisational and group constraints on evaluation processes. The empirical part of the project consists of in-depth studies of the work of six expert panels, concentrating on the bases of judgements and the decision-making. The role of the commissioning research council and the influence of the organisational setting on the panels' decision-making also receive special attention in the analysis.

The answers to the central questions posed at the outset of the study (Section 1.3) are summarised below (the questions are used as headings). Empirical findings with focus on the major weaknesses and sources of bias are summarised at the end of the section.

7.1.1 Are there neutral criteria of good research?

The question is critical. If there is nothing that constitutes good research as such, everything is equally valid and it does not make much sense to speak of neutral criteria. It is obvious that everything is not equally valid in research – research is assessed and some contributions are praised, others refused. We may therefore say that there *is* a notion of good research. The question is *how* the notion of good research is constituted – what status does the notion have?

This question has no conclusive answer. What constitutes good research, and whether there are *identifiable* neutral criteria for good research depends on whether we rely on realism or idealism.[148] *Realism* denotes the view that reality exists independently of being experienced or conceived. The realist view of research quality means that there are standards which are constitutive of good research, unrelated to what evaluators might define as good research. In this view, good research might be something quite different from what the research community defines as good research. However, according to realism, there are *no clear indicators* of correct evaluations (unless we adopt the naïve assumption that peer review consistently reveals the 'truth' about research quality). *Idealism*, on the other hand, means that experience and thought constitute reality. An idealistic concept of research evaluation implies that the meaning of 'good research' is constituted through the evaluation process. According to idealism, *acceptability* by the actors involved may serve as a good indicator of properly based evaluations.

7.1.2 What does 'unbiased' evaluation imply? Is it attainable, and is it definitely desirable?

Four categories of sources of bias in research assessments have been identified (analytically). According to *realism* all four categories must be said to be bias: research interests, personal interests, the constraints of a professional platform and general/personal cognitive constraints. Personal/general cognitive constraints and personal interests are likely to be defined as bias also according to *idealism*.[149]

Peer judgements on research are found to be tacit and subtle, and based on diverse and non-easily operationalised criteria. This implies that guarantees for unbiased judgements are impossible. The attainability and desirability of unbiased assessments depends on whether we take the point of view of realism or idealism. Realism says that there is something that *is* constitutive of 'good research'; idealism says that 'good research' is constituted culturally and socially. From the point of view of realism all kinds of bias are undesirable. From the point of view of idealism some sort of professional bias may be desirable (based in professional preconceptions of good and valuable research).

Idealism, with the view that peer review is part of a continuous process defining quality, implies that low inter-reviewer consensus on

an evaluation panel is no indication of low validity of the assessments. In fact, lack of consensus may indicate that the panel as a whole is highly competent to make valid assessments because the panel represents a large scope of the various views on what is good and valuable research. In this view, the evaluation *process* may be a better indicator of peer review validity than the outcome of the evaluation. What is desirable here is broad representation of divergent judgements within the evaluation panel, and open debate on criteria and assessments. There is no apparent reason for this not to be attainable (see the discussion of ideal types Chapter 6).

Attainability of unbiased judgements appears different from the point of view of realism. Realism says that there is *something* that is constitutive of good research, but it does not say what it is. The evaluator's task is to know the answer. In a realistic perspective, procedural rules are in themselves irrelevant for whether an evaluation is 'right' or 'wrong'. Rules specifying standards for 'right' assessments and how to reach the 'right' conclusions may be essential as means to a correct evaluation, but there are no clear indicators of correct outcomes (see the preceding question). The indicators pointed out by idealism (peer acceptability and broad representation of views) are irrelevant from the point of view of realism.

7.1.3 How is good research identified, and what 'professional' and social norms affect judgements?

Peer review depends on tacit knowledge and craft skills internalised through socialisation processes – rendering some unity in the basis of peer review. Informal hierarchies, 'gatekeepers', dependence on judgements made by others, and the overlap and dependencies between research areas also contribute to unity. Moreover, there seems to be a common 'language' for peer review – a certain set of criteria that reviewers (more or less explicitly) pay attention to. In this way there are limits as to what may be seen as acceptable evaluations. We may add that to keep within such limits an evaluation at least has to appear to be rigorous and based on impersonal criteria of scientific merit.

We cannot specify rules, standards or (exhaustive) criteria of peer review, however. Studies of peer review find low inter-reviewer agreement, indicating that evaluators either use different criteria,

emphasise the various criteria differently or interpret the criteria differently – all leading to divergent assessments. Reliance on tacit knowledge, craft skills, and the lack of explicit criteria, emphasises that there will normally be a large grey area of acceptable evaluations, i.e. evaluations not clearly definable as lax or partial.

The data display an extensive tacitness in *criteria and standards* of judgements. Criteria were not substantially discussed on the panels or described in the evaluation reports, nor were the informants able to describe them fully when asked about them. The panel members had no problem pointing out criteria for judging scientific quality. They had problems explaining how they *use* these criteria when they evaluate research. Operationalisation of such criteria seems to be based on profoundly tacit knowledge.

Various rules and norms affecting judgements have been identified. On the one hand there are social norms prescribing impartiality and thoroughness, and on the other there are (informal) rules allowing pragmatism and prescribing some 'social sensitivity' in the assessments. The interviewees' accounts reflect this ambiguity. The objectivity of the assessments was emphasised. At the same time it is clear that promoting the research in the field under review was understood as an obvious and legitimate mission of government commissioned research evaluations. Most of the evaluators were careful not to write anything that might harm the resource situation of the evaluees. In addition, the context called for a pragmatic approach to thoroughness. Limited time and a large scope of evaluation material set limits as to what could be done and reduced the evaluators' ambitions to do a rigorous review of the material.

7.1.4 How may the group setting of expert panels evaluating research affect the outcome?

Four possible group effects have been outlined:

(1) The interaction has qualities that enhance the review work, for example that more ideas/information are considered by each member, or that the group members gain new insights through dialogue.
(2) The group members try to impress each other and therefore work harder (or appear tougher) than when working alone.
(3) Shared responsibility results in collective shirking.

(4) The group situation leads to uniformity/groupthink, including impairment of critical thinking, less rigorous review and suppression of minority opinions/false consensus.

The data point towards various group effects, but there is not much clear evidence for any of the effects sketched above (see Section 5.2.3 and Tables 5.2–5.7). In general, the processes on the evaluation panels were characterised by division of tasks, little interaction on assessments and mostly tacit compromises in case of disagreements among panel members. Consequently, the opportunities for *dialogues leading to new insights* were limited.[150] It should be noted that this conclusion concerns the effects of interaction in the panels on the *assessments* of the research under review. There seems to have been much more interaction and potential for fruitful dialogue on how to carry out the evaluation and on the policy conclusions than on the assessments of the research.

The second kind of group effects, *more thorough review to make a good impression*, seems to have appeared in one case – the case of a 'junior' panel member with no prior experience with this kind of evaluation. In addition, there were cases where the group context made panel members work harder to get his/her point of view into the report. This is another kind of group effect. More weight was put on good argumentation, but not necessarily on more thorough review work.

The material contains two cases of individual *shirking*, but no evidence for collective shirking. With regard to *groupthink*, the kind of setting and processes found – division of tasks and little interaction on assessments – tell us that groupthink was not much likely.

In sum, the kinds of group effects sketched in Chapter 3 seem to have been of minor importance for the expert panels, mainly because of the high degree of task division on the panels. To study the processes *defining* task division and the ways of tacit co-ordination between members of a task divided panel are more important in this setting.

Task division on the panels seems to rely on mostly *tacit* compromises on who ought to bow to whose judgements.[151] In addition to the authority and expertise areas of the panel members, their time and interests here seem important for how the borderlines are drawn between panel members' areas. Not all the research under review clearly belongs to the area of (only) one of the panel members. The

panel members take special interest in different evaluation units/fields and the time they are willing to devote to the evaluation work, also vary. In this way borderlines are more 'personally' adjusted than they would be without a panel, that is compared to single evaluators writing separate evaluation reports without any interaction.

Tacit co-ordination between panel members also adjusts the outcome of reviews. Ideal Type I (Section 6.5.1) illustrates how this can be done between members of a strictly task-divided evaluation panel. With the help of game theory it is concluded that each panel member will produce a review of their respective areas that *put weight on what is good and promising* (Figure 6.2). A central condition for this outcome is that it is commonly known among the panel members that they all take special interest in promoting research in their own areas. Each panel member expects the other panel members to write a review that places their own area in a good light, and in case some of them do not choose that strategy, the better the reason for the rest to do it. This kind of adjustment is even more likely when evaluators are not part of a panel, but work individually. In such a situation each evaluator may likewise suppose that the other evaluators will write a positive review and consequently write a positive review himself.

The differences between the two situations (group or individual evaluators) are the information on the other evaluators (who evaluates what, and what are their interests) and also the possibility of adjusting one's assessments after reading the assessments of the other evaluators. If one or more panel members initially choose to write a more nuanced/balanced or critical report than the others, this possibility of adjustments is important. It should be noted here that the question of positive or nuanced assessments is a question of *degrees* of emphasises on positive aspects, and not a choice between two different, clearly defined approaches. Some adjustments will therefore normally have to be done if panel members are concerned about how the assessments of the various areas relate to each other. What will be the result of such adjustments? The data here point to adjustments towards more positive assessments as the likely outcome. The panel members are not in a position to demand that other panel members write more critically about their area, but may of course say that after having read the other assessments they need to reword their own assessments to make them comparable.

7.1.5 How may the research council influence the outcome?

In the meaning of not foreseeing the influence of one's choices, the commissioner of a research evaluation may be 'neutral', but not in the meaning of not influencing the outcome. Central factors are set when commissioning and organising a research evaluation: scope and subject of evaluation, time and resources, mandate, signals about planned use of the report, and selection of evaluators.

The *number and size* of institutes, fields or programmes selected for evaluation might influence both the evaluation process and the content of the evaluation report. In the present material there is not enough variation in the size of the evaluation task to see effects, for example, on what unit/level of assessment the panel adopts (e.g. research groups, institutions or research areas). On the other hand, it is found that whether institutes, programmes or fields are evaluated does not necessarily influence what levels or units the panels focus on when assessing the research. Both the programme evaluations and the institute evaluations differ with regard to focused units/levels.

The *time and resources* available for the evaluations set important limits for the thoroughness of review. In general, it seems that the panels ability (and/or willingness) to devote time to thorough review, collaborative assessments or discussions of the assessments was marginal – with regard to the demands that should ideally be met in these kinds of evaluations.[152] The effects of time limits were most obvious in the case where the report was to be finished during the site visits and there was no time for discussion on the individual reviews.

The *terms of reference* generally gave no particular guidelines on the evaluation approach, nor the standards or criteria for evaluation, except stating that the research should be evaluated in an international perspective or that extra-scientific relevance was to be assessed. The mandates' potentials for determining the focus of the assessments were therefore limited. Nevertheless, for the two programme evaluations the terms of reference seem vital to the approach chosen by the panel. Different mandates can explain why one of these panels focused on programme effects whereas the other did not. In the two field evaluations the commissioning bodies provided the panels with copies of previous evaluation reports and in this way indicated what kind of report they expected. The access to these previous evaluation reports clearly influenced the approach adopted by the panels. In the institute and programme evaluations it seems that there was no

tradition for the kind of evaluation in question; the commissioning body did not know precisely what was wanted and therefore gave no signal on, for instance, what should be the units of assessments. In conclusion: The commissioning body influenced the aspects being considered and assessed by the panel, and the unit of analysis of the evaluation report, in the cases where clear signals/instructions were given on these matters.

The *composition of the panel* is found to have great importance. The composition set the potential for interaction, divergent opinions and conflict in addition to the fact that the selection of panel members determines what scholarly positions and opinions are allowed access to the evaluation report. The selection of panel members may explain the *potential for controversies* on the panels. To some degree the commissioning bodies designed for panels characterised by consensus or by divergent opinions. Unanimous panels were obtained by appointing a small group of experts already known to the body organising the evaluation or its trusted advisers. A broad representation of opinions, on the other hand, was obtained by letting all parties involved getting their candidate on the panel. With the last kind of appointment process, one risks appointing evaluators with loyalty relations to the evaluees. This happened in some of the cases.

7.1.6 Summary of empirical findings with focus on the major weaknesses and sources of bias

When we focus on the constitutive aspects of research quality we risk giving a one-sided account – to leave out the contextual and decision-making aspects and the possible informal rules of assessments. On the other hand, the understanding of constitutive aspects is important for the interpretation of such contingent aspects.[153] The present monograph has used a discussion of the constitutive aspects of good research as the point of departure to study decision-making in research evaluation and possible bias due to factors at the organisational level, the panel level, and the level of the individual evaluator. 'Bias' was found at all levels:

- *Bias on the organisational level* was basically related to the selection of panel members – i.e. which scholarly points of view were invited to participate in the process – but such factors as time limits were also found to have influenced the outcome.

- *Bias on the panel level:* No clear cases of group effects resulting in biased outcome were found. Tacit negotiations and compromises dominated much of the decision-making. Such decisions give a more narrow representation of the reviewers' opinions than either a process based on open confrontation of the divergent views or a process based on independent reviews. Consequently, from the point of view of idealism central conditions for 'unbiased' outcomes were not present. Broad representation of divergent views and open discussions are the best ways to avoid bias and ensure acceptability of the outcome.

- *Bias on the individual level:* Most of the assessments of research quality were written by individuals without much intervention from other panel members. When there were disagreements between panel members, these were solved in some cases by minority decisions. Both individual decisions and minority decisions give ample room for the research interests and the scholarly points of view of single evaluators to be decisive for the outcome.

These cases of 'bias' imply that the views of the individual panel member selected by the research council play a central role. Individual views may be modified due to the panel context – by anticipating objections or other kinds of tacit negotiation – but the scholarly division of tasks give each panel member the major say on his/her own fields. In addition, the large scope and limited time of the evaluations does not allow thorough review or the kind of panel processes that best prevent bias.[154] The likely results are vague and weak reports and/or reports in which the panel members' prior impressions of the research and researchers under review are important.

There was no significant guard against the possible bias of the individual panel members. However, norms or considerations that modify or erase harsh assessments were important. Most evaluators wanted to help to promote the research under review, and took care not to write anything that could harm the units under review. Site visits gave the evaluees ample room to promote their case. There were more effective guards for the single panel members against expressing their antipathies than against expressing their sympathies in the evaluation report.

Bottom-up processes and a substantial element of chance are major implications of such research evaluations:

- Research evaluations are based on information from the researchers under review and the opinions of their peers (qualified evaluations). The evaluees are heard (as they are listened to by their peers/those writing the evaluation reports). Evaluees acquire increased status ('good evaluation') and/or their problems get attention.
- A substantial element of chance: one expert in an area gets the major say (the one appointed by the research council and willing to spend time on the task). The points of view of these experts get increased status (possible Matthew-effect for future evaluation tasks).

Table 7.1 *Central findings*

Central combined theoretical and empirical findings include:
• a common set of criteria of good research
• tacit bases of judgements and divergent judgements (tacit and divergent operationalisation and use of criteria)
→ large grey area of acceptable outcomes of evaluations
Central empirical findings include:
• the composition of the expert panel, the organisation of its work, its time limits, and the (lack of) group interaction may be decisive for the conclusions of the evaluations
• little overlap of competence and clear scholarly division of tasks
→ tacit decision-making on panels, little interaction on judgements and ample room for scholarly 'bias' and a substantial element of chance

Table 7.1 gives a brief summary of central conclusions. The combination of a broad scope of acceptable outcomes and an ample room for scholarly 'bias' emphasises the importance of the 'luck of the reviewer draw' (a substantial element of chance).

7.2 Research design and analytical tools in retrospect

The present study is explorative. It contains detailed analysis of selected cases, focuses on a broad group of factors and uses a variety of different approaches. This section discusses strengths and weaknesses in the chosen design and analytical tools with regard to answering the research questions.

Design

The empirical basis of this multiple case study is six evaluation processes. As the study is explorative and not set out to test specific hypotheses, no strict comparative design was used for the selection of cases. A 'mixed strategy' with some variation and some similarities between the cases was adopted (see Section 1.4.1). The cases cover a variety of different panels, research areas, different kinds of evaluation units and evaluations commissioned by several different research councils. The strength of this strategy proved to be that it allowed conclusions on the more general characteristics of expert panel evaluations of research, regardless of varying contexts. Interesting features common to all the cases are found: clear scholarly task division, little interaction on judgements and (mostly) tacit decision-making, ample room for scholarly bias and a substantial element of chance. The variation between the cases inspired the construction of some 'ideal types', which also describes contexts that are likely to give evaluations deviating from some of these characteristics (Section 6.5).

A central question when doing a multiple case study is whether more or other cases would have improved the possibilities for drawing general conclusions. To collect in-depth interview data to reconstruct decision-making processes have been essential for the conclusions of this study. Given the resources of the project, more cases would imply less in-depth study, and would probably not have provided this kind of understanding of decision-making processes. To include *other* cases, on the other hand, might have improved the basis for conclusions. If it had been possible to include some cases clearly deviating from the 'vague and kind assessments'-approach, this would allow a better understanding of processes resulting in more explicit and harsh assessments and also the background for different evaluation strate-

gies. As explained in Section 1.4.1, such cases were not available when this study commenced.

A variety of data sources has been combined: the evaluation reports, the research councils' files on the evaluations, interviews with panel members, oral information from the secretaries of the panels, and in some cases the evaluators' private notes and draft reports. Some scholars advocate, while others criticise, 'triangulation' of data sources as a test of the validity of qualitative data. The argument for triangulation is that a conclusion is 'more convincing and accurate if it is based on several different sources of information' (Yin 1989:97). The opposing argument is that it is naïve to assume 'that the aggregation of data from different sources will unproblematically add up to produce a more complete picture', and that it often means to use 'one account to undercut the other, while remaining blind to the sense of each account in the context in which it arises' (Silverman 1993:157, 158). An account being corroborated or refuted by other accounts should, of course, not be taken as a simple test of whether it is true or false. For the present study, I am convinced that drawing on different data sources has given a better basis for conclusions. For the kind of processes in focus, a broad set of data has been essential to get a better understanding of the context in which to analyse each account.

Direct observation of decision-making processes is not among the data sources. The question of whether or not to collect observational data is complex. The less routine the activity, the more likely that the presence of an observer taking notes to study the process may affect the participants and the processes. In another study of more routine peer review (of grant proposals) I found direct observation of panel meetings a valuable data source in the study of what affected the judgements and decision-making (Langfeldt 1998). Here, I experienced that the less the routined the panel was, the harder it was to gain access to observe meetings. In addition to the problems of access and interference, direct observation may complicate a comparative design. Direct observation means observations of processes with an unknown outcome, and less prior information of characteristics of the chosen cases complicates the choice of an appropriate comparative design (cases with different or similar outcomes).

It should be added that while observation is not among the explicit data sources of the study, direct observation has contributed to forming my general understanding of the decision-making process on peer panels, and thereby form a more general input to the study. The

above-mentioned study of grant peer review based on direct observation was undertaken while working with the present study, and served as a valuable source providing insight and understanding of central aspects of the evaluation of research and the character of judgements. In the final phase of the dissertation work I also served as a secretary of an expert panel evaluating research institutes, a job that allowed me to come very close to, and be part of, the kind of processes studied.

Theoretical approaches

The analytical tools included input from a variety of different areas and approaches: the sociology of science, the philosophy of science, theory of group behaviour, game theory, analysis of goal structures and ideal types. The philosophy and sociology of science, theory of group behaviour and game theory proved useful in the theoretical discussion, and were central in pointing out factors for analysis. They also provided an overarching frame for the empirical analysis. Below, the different ways that the various approaches contributed are briefly summarised.

- *The philosophy and sociology of science* were used to understand the constraints on actors and the bases of assessments. The main conclusions here are that there seems to be a common set of general criteria for good research, but no common understanding of such criteria. The bases of judgements are tacit and we may find a dual set of rules for assessments.

- *Theory of group behaviour* was used to study possible 'bias' at the group level. Conditions for good group work were identified. These conditions were not present on the studied panels as they had little overlap of competence and interests, clear scholarly division of tasks and little interaction of judgements. A substantial room for chance and bias were found.

- *Game theory* was used to outline the logic of different settings and implications of various constellations of interests on a panel (hypothetically in Chapter 3). Game theory also contributed to the understanding of the outcome of the controversies on assessments in some of the empirical cases. Moreover, game theory was used to explain the pull toward positive assessments on a heterogeneous peer panel (ideal Type I).

- *The analysis of goal structures* concluded that evaluators, evaluees and commissioning bodies had a common interest in the chosen approach (mild assessments and arguments for better research conditions). This analysis is input to an understanding of the evaluation strategies as well as the relation between evaluators, evaluees and commissioners.
- *Ideal types* were used to pinpoint central findings regarding the logic of decision-making of expert panel evaluations, and to illustrate how the policy-setting and the composition of the panel set central premises for panel interaction and for the content of the evaluation report.

The theoretical and empirical parts of the project are not closely integrated in the sense that the theoretical discussion points out hypotheses to be tested. This would restrict the focus in a way that is not fruitful for this kind of explorative study. In retrospect, with the insight on decision-making processes on expert panel evaluations gained through the study, theory and data might, of course, be better matched. For example, data collection and interview questions might be better designed in order to analyse the role of the commissioning body, to undertake a more detailed analysis of task division and overlapping competencies and interests on the panels, and also to study tacit co-ordination between panel members.

7.3 Policy implications

Several factors indicate that the kind of evaluations studied have limited value as a basis for policy-making. Firstly, there are no significant guards against the possible bias of individual panel members and the conclusions rest upon a substantial element of chance (i.e. *who* is responsible for the assessment of an area). Secondly, the reports are rather vague – they do not offer nuanced assessments of research quality, nor comparisons of the units under review. Organisational constraints set by the commissioners themselves contributed to this kind of outcome. The scope of the evaluations and the limited time available for panel interaction set clear limits to the thoroughness of review. In addition, the terms of reference (mandates) given the panels were not helpful with regard to

writing a report suited for policy-making – the intended use of the report was not specified, neither were there guidelines for approach or units of analysis.

Did such weaknesses affect the use of the evaluations? Lack of implementation of recommendations because of the Research Council's lack of trust in the recommendations is found in one case only. This was related to lack of trust in the representativity of the recommendation (due to minority decision). The Council's knowledge about possible bias and weaknesses here seems to have affected the use.

It is not obvious whether or how vague assessments affected the use of the reports. It should be noted that *recommendations* were clear and that there was no stated intended use that demanded nuanced assessments and clear comparisons. In addition, vague assessments may hinder evaluees' mobilisation against the report – a situation that may limit the possibilities of implementing the recommendations. Effects on policy and reasons for lack of implementation of recommendations consequentially need to be studied more broadly than the question of vague and feeble assessments and conclusions resting on a substantial element of chance and potential bias.

In general, the effects on policy of the studied evaluations are not obvious. Those who agreed with the conclusions of the evaluation reports received better arguments for more resources for long-term basic research and/or better conditions for particular areas/groups. These were the main recommendations of the evaluation reports. Concrete effects on policy of these recommendations demanded receptiveness by funding authorities.[155] The authority of the recommendations and the receptiveness from funding authorities varied. In most cases major recommendations were not implemented. One programme evaluation is the clearest exception.[156] Here, the main ideas of the evaluation report were influential on the continuation of the programme. Programmes are supposed to be 'dynamic' and temporary, and evaluations of programmes therefore stand a better chance of making a real impact than evaluations of research disciplines or research institutions.[157]

Given the lack of implemented recommendations, the reports have probably played a more substantial role in distributing status and credits than in reorganising research or reallocating resources.[158] Evaluations are part of scholars' credibility cycle (see Latour &

Woolgar 1986; Rip 1994). The points of view of some scholars get increased status, and the research of some evaluees get increased status ('good evaluation'). This implies an important function of research evaluations regardless of their policy implementation. It should be noted that the Matthew effect of credibility cycles increases the consequences of chance. Those who partly by chance get a good evaluation (by 'luck of the reviewer draw') increase their chances of future good evaluations.[159]

There are various reasons why recommendations may not be implemented. As mentioned, there are indications of lack of trust in the recommendations in one concrete case. More generally, the increase in research evaluations commissioned by funding authorities in the last twenty years may be taken as an indication of decrease in trust in the self-organisation processes of science (van der Meulen 1998:400). If lack of trust is the problem, peer dominated evaluation panels, and substantial input from the evaluees may in itself reduce the chances of the funding authorities' being receptive to the recommendations of the evaluation reports, and consequently make such evaluations less suitable as a basis for policy measures.

Another explanation of the lack of policy effects may be that the focus of the evaluations did not correspond to that of meet the commissioning research councils. A study of 27 evaluations commissioned by the Research Council of Norway concludes that they were better tools for administrative needs than for strategic research policy-making. Individual research evaluations are a poor instrument for overall and general research policy and strategic priorities. For this purpose evaluations on broader areas and whole research sectors focused on policy questions would be better policy tools (Brofoss 1998).

'More basic research', a central recommendation of the reports, *is* a central policy question on strategic priority (but here restricted to specific areas or units). Put forward by peers to the researchers under review, it is also a *predictable* recommendation, and might be regarded as 'peer bias'. I have found no indications of lack of trust or understanding by the commissioning research councils of the importance of basic research. Such lack of trust or willingness to priority might be found at the government and political levels, and lack of receptiveness at this level might have reduced the possibilities of implementation. If this is the case, peer panel evaluations have

weaknesses as policy tools: they recommend (predictable) policy measures with low chances of implementation.

7.3.1 Overlap of competence – a central factor to be improved

The degree of overlap of competence on the panels is a central factor that influences the possibility for bias on all levels:

- Organisational level: Picking more than one expert in a field would give a broader spectrum of expert opinions access to the process.
- Panel level: Interaction among panel members may give more thorough and less biased judgements, but is conditioned by overlap of competence.[160]
- Individual level: Overlap of competence and interaction among panel members would reduce the possibility of individual bias affecting the final evaluation.

Overlap of competence is more costly and implies more time-consuming processes. There is, as already mentioned, a limit to the time and resources that can or should be spent on preventing arbitrariness. Even with several experts with overlapping competence and good time for discussion, there may always be an element of arbitrariness influencing the outcome (e.g. depending on the decision-making method used when experts disagree (see Langfeldt 2001a). Yet, processes that define good research are important, and should be given the needed time and resources for proper work. Two experts assessing each unit under review and some time for discussing the results, would be the minimum needed if expert panel evaluations are to have some function exceeding individual review reports when it comes to assessing the quality of research.

The issues at stake regarding overlap of competence are the degree of chance and the thoroughness and legitimacy of research evaluations. These are central factors for the foundation on which the concept of 'good research' is constituted. The Research Council of Norway has research quality as a major policy aim, which should imply a responsibility to secure that the concept of 'good research' is founded on solid ground. Even if research councils do not make any explicit use of the conclusions of evaluation reports, evaluations are part of the processes defining what is good and worthwhile research.

To commission evaluations means to take part in these processes, and the commissioning body should try to avoid all obvious sources of possible bias. Moreover, the degree of chance and the thoroughness and legitimacy of the conclusions are central when evaluations are input to policy-making. From all the above points of view, overlap of competence on evaluation panels is a central element requiring improvement. To my knowledge, the composition of more recent evaluation panels does not differ from the cases studied with regard to overlap of competence.

It might be objected that overlap of competence would increase the problem with vague evaluation reports. In the data there is no indication of such effects. On the contrary, the least vague report was found in a case with some overlap of competence and interaction on assessments (and open confrontation of divergent views). Previous literature has suggested that 'rational and creative disagreement' may improve peer review (Harnad 1985).[161]

7.4 Unanswered questions

Scholarly 'bias'? Realism or idealism?

No conclusions are drawn on whether assessments based on a scholarly viewpoint or scholarly preconceptions of what is good and bad research are 'biased' or not. The answer depends on whether we adopt the point of view of idealism or realism. Scholarly 'bias' is bias according to realism, but is unlikely to be seen as bias from the point of view of idealism.

The present study does not draw any philosophical conclusions on realism versus idealism.[162] The existence and content of standards of good research, unrelated to how we assess good research (realism), are open questions. Nevertheless, further elaboration on intrinsic characteristics of good research (i.e. realism) is possible. Some elements of research quality may be said to be part of our definition of research as such (although the interpretation of such intrinsic criteria may still depend on context), other elements may have a clear context dependent character (e.g. scholarly value and relevance).

How do expert panels operate in other policy contexts?

The evaluation policy of the Research Council has changed since the evaluations studied here were carried out. Budgetary consequences in favour of the best units are now recommended in policy documents on institute evaluation (*Norges forskningsråd* 1994). An evaluation approach of general praise, no clearly negative criticism and vague assessments on 'not-so-good' units, combined with general arguments for the need of better research conditions, can hardly satisfy such ambitions.

Does altered policy change evaluation processes and outcomes? Are peer panels loyal to commissioners which demand clear ranking of evaluated units as tools for reallocations? The ideal types based on the findings in this study should be valuable for understanding all kinds of expert panel evaluation of research (Section 6.5). In one of the sketched ideal types, the policy context is to provide a basis for reallocation in favour of the best units, and the predicted outcome is a report with clear comparisons or rankings of the units under review.

It is an open question how evaluators tackle such a demand in practice. The ideal types of Section 6.5 are analytic constructs to pinpoint the logic and mechanisms of expert panel evaluation of research. They do not predict the outcome of real cases as real cases will not have the extreme and simple outset conditions of the ideal types. If the present study was supplemented by new cases with reallocations in favour of the best units as part of the outset conditions, it would be easier to elucidate the question on basis of comparative method. Such cases are still difficult to find in the Norwegian research policy context. As mentioned, the policy regarding institute evaluations has changed in the direction of reallocations in favour of the best units. However, the recommended budget implications in favour of the best units do not seem to have been implemented. Neither do the evaluation reports allow for comparisons between institutes evaluated by different panels (Brofoss & Langfeldt 1999). There are also varying degrees of ranking or comparisons of the units evaluated by the same panel.

Nevertheless, interesting cases deviating from those of the present study can be found among more recent field evaluations in the natural sciences commissioned by the Research Council of Norway. In particular, the evaluation of physics research has much more explicit

assessments than has been usual in Norwegian evaluation reports (The Research Council of Norway 2000). The report grades individual researchers in the evaluated field from 'poor' to 'excellent'. What kind of processes led to this result? According to ideal Type II, central factors leading to this result are: (a) panel members perceive the purpose of the evaluation as to provide a basis for reallocation in favour of the best units, (b) the panel does *not* include members from all research directions under review, (c) there are no conflicting research directions represented on the panel, and (d) there is overlapping competence between the panel members' areas. No conclusions as to whether such conditions were present or not may be drawn without comprehensive data on the panel members' positions and views. From the report and from the public debate that the report provoked, we may find tentative answers to three of these conditions: From the CV of the seven panel members (included in the evaluation report) there seems to have been some overlap of competence with regard to the eight sub-disciplines into which the evaluated research is divided in the report. According to some of the evaluated researchers, the panel did not cover all areas and directions under review (Rekstad et al. 2000). The mandate given to the panel does not say that purpose of the evaluation is linked to reallocation of resources, but such a purpose might be read out of the letter sent the evaluees prior to the evaluation (the letter is enclosed the evaluation report).

It should also be noted that in the UK interesting data might be found for the study of how expert panels tackle a demand for reports suited as tools for reallocation. In the UK a large part of research funds from the higher education funding bodies are allocated on the basis of the ratings given by expert panels (the so-called RAE – Research Assessment Exercise). Various panels rate the university departments at intervals of four to five years, and they do give precise grades (a 7-point scale is currently used). I am not aware of any studies of the work of these panels.[163]

Decision-making processes on the borderline between science and politics

The boundaries, relations and interactions between science and politics are complex questions to investigate. In the present study, science and politics are found to interact in the sense that the commissioning body (that is, the research council which represents the government) affects panels' conclusions and recommendations.

The composition of the panel sets premises for panel interaction and for conclusions. The commissioning body is also likely to affect the evaluation approach adopted by the panel if clear signals are given on such matters. It is also important to note that the various parties directly involved in the evaluations – the commissioners, the evaluators and the evaluees – seem to have had central interests in common: they had reasons to help forward the research under review and to welcome good arguments on how to improve research and research conditions. These common interests were served by the kind of conclusions and recommendations that dominated the studied reports: mild assessments and arguments for better research conditions.

The composition of the evaluation panel is found to be of vital importance for the outcome of evaluations. Two different methods of selection were applied:

- The evaluees were asked to propose evaluators and were in this way given formal input on the selection.
- The council found suitable candidates without any input from the evaluees.

The first model is found to enable broad representation of opinions on the panels, whereas the latter gives a much higher degree of consensus.

As my focus has been on the work of the evaluation panels, there are central questions raised by the findings that are not answered: What are the reasons and conditions for the choice of different ways of selecting panel members? Who are the actors setting the premises for the selection of panel members? What is the role of research council staff, council members and evaluees in such selections? To what degree are the commissioning bodies (policymakers/staff) aware of the consequences of their choices? How does the interaction between commissioners and evaluators set conditions for implementation of recommendations? Given the importance of the composition of the evaluation panel found here, these should be central questions for future studies of expert panel evaluations of research.

Results from such studies of the politics of research evaluation may also prove fruitful for the study of decision-making processes on the borderline between science and politics in other contexts. With regard to such factors as the importance of personal and professional

networks and the importance of evaluations/research for symbolic purposes, such studies should be especially useful for the study of use, or non-use, of science and scientific expertise in public policy-making.

Appendix A Definitions of central terms

This appendix deals with some central terms that are used in a specific meaning in the present study. The concepts of norms, rules, criteria and standards are used in relation to different phenomena and should be separated from each other. Furthermore, 'tacit' is used with two different meanings.

Norms, rules, criteria, indicators and standards

Social and professional norms refer to particular kinds of social expectations. Rules, criteria, indicators and standards, on the other hand, refer to more cognitive or technical regulations, and may or may not involve social expectations.

- *Social expectations* refer to the way people in general expect others, or certain groups of people, to react or behave, including the kind of opinions they expect others/certain groups to have.
- *Social norms* are social expectations that are sustained by the feelings of embarrassment, guilt, anxiety or shame that a person suffers at the prospect of being caught violating them. Social norms are shared by the members of a society or a group and supported by sanctions, either positive or negative sanctions (Elster 1989:99). Sanctions may range from subtle to severe expressions of approval or disapproval. Social norms may be more or less internalised and are often unspoken.[164]
- *Professional norms* are shared by members of a professional or scholarly group or community. The explicity and internalisation of the norms, as well as the embarrassment when caught violating the norms, may vary. Negative sanctions may be subtle, like loss of esteem, respect, or credibility.
- In contrast to norms, *rules* do not need to be related to social expectations and are not necessarily followed by sanctions. Rules may be voluntary in the sense that it may be up to the individual or group to choose what rule(s) to adopt in a particular situation,

i.e. rules may be regarded as a voluntary aid for regulating social or cognitive processes. Rules deal with how to precede or conduct in specific situations, or how to solve certain kinds of problems, such as rules of decision-making. In relation to assessments of research, rules may, for instance, advise on the criteria to use or how to combine or balance various criteria.[165]

- *Criteria* are attributes used as a basis for defining and judging a quality. One may operate with a hierarchy of criteria, e.g. truth as a criterion of good research, and consistency, profundity and completeness as criteria of truth.

- *Indicators* are more operational than criteria. For example, citations may be understood as quantitative indicators of intra-scientific relevance and contribution to knowledge.

- *Standards* state what score on indicators or criteria must be obtained for a certain kind of assessment (excellent, good, mediocre, etc.). For the evaluation of research there are few if any explicit standards – standards for assessing the quality of scholarly research are generally tacit.

Tacit knowledge and tacit decision-making

'Tacit' has different meanings depending on whether it is used about knowledge and basis of judgements (including criteria etc.) or about group decision-making.

- *Tacit knowledge* refers to personal, internalised knowledge, not theorised or written down, i.e. knowledge not easily made explicit. Tacit knowledge may include craft skills, 'know how' and intuitive knowledge. '[W]e can know more than we can tell. ... The skill of a driver cannot be replaced by a thorough schooling in the theory of the motorcar; the knowledge I have of my own body differs altogether from the knowledge of its physiology; and the rules of rhyming and prosody do not tell me what a poem told me, without knowledge of its rules' (Polanyi 1966/1983:4, 20).

- When used about *judgements or evaluations*, 'tacit' refers to the *basis* of assessments. Tacit judgements mean that the basis of judgements cannot fully be described or explained. The 'tacitness' of

judgements is a question of degree. Few judgements are based on decisive rules alone and few are based on 'blind' intuition.[166] The more difficult to make explicit, the more tacit is the basis of judgements.

When it comes to group decision-making tacitness does not refer to whether the basis may be made explicit or not, but whether the basis *is* explicit or not. *Tacit group decision-making* includes tacit negotiations, bargaining or logrolling, tacit compromises and implicit voting. In all cases the rules of decision-making are tacit, i.e. unspoken and implicit, and not necessarily understood in the same way by all participants.

- *Tacit controversy, tacit disagreement or tacit dissension* refers to unspoken, concealed differences of opinions, differences which might be known or unknown to the actors.

- *Tacit negotiations or tacit bargaining* refer to bargaining concealed as co-operative discussions: the fact that the outcome of the process is subject to bargaining between parties with (at least partly) conflicting interests, is unstated and concealed by the parties – meaning that claims or proposals for compromises must be put forward with use of another set of arguments than used in straightforward explicit bargaining. To bargain some sort of awareness is needed: it cannot be claimed that (tacit) bargaining is going on if none of the parties are aware of seeking a compromise/agreement on the basis of (at least partly) conflicting interests. However, the distinction between tacit bargaining and discussions may be subtle and hard to define.

- *Tacit compromise* refers to the result of (tacit) bargaining, concealed as agreement in the sense of concurring opinions, i.e. that there has been no 'giving and taking' to reach the outcome.

- *Tacit or implicit voting* refers to implicit majority decisions, i.e. a process where decisions are reached by sounding for a majority opinion. In contrast to explicit voting, implicit voting may conceal the existence of a minority.

Appendix B Interview guide

1 **Had you done this kind of evaluation before?**/have previous evaluation experience?

2 **Why do you think you were chosen to do this work?**/Who proposed you?

3 **Why did you take the job?**

4 There were various aspects of ..... which were to be evaluated. **I have some questions about the division of tasks:**

- Was it clear to you what were your tasks were? How?/when?/Who decided this?/formal/informal
- Could you have done other tasks than those you were given?
- In which of the fields that were evaluated did you have the best oversight?
- Were there questions that you fully or partly left to the others in the panel?

5 **Did the panel, formally or informally, receive assessments from external experts?**

6 **Did you have any advance knowledge about ..... in Norway?**

knew by reputation
have close colleagues who collaborate with
have collaborated/published with/supervised

If yes, was this information useful for the evaluation? Could you have done the same evaluation without advance knowledge? (Why/why not?)

7 **..... probably had some advance knowledge. Was this useful for the evaluation?**

- What about the other members of the panel, did they have such sources of information?
- If yes, did this influence his/her attitude to these researchers? (in what way? – advocate, oppose or withdraw from the discussion saying he was disqualified)

Did such sources of information have any influence on the content of the evaluation report?

8 **Was it an advantage being a foreigner** in the sense that you were free to write what you meant without thinking that some of your colleagues might not like it, and that you were to write a public report to be used by the Research Council?

- Did you write anything in the report that you wouldn't have written about research groups in your own country?

9 **What was the most important information for the evaluation?** (publications, site visits, personal experiences and contacts, external expertise, other sources)

- How did you get this information?/Who provided this information?
- Who decided what written material you were to examine?

10 **Was the evaluation work properly organised by the Research Council?** (enough information about the purpose of the evaluation, enough time for discussing and writing)

11 This is a list of criteria for evaluating scientific work (card is given):

- Pure scientific criteria: consistency, correctness, stringency, profundity, etc.
- Results: theoretical contributions, productivity, publications (how much, where)
- Intra-scientific relevance: relevance of subject, novelty, originality, cumulativity, citations
- Extra-scientific relevance: applicability, use, effects
- Properties of the researchers: achievement, motivation, ambitions, reputation, international position
- Properties of the surroundings: equipment, freedom, group size, financing, organization

What would you put on top if you were to rank these criteria?

Are any of them irrelevant?/Which of these are the least important?

Would the other members of the panel agree?

12 What criteria were the main considerations in the evaluation of.....?

- Why?/Was there any discussion about criteria?
- How did you decide what criteria to use?
- Did any of the panel members disagree?

13 How did you reach agreement on the assessment of the various research groups, etc.?

- In what way was the discussion conducted?/Who were the most active participants?
- How were your assessments/views received in the group?
- Did you get any comments?
- How was the evaluation report put together?
- (Case specific question: why are there few evaluative statements in the report/so few clear comparisons?)

14 To what degree did the terms of reference steer the evaluation work and the content of the evaluation report?

- Would the result have been the same without the terms of reference?
- Were the terms of reference precise?
- What about oral statements from the Research Council?
- What about the secretary?

15 Do you think the content of the evaluation report was predictable for the Research Council/the research groups/institutes?

- Did they get what they wanted?

16 What reactions have you got on the report?

- Was the work meaningful/what did you learn?

17 Peer review is traditionally made at the micro-level – on individual researchers – and is used for decisions about employment, promotion, publication, and funding of research projects. In the last twenty years or so, it has been usual to set up peer panels to evaluate larger units, like research groups, research disciplines, institutes and research programmes – and such evaluations are used for reallocations and setting other research policy priorities.

- Is this new kind of evaluation useful according to your experience?
- Are such evaluations appropriate as bases for setting research policy priorities?

	useful	appropriate for priority setting
Research groups		
Research disciplines (nationally)		
Institutes/institutions		
Research programmes		

- Why? Weaknesses?

18 Should peer review in these contexts

- be used alone?
- not at all?
- together with other indicators? (Publication analysis/Citation analysis/analysis of use and effects/other possibilities?)

19 How would you have undertaken a similar evaluation task today?

20 Have you retained notes or documents?

References

ABRC (1990): *Peer Review. A Report to the Advisory Board for the Research Councils from the Working Group on Peer Review.* London: Advisory Board for the Research Councils, November 1990.

Ajenstat, Jacques (1993): 'Empirical test of a computer-based support system for the evaluation of research grant proposals.' *Research Evaluation,* 3(2):68–74.

Albæk, Erik (1988): *Fra sandhed til information: evalueringsforskning i USA før og nu.* København: Akademisk Forlag.

Barnes, C. B. & R. G. A. Dolby (1970): 'The Scientific Ethos: A deviant viewpoint.' *Archiv. europ. sociol.,* XI:3–25.

Becher, Tony (1989): *Academic Tribes and Territories. Intellectual enquiry and the cultures of disciplines.* Buckingham: Open University Press.

Bozeman, Barry (1993): 'Peer review and evaluation of R&D impacts.' In Barry Bozeman and Julia Melkers (eds.): *Evaluating R&D impacts: methods and practice.* Boston: Kluwer Academic Publishers.

Brofoss, Karl Erik (1997): 'Trenger Forskningsrådet en ny evalueringspraksis?' *Forskningspolitikk* 2/97:10-11.

Brofoss, Karl Erik (1998): 'The Research Council of Norway's use of research evaluation: an assessment of research evaluation as a strategic tool.' *Research Evaluation,* 7(3):134-140.

Brofoss, Karl Erik & Liv Langfeldt (1999): *Forskningsrådets instituttevalueringer 1995–1999. En analyse av innhold og brukspotensial for Norges forskningsråd, samt bruk og effekter ved instituttene.* Oslo: Norsk institutt for studier av forskning og utdanning, U-notat 5/99.

Burnham, John C. (1992): 'How journal editors came to develop and critique peer review procedures.' In Henry F. Mayland & R. E. Sojka (eds.): *Research Ethics, Manuscript Review and Journal Quality.* Madison: ACS Miscellaneous Publication.

Campanario, Juan Miguel (1998a): 'Peer Review for Journals as It Stands Today – Part 1.' *Science Communication,* 19(3):181–211.

Campanario, Juan Miguel (1998b): 'Peer Review for Journals as It Stands Today – Part 2.' *Science Communication,* 19(4):277–306.

Ceci, Stephen J. & Douglas P. Peters (1982): 'Peer Review: A study of reliability.' *Change,* 14(6):44–48.

Christiansen, John and Lene Christiansen (1989): *Research on Research: Evaluation of evaluations in the Nordic Countries.* Copenhagen: COS Forskningsrapport 3/89.

Chubin, Daryl E. & Edward J. Hackett (1990): *Peerless Science.* New York: State University of New York Press.

Cicchetti, Domenic V. (1991): 'The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation.' *The Behavioral and Brain Sciences,* 14(1):119–186.

Cole, Jonathan R., & Stephen Cole (1985): 'Experts' "Consensus" and Decision-Making at the National Science Foundation.' In Kenneth S. Warren (ed.): *Selectivity in Information Systems. Survival of the Fittest.* New York: Praeger.

Cole, Stephen (1983): 'The Hierarchy of the Sciences?' *American Journal of Sociology,* 89:111–139.

Cole, Stephen, Jonathan R. Cole and Gary A. Simon (1981): 'Chance and Consensus in Peer Review.' *Science,* 214(20 November):881–886.

Cole, Stephen, Leonard Rubin and Jonathan R. Cole (1978): *Peer Review in the National Science Foundation. Phase one of a study.* Washington D.C.: National Academy of Science.

Collins, Harry M. (1985): *Changing Order. Replication and Induction in Scientific Practice.* London: Sage Publications.

Collins, Harry M. (2001): 'Tacit Knowledge, Trust and the Q of Sapphire.' *Social Studies of Science,* 31(1):71–85.

Cozzens, Susan (1990): 'Options for the Future of Research Evaluation.' In Susan E. Cozzens et al. (eds.): *The Research System in Transition.* Dordrecht: Kluwer Academic Publishers.

Daniel, Hans-Dieter (1993): *Guardians of Science. Fairness and Reliability of Peer Review.* Weinheim: VCH.

Davis, James H. (1992): 'Some Compelling Intuitions About Group Consensus Decisions, Theoretical and Empirical-Research, and Interpersonal Aggregation Phenomena – Selected Examples, 1950-1990.' *Organizational Behavior and Human Decision Processes,* 52(1):3–38.

Edge, David (1995): 'Reinventing the wheel.' In Sheila Jasanoff et al. (eds.): *Handbook of Science and Technology Studies.* London: Sage Publications.

Elster, Jon (1983): *Sour Grapes.* Cambridge: Cambridge University Press.

Elster, Jon (1989): *The cement of society. A study of social order.* Cambridge: Cambridge University Press.

Etzkowitz, Henry & Andrew Webster (1995): 'Science as Intellectual Property.' In Sheila Jasanoff et al. (eds.): *Handbook of Science and Technology studies.* London: Sage Publications.

Farr, James (1985): 'Situational Analysis: Explanation in Political Science.' *The Journal of Politics,* 47:1085–1170.

Fisher, Martin, Stanford B. Friedman & Barbara Strauss (1994): 'The Effects of Blinding on Acceptance of Research Papers by Peer Review.' *JAMA,* 272(2):143–146.

Frendreis, John P. (1983): 'Explanation of Variation and Detection of Covariation. The Purpose and Logic of Comparative Analysis.' *Comparative Political Studies,* 16(2):255–272.

Fürst, Elisabeth (1988): *Kvinner i Akademia – inntrengere i en mannskultur?* Oslo: NAVFs sekretariat for kvinneforskning.

GAO (1994): *Peer Review. Reforms Needed to Ensure Fairness in Federal Grant Selection.* Washington, D.C.: United States General Accounting Office, GAO/PEMD-94–1.

Garfinkel, Harold (1967): 'Some rules of correct decision making that jurors respect.' In Harold Garfinkel: *Studies in Ethnomethology.* Englewood Cliffs, N.J.: Prentice-Hall, Inc.

Garfunkel, Joseph M., Martin H. Ulshen, Harvey J. Hamrick & Edward E. Lawson (1994): 'Effect of Institutional Prestige on Reviewers' Recommendations and Editiorial Decisions.' *JAMA,* 272(2):137–138.

Giddens, Anthony (1984/91): *The Constitution of Society. Outline of the Theory of Structuration.* Cambridge: Polity Press.

Gilbert, G. Nigel & Michael Mulkay (1984): *Opening Pandora's Box. A sociological analysis of scientists' discourse.* Cambridge: Cambridge University Press.

Goethals, George R. & John M. Darley (1987): 'Social Comparison Theory: Self-Evaluation and Group Life.' In Brian Mullen & George R. Goethals (eds.): *Theories of Group Behaviour.* New York: Springer-Verlag.

Grigson, Dianna & Terry Stokes (1993): 'Use of peer review to evaluate research outcomes.' *Research Evaluation,* 3(3):173–177.

Gulbrandsen, Johan Magnus (2000): *Research quality and organizational factors: An investigation of the relationship.* Trondheim: NTNU, Doktor ingeniør-avhandling 2000:9.

Gulbrandsen, Magnus & Liv Langfeldt (1997): *Hva er forskningskvalitet? En intervjustudie blant norske forskere.* Oslo: Norsk institutt for studier av forskning og utdanning, Rapport 9/97.

Hansen, Hanne Foss & Birte Holst Jørgensen (1995): *Styring af forskning: Kan forskningsindikatorer anvendes?* Frederiksberg: Samfundslitteratur.

Harnad, Steven (1985): 'Rational Disagreement in Peer Review.' *Science, Technology & Human Values,* 10(3):55–62.

Hellstern, Gerd-Michael (1986): 'Assessing Evaluation Research.' In Franz-Xaver Kaufmann, Giandomenico Majone & Vincent Ostrom (eds.): *Guidance, Control, and Evaluation in the Public Sector.* Berlin/New York: Walter de Gruyter.

Hemlin, Sven (1991): *Quality in Science. Researchers' Conceptions and Judgements.* Göteborg: University of Göteborg, Department of Psychology, Doctoral Dissertation.

Hernes, Gudmund (1978): *Makt og avmakt.* Oslo: Universitetsforlaget.

Hovi, Jon & Bjørn Erik Rasch (1993): *Strategisk handling. Innføring i bruk av rasjonalitetsmodeller og spillteori.* Oslo: Universitetsforlaget.

Hull, David L. (1988): *Science as a process.* Chicago/London: The University of Chicago Press.

Janis, Irving L. (1982): *Groupthink. Psychological Studies of Policy Decisions and Fiascoes.* Boston: Houghton Mifflin Company.

Jasanoff, Sheila (1990): *The Fifth Branch.* Cambridge: Harvard University Press.

Kerr, Norbert L., Robert J. MacCoun, and Geoffrey P. Kramer (1996): 'Bias in Judgment: Comparing Individuals and Groups.' *Psychological Review,* 103(4):687-719.

Knott, Jack H., & Gary J. Miller (1987): *Reforming Bureaucracy. The Politics of Institutional Choice.* Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

Kuhn, Thomas (1962/1970): *The Structure of Scientific Revolutions.* Chicago: The University of Chicago Press.

Langfeldt, Liv (1997): 'Omfordelinger, styringsproblemer og rettferdighet.' In Tom Christensen and Knut Midgaard (eds.): *Universitetet som beslutningsarena.* Bergen: Fagbokforlaget.

Langfeldt, Liv (1998): *Fagfellevurdering som forskningspolitisk virkemiddel. En studie av fordelingen av frie midler i Norges Forskningsråd.* Oslo: Norsk institutt for studier av forskning og utdanning, Rapport 12/98.

Langfeldt, Liv (2001a): 'The Decision-making Constraints and Processes of Grant Peer Review, and Their Effects on the Review Outcome.' *Social Studies of Science,* 31(6):820–841.

Langfeldt, Liv (2001b): 'Fagfellevurdering.' In Bjørn Stensaker (ed.): *Kunnskaps- og teknologivurdering. Perspektiver, metoder og refleksjoner.* Oslo: Cappelen Akademisk Forlag, Kunnskapspolitiske studier.

Laband, David N. & Michael J. Piette (1994): 'A Citation Analysis of the Impact of Blinded Peer Review.' *JAMA,* 272(2):147–149.

Larsen, Bøje (1985): 'Forskningsevaluering – Problemer og muligheder.' In Egil Fivesdal (ed.): *Nærbilleder af forskning: Organisasjonssociologiske studier.* Copenhagen: Nyt fra Samfundsvitenskaberne.

Latour, Bruno (1987): *Science in Action.* Cambridge, Mass.: Harvard University Press.

Latour, Bruno & Steve Woolgar (1986): *Laboratory Life. The Construction of Scientific Facts.* Princeton, N.J.: Princeton University Press.

Luukkonen, Tertu (1995): 'The impacts of research field evaluations on research practice.' *Research Policy,* 24(3):349–365.

Lübcke, Paul (ed.) (1983): *Politikens filosofi leksikon.* Copenhagen: Politikens Forlag.

Lysgaard, Sverre (1961/1985): *Arbeiderkollektivet.* Oslo: Universitetsforlaget.

Mahoney, Michael J. (1977): 'Publication Prejudices: An Experimental Study of Confirmatory Bias in the Peer Review System.' *Cognitive Therapy and Research,* 1(2):161–175.

March, James G & Johan P. Olsen (1989): *Rediscovering Institutions. The Organizational Basis of Politics.* New York: The Free Press/Macmillian.

Martin, Ben, James E. F. Skea and E. Nigel Ling (1992): *Performance indicators for academic scientific research.* Brighton: Science Policy Reearch Unit, University of Sussex.

Mathisen, Werner Christie (1994): *Universitetsforskeres problemvalg – akademisk autonomi og styring gjennom forskningsprogrammer.* Oslo: Utredningsinstituttet for forskning og høyere utdanning, Rapport 7/94.

McLean, Iain (1987): *Public Choice. An Introduction.* Oxford: Basil Blackwell.

McNay, Ian (1999): 'The Paradoxes of Research Assessment and Funding.' In Mary Henkel and Brenda Little (eds.): *Changing Relationships between Higher Education and the State.* London: Kingsley.

Merton, Robert K. (1942/1973): 'The Normative Structure of Science.' In Norman W. Storer (ed.): *The Sociology of Science. Theoretical and Empirical Investigations.* Chicago and London: The University of Chicago Press.

Merton, Robert K. (1957/1973): 'Priorities in Scientific Discovery.' In Norman W. Storer (ed.): *The Sociology of Science. Theoretical and Empirical Investigations.* Chicago and London: The University of Chicago Press.

Merton, Robert K. (1963/1973): 'The Ambivalence of Scientists.' In Norman W. Storer (ed.): *The Sociology of Science. Theoretical and Empirical Investigations.* Chicago and London: The University of Chicago Press.

Merton, Robert K. (1968): 'The Matthew Effect in Science.' *Science,* 159(5 January):56–63.

Merton, Robert K. (1988): 'The Matthew Effect in Science, II. Cumulative Advantage and the Symbolism of Intellectual Property.' *ISIS,* 79(299):606–623.

Meyer, John M. & Brian Rowan (1977): 'Institutionalized Organizations: Formal Structure as Myth and Ceremony.' *American Journal of Sociology,* 83(2):340–363.

Midgaard, Knut (1965): 'Co-ordination in 'Tacit' Games: Some New Concepts.' *Cooperation and Conflict,* 2:33–52.

Midgaard, Knut (1976): 'Cooperative negotiations and bargaining: Some notes on power and powerlessness.' In Brian Barry (ed.): *Power and Political Theory: Some European Perspectives.* London: John Wiley.

Midgaard, Knut, Halvor Stenstadvold & Arild Underdal (1973): 'An Approach to Political Interlocutions.' *Scandinavian Political Studies,* 8:9–36. Also published in (and here cited from) Knut Midgaard (1993): *Strategi, politisk kommunikasjon og forhandlinger.* Oslo: Institutt for statsvitenskap, Universitetet i Oslo.

Mitroff, Ian I. (1974): 'Norms and counter-norms in a select group of the Apollo moon scientists: A case study in the ambivalence of scientists.' *American Sociological Review,* 39(August):579–595.

Mulkay, Michael J. (1977): 'Sociology of the Scientific Research Community.' In Ina Spiegel-Rösing & Derek de Solla Price (eds.): *Science, Technology and Society.* London: Sage Publications.

Mullen, Brian & George R. Goethals (eds.) (1987): *Theories of Group Behaviour.* New York: Springer-Verlag.

NAVF (1992): *Informatikk: Research and Teaching in Norway. A Critical Evaluation.* Oslo: The Council for Natural Science Research.

NIH (1996): *Report of the Committee on rating grant applications.* [US]: National Institutes of Health, Office of Extramural Research.

Niinilouto, Ilka (1987): 'Peer review: problems and prospects.' In *Evaluation of Research. Nordic Experiences.* Nordic Science Policy Council, FPR-publication No. 5 (NORD 1987:30).

NORAS (1992): *Evaluering av norsk arbeidslivs- og aksjonsforskning.* Oslo: NORAS evalueringsrapport 1/92.

Norges forskningsråd (1994): *Evaluering og finansiering: Prinsipper for årsrapportering, måltall, grunnbevilgning og ekstern evaluering. Rapport nr. 2 fra prosjekt om instituttpolitikk i Norges forskningsråd.* Oslo: Norges forskningsråd.

Norges forskningsråd (1996): *Evalueringshåndbok for Norges forskningsråd.* Oslo: Norges forskningsråd.

NSF (1996): *National Science Board and National Science Foundation Staff Task Force on Merit Review. Discussion Report.* [US]: National Science Foundation, NSB/MR-96–15.

Nylenna, Magne, Povl Riis & Yngve Karlsson (1994): 'Multiple Blinded Reviews of the Same Two Manuscripts: Effects of Referee Characteristics and Publication Language.' *JAMA,* 272(2):149–151.

OECD (1987): *Evaluation of research. A selection of current practices.* Paris: Organisation for Economic Co-operation and Development.

Olsen, Johan P. (1972): 'Alternative beslutningsprosedyrer i organisasjoner.' *Tidsskrift for samfunnsforskning,* 13(1):25–50.

Page, Ewan S. (1997): *Data Collection for the 1996 Research Assessment Exercise: Review.* [Bristol]: Higher Education Funding Council for England, M2/97.

Peters, Douglas P. & Stephen J. Ceci (1982): 'Peer-review practices of psychological journals: The fate of published articles, submitted again.' *The Behavioral and Brain Sciences,* 5(2):187–255.

Pinch, Trevor (1986): *Confronting nature: The sociology of neutrino detection.* Dordrecht: Reidel.

Polanyi, Michael (1966/1983): *The Tacit Dimension.* Gloucester, Mass.: Peter Smith.

Ravetz, Jerome R. (1971): *Scientific Knowledge and its Social Problems.* Oxford: Clarendon Press.

Rawls, John (1971): *A Theory of Justice.* Oxford: Oxford University Press.

Rekstad, John, Einar Sagstuen, Eivind Osnes & Bernhard Skaali (2000): 'Sviktende evaluering av norsk fysikk.' *Forskningspolitikk 4/2000:*20–21.

Richards, Evelleen (1991): *Vitamin C and Cancer: Medicine or Politics?* London: Macmillan.

Rip, Arie (1990): 'Implementation and Evaluation of Science & Technology Priorities and Programs.' In Susan E. Cozzens et al. (eds.): *The Research System in Transition.* Dordrecht: Kluwer Academic Publishers.

Rip, Arie (1994): 'The Republic of Science in the 1990s.' *Higher Education,* 28(1):3–23.

Silverman, David (1993): *Interpreting Qualitative Data. Methods for Analysing Talk, Text and Interaction.* London: Sage Publications.

Speck, Bruce W. (1993): *Publication Peer Review. An Annotated Bibliography.* Westport/London: Greenwood Press.

The Research Council of Norway (2000): *Physics Research at Norwegian Universities, Colleges and Research Institutes. A Review.* Oslo: The Research Council of Norway.

Travis G. D. L. & Harry M. Collins (1991): 'New Light on Old Boys: Cognitive and Institutional Particularism in the Peer Review System.' *Science, Technology & Human Values,* 16(3):322–341.

Tranøy, Knut Erik (1988): *Vitenskapen - samfunnsmakt og livsform.* Oslo: Universitetsforlaget.

Turney, Jon (1990): 'End of the peer show?' *New Scientist,* 22 September:38–42.

van der Meulen, Barend (1998): 'Science policies as principal-agent games – Institutionalization and path dependency in the relation between government and science.' *Research Policy,* 27(4):397–414.

Vedung, Evert (1997): *Public policy and program evaluation.* New Brunswick, N.J.: Transaction Publishers.

Weber, Max (1949): *On The Methodology of the Social Science.* Illinois: The Free Press of Glencoe.

Weinberg, Alvin M. (1963): 'Criteria for scientific choice.' *Minerva,* 1(2):159–171.

Weiss, Carol H. (1972): *Evaluation research: methods for assessing program effectiveness.* Englewood Cliffs, N.J.: Prentice-Hall.

Wennerås, Christine & Agnes Wold (1997): 'Nepotism and sexism in peer review.' *Nature,* 387(22 May):341–343.

Whitley, Richard (1984): *The Intellectual and Social Organization of the Sciences.* Oxford: Clarendon Press.

Wood, Fiona Q. (1995): *Issues and Problems in the Public Funding of University Basic Research.* Armidale: University of New England, Doctoral Dissertation.

Wood, Fiona Q. (1997): *The Peer Review Process.* Canberra: Australian Research Council. National Board of Employment, Education and Training. Commissioned Report No. 54.

Yin, Robert K. (1989): *Case Study Research. Design and Methods.* Newbury Park, CA.: Sage Publications.

Ziman, John (1990): 'Research as a Career.' In Susan E. Cozzens et al. (eds.): *The Research System in Transition*. Dordrecht: Kluwer Academic Publishers.

Ziman, John (1987): *Knowing Everything about Nothing: Specialization and change in scientific careers*. Cambridge: Cambridge University Press.

Ziman, John (1984): *An introduction to science studies. The philosophical and social aspects of science and technology*. Cambridge: Cambridge University Press.

Zuckerman, Harriet (1988): 'The Sociology of Science.' In Neil J. Smelser (ed.): *Handbook of sociology*. Newbury Park, CA.: Sage Publications.

Zuckerman, Harriet & Robert K. Merton (1971): 'Patterns of Evaluation in Science: Institutionalisation, Structure and Functions of the Referee System.' *Minerva,* 9(1):66–100. (Also published as 'Institutionalized patterns of Evaluation in Science.' In Norman W. Storer (ed.) (1973): *The Sociology of Science. Theoretical and Empirical Investigations*. Chicago and London: The University of Chicago Press.)