Article

S|S|S

Social Studies of Science
0(0) 1–25

DOI: 10.1177/0306312712458478
sss.sagepub.com
SAGE

Communicating and compromising on disciplinary expertise in the peer review of research proposals

Katri Huutoniemi
Department of Social Research, University of Helsinki, Finland

Abstract

This paper analyses peer review deliberations in four evaluation panels that differ in terms of scope and disciplinary heterogeneity. Based on evaluation reports and discussions with panel members, it illustrates a variety of ways in which reviewers bridge their epistemological differences and achieve consensus on the quality of research proposals. The analysis demonstrates that peer review panels are forums in which communication across disciplinary boundaries occurs and interdisciplinary judgments arise. At the same time, disciplinary gate-keeping and incommensurabilities may set limits on such communication. The comparison of deliberative processes sheds light on how collective judgments are shaped and constrained by the disciplinary set-up of the panels in which the reviewers operate and in which the intersubjective dynamics of the deliberations unfold. Based on these findings, the paper considers conditions that may enhance disciplinary interaction and complementary judgments in the peer review of proposals, and thereby expands the prospects for interdisciplinary research.

Keywords

criteria, discipline, evaluation, expert judgment, interdisciplinarity, peer review

This study is concerned with the internal functioning of peer review, the practice through which scholarly work is evaluated by those with demonstrated competence. It analyses the ways in which peer review panels produce consensual judgments on the quality of research proposals and how reviewers are able to bridge their epistemological differences in this process. The topic is of interest, especially for those of us who are concerned with the status and fate of less established forms of inquiry – most typically,

Corresponding author:
Katri Huutoniemi, Department of Social Research, University of Helsinki, PO Box 18, 00014, Helsinki, Finland
Email: katri.huutoniemi@helsinki.fi

interdisciplinary research in its various forms. Thus, the aim of this paper is to investigate the disciplinary interaction that occurs in peer review deliberations, and to consider its effects on the evaluation of research proposals. The practical goal is to find a way to compensate for implicit biases in peer evaluation and to broaden the criteria used in research assessments. Despite the widespread concern about interdisciplinary research and the uncertainties in its evaluation, it is rarely asked how disciplinary interaction in peer review can be strengthened.

'Discipline', understood as an intellectual structure, denotes a set of codified rules, beliefs, perceptions and procedures with regard to producing and evaluating knowledge. These often tacit standards determine the criteria for admission into a community of scholars, the range of problems considered important, the approaches considered appropriate and the criteria for new knowledge (Russell, 1983). The existence of rules is claimed to provide clear indicators for assessing performance within a discipline (Feller, 2006). Whenever research expands, integrates with, or challenges the disciplinary canon, the problem arises that the epistemological standards of the disciplinary community may prove insufficient or even counterproductive (Huutoniemi, 2010; *Research Evaluation*, 2006). Interdisciplinary research is, by definition, a hybrid compound of more than one discipline, and is thus a form of scholarship that is not easily amenable to evaluation.

Not only does interdisciplinary research present difficulties for evaluation, studies of researchers' conceptions of quality also have shown that quality criteria even *within* disciplines are rarely expressed in unambiguous terms (Hemlin, 1991). Interviews with experienced evaluators suggest that good research is something that they 'feel' or 'experience' as much as 'analyse', and many experts state, 'I know good research when I see it' when asked to explain how they go about identifying quality (Gulbransen, 2000; Lamont, 2009; Lamont et al., 2007; Langfeldt, 2004; Thorngate et al., 2009). It is thus crucial to look beyond the criteria defined in methodological textbooks in order to explicate how quality is assessed in concrete evaluation settings. A study of evaluation practices in general, and the practices of peer review panels in particular, can inform us about how standards are intersubjectively constructed and how they determine what is prized in research.

Of prime concern in the literature of peer review has been the norm of universalism, as opposed to different forms of particularism, and the extent to which universalism is realized in practice (Cole, 1992; Cole and Cole, 1981; Cole et al., 1978; General Accounting Office, 1994; Merton, 1996; Roy, 1985). The questions posed by most researchers imply that a unified and fair process of evaluating knowledge can be put in place once particularistic considerations are eliminated. A smaller branch of the literature has, however, revealed various intrinsic 'biases' in peer review, such as 'cognitive homophily' (Lamont, 2009), 'cognitive particularism' (Travis and Collins, 1991), 'favoritism for the familiar' (Porter and Rossini, 1985), 'professional bias' (Langfeldt, 2004) or just 'peer bias' (Chubin and Hackett, 1990; Fuller, 2002). In short, these studies suggest that the particular cognitive and professional platforms that are the basis for authoritative evaluations are, at the same time, sources of potential bias. In addition, professional expertise is often accompanied by embodied knowledge, views and expectations, which are extremely difficult if not impossible to differentiate from well considered, 'unbiased' judgments. It is thus debatable whether one can talk about 'bias' with respect

to these 'emotional-cognitive' (in the term used by Thorngate et al., 2009: 16) aspects of judgment, and rather than being 'controlled away', they should be considered as an inherent part of evaluation.

In the light of the latter discussion, the evaluation of interdisciplinary research concerns a more complex issue than merely a lack of explicit criteria or established procedures (cf. Boix Mansilla et al., 2006; Shimada et al., 2007). While a more transparent formalization of the evaluation procedure would serve to institutionalize interdisciplinary criteria (e.g. Huutoniemi et al., 2010; Klein, 2008; Laudel and Origgi, 2006), it would probably not change the less explicit repertoires scholars employ when identifying quality. What seems most important, and yet understudied in research on interdisciplinary evaluations, is how reviewers' cognitive affiliations with particular disciplines unfold in the evaluation process. Also important is how such aspects of their judgments can be 'compensated for' in order to strengthen interdisciplinary deliberation.

Among the frequently cited cognitive tendencies in peer review, two stand out as particularly prejudicial for interdisciplinary research. First, evaluators are likely to perceive excellence in a way that resonates with the work that they themselves are conducting ('cognitive homophily'). Second, evaluators tend to be conservative and to protect their own epistemic territory from new perspectives and approaches ('gate-keeping'). These tendencies imply that the prospects for interdisciplinary proposals are not very good, since they may be regarded as unorthodox examples of disciplinary research (Laudel, 2006; Travis and Collins, 1991).

A general consensus prevails that a panel of experts from different fields is better equipped than any individual expert to assess the quality of interdisciplinary endeavours. It is also assumed that collective deliberation by such a panel produces a more comprehensive and balanced evaluation than a composite review by several independent evaluators (Boix Mansilla et al., 2006; Langfeldt, 2006). On the one hand, in-depth ethnographic research on multidisciplinary peer review panels confirms that a group context may indeed prevent reviewers from institutional bias and encourage methodological pluralism, since panellists often lose their credibility with colleagues if they push their own fields too strongly (Lamont, 2009; Lamont et al., 2006). On the other hand, studies on the decision-making of evaluation panels suggest that collective evaluation processes often involve a clear division of scholarly tasks, little interaction, and tacit compromises (Grigg, 1999; Langfeldt, 2004; Olbrecht and Bornmann, 2010). Thus, while attempts are often made to manage the disciplinary 'bias' of peer review by striving for high group diversity, the social or interpersonal context of judgment creates its own challenges.

The present study continues the effort to open the 'black box' of quality judgment by analysing the ways in which experts in differently composed groups evaluate research proposals. It focuses on the disciplinary interactions and the bargaining that occur between peer reviewers in different panel compositions. The analysis is meant to contribute to our understanding of variations in evaluation processes across scholarly contexts. In the emerging literature on evaluation practices, there has been very little comparative work (see, however, Lamont and Huutoniemi, 2011). A particular goal of this study is to consider which conditions may enhance disciplinary interaction and complementary judgments in proposal peer review, thereby enhancing the prospects for interdisciplinary research.

In the section that follows, I describe the materials and methods of this study. I then go on to examine the intersubjective process of evaluation in four different peer review panels. My analysis pays attention to the ways in which the members of the different panels brought together their different areas of expertise and quality criteria, and how they perceived and justified this process. I then discuss my findings in a more theoretical manner and ask what they offer for interdisciplinary evaluation. The concluding section summarizes the findings and raises pragmatic issues about designing the peer review process.

Data and methods

The study draws on a small sample of peer review panels organized in a recent year by the Academy of Finland, a national funding agency in Finland.[1] The panels were put together to evaluate research proposals submitted to one of the Academy's key funding instruments, 'Academy Projects'. In the peer review model used by the Academy, recognized scholars from all over the world (but mostly from other European countries) are invited to a panel meeting (usually for one to two full days) where they collectively rate proposals according to their scholarly quality. Although funding officers instruct panellists about evaluation criteria, panellists are given full sovereignty over rating the proposals. Each panel is assigned a subset of proposals emanating from the same research area or concerned with similar subject matter. The proposals are grouped and the panel members selected by the Academy staff, helped by experts of the four Research Councils. Before the meeting, individual panel members are asked to make a preliminary review, including a suggested score (1–5), for several applications. Each application is assigned to two (or sometimes more) panellists by the funding officials. All of the preliminary reports as well as the applications are available online for the whole panel prior to the meeting, but the panellists are not obliged to read them. During face-to-face deliberation, the panellists are required to achieve consensus in their ratings and collectively write an evaluation report of each proposal. The evaluation reports are sent back to the respective applicants and forwarded to one of the four Research Councils of the Academy, and that Council makes funding decisions for a wider range of proposals considered by a number of panels. In the year of this study, approximately every fourth proposal was funded.

I selected four panels for this study; the Academy designated the fields of these panels as *Environmental Ecology, Environmental Sciences, Social Sciences, and Environment and Society*. The panels were selected from the approximately 20 panels that were invited to evaluate the project proposals in the given round. The idea was to include panels with varying degrees of specialization in the research fields I am familiar with. The Environmental Ecology panel (ECO-ENV) was quite unidisciplinary: it operated with a thematically and epistemologically coherent set of submitted applications emanating from one broad field, the ecology of aquatic environments, and all the panellists were ecologists of some sort. The Environmental Sciences panel (ENV) evaluated proposals that dealt with natural processes in various environments, including forests, soils, peat lands and vegetation. It was multidisciplinary, since both the proposals and the panellists spanned multiple fields. The Social Sciences panel (SOC)

was also multidisciplinary, and considered proposals from sociology, social psychology, social policy, social theory, social work and cultural studies. It was composed of specialists from these various fields. The Environment and Society panel (ENV-SOC) was clearly interdisciplinary: it included a heterogeneous group of panellists working with a diverse set of proposals. The panellists often had degrees in more than one discipline and were knowledgeable on a wide range of interdisciplinary topics. They considered multi- or interdisciplinary proposals that dealt with environmental issues or with social-environmental interactions from a social, political, economic, technological or other perspective beyond the natural sciences. The scope and composition of each panel is summarized in Table 1.

Various forms of empirical data were collected on each panel, including: the project proposals under review (n=109 – the total for all panels); the preliminary reviews and scoring of the proposals prior to the panel meetings; and the evaluation reports and scores as defined collectively by the panels. Using these data as a guide, I then conducted phone interviews with 18 (out of 27) panel members (designated in this paper as P1–P18) shortly after the panel deliberations. The selection of interviewees included the majority of those panellists who were willing to be interviewed within a few weeks after the panel meeting. As all the panellists could not be interviewed, the views collected do not necessarily represent the group as a whole, and the self-selection of interviewees may have caused a bias in how the group process was represented. However, given the number of commitments and busy schedules of these high-ranking academics, this seemed to be the only conceivable strategy to get first-hand information on the evaluation process. This was the case, especially because I was not allowed personal access to panel meetings by the funding agency, and so direct observation was not possible. Access to these confidential procedures is commonly restricted; unfortunately, the resulting paucity of evidence sets limits on this study.

Table 1. The four evaluation panels compared: the number and scope of proposals under evaluation, and the number and expertise of panel members in terms of educational background

	ECO-ENV Environmental Ecology	ENV Environmental Sciences	SOC Social Sciences	ENV-SOC Environment and Society
Proposals	n=19 Ecology of aquatic environments	n=46 Natural processes in various terrestrial environments	n=22 Various social phenomena	n=22 Social–environmental interactions
Panel members	n=5 Limnology (2) Microbiology (1) Microbial ecology (1) Stream ecology (1)	n=11 Forest ecology (2) Plant physiology (2) Soil ecology (2) Mycology (2) Chemistry (1) Microbiology (1) Micrometeorology (1)	n=5 Sociology (2) Social policy (1) Social work (1) Social sciences (1)	n=6 Social sciences (2) Economics (2) Ecology (1) Mathematics–psychology (1)

The interviews lasted approximately an hour, and they loosely followed a schedule structured beforehand. The interviews were conducted in English, although only a fraction of the panellists were native English speakers. During the interviews, the panellists were asked to profile their personal expertise as well as the collective expertise of their panel, to explain what happened in each controversial case, to describe the arguments they themselves made about a range of proposals, to contrast their own arguments with those of other panellists, and to consider the meaning of the deliberation process in general. In addition to conducting the phone interviews with the panellists, I conducted face-to-face interviews with two Finnish funding officers who convened and attended the panel meetings. My main goal with these two interviews was to find out the formal procedures of evaluation in the specific panels. I conducted similar interviews with eight funding officers during an earlier project (Bruun et al., 2005) that guided the design of this study. All interviews were recorded and transcribed.

I used a qualitative research design to analyse data from the panels and to explore in depth the intersubjective aspects of evaluation that emerged from the deliberations among panel members. The empirical focus was on how each evaluator evaluated proposals and negotiated with the other panellists, and especially how the composition of the panel and the substance of the proposals created constraints on how the negotiation proceeded (see Boltanski and Thévenot, 1999; Muller-Mirza et al., 2009). I began by focusing on proposals that had aroused different opinions, in an effort to make sense of those disagreements on the basis of the evaluation reports and the panellists' own accounts. I then pinned down the strategies through which consensus was reached in each panel as well as the meanings panellists assigned to this process. Throughout this analytical work, I compared and contrasted accounts of each of the four panel's processes, in order to illustrate how the negotiations were shaped and constrained by the scholarly settings in which the panellists operated.

Perceiving and bridging expertise

Since the task of peer review panels is to produce evaluations of proposals, the panellists' first effort is to put their collective expertise to efficient use. While most panel members had some knowledge of all of the issues covered in the proposals, they tended to be good at different things, and so each had something unique to bring to the table. Given that the panel members typically did not know one another before the meeting, they came to discover their competences in an incremental way as the work went along. Their eventual views, however, were constrained by the proposals at hand and the discussions that followed. Consequently, substantial differences occurred in how panellists in each committee came to understand, utilize and coordinate their evaluations of the proposals, as well as in how they assigned meanings to this process (for a summary, see Table 2).

Given the abundance and compatibility of expertise in relation to the evaluative tasks at hand, the ECO-ENV panellists did not need to bother themselves much with assessing each other's competences. When opinions differed, the panellists justified their stances on the basis of what they knew about the topic. In most cases, a panellist with more competence on the topic was able to convince the others that, for example, the proposed methodology was not appropriate, or the research problem was not original. In the

Table 2. Components of collective judgment in the context of each panel

	ECO-ENV	ENV	SOC	ENV-SOC
Bridging expertise	Pooling, aggregating	Integrating	Deferring to the best expert(s)	Generalizing
Agreeing on rating	Equalization	Calibration	Contextualization	Commensuration
Perceived value of deliberation	Accuracy, consistency	Robustness	Fairness	Empowerment

following incident, an expert in microbiology explained why he deferred to other panellists who knew more on a topic:

> I think [the debate] was mainly about the relative importance [of a microbiology proposal] in this field. This sea ice business is a rather specific field that I don't know so much about, and I considered it an interesting area. But then it turned out that there is already broader background information existing in this area, and from that point of view, the novelty of this approach, or of this project here, was lower than I had assumed. (P1)

In this way, these panellists were able to combine their areas of expertise and fill gaps in each others' knowledge. Pooling expertise was believed to provide highly refined evaluations and was thus considered a crucial aspect of the panel deliberation. When I asked the panellists to explain what was most important for the evaluation process, they noted unequivocally that the discussions enabled their evaluations to get 'much more detailed, much fairer' (P2). A panellist explained that, since reviewers' judgments were often very similar, 'the discussion I think is important to validate the views of the individual evaluators. So it's a sort of a moderation of the quality assurance process, really' (P4). At the same time, this panellist felt it unnecessary for his evaluations to be 'validated' by others when he already was sure of himself: 'I feel most competent in subjects that are closest to my own. So, for example in terms of molecular work, I feel very confident of my position, *it's gonna be correct*' (P4).

The interviewees from the ECO-ENV panel were not convinced that the deliberation itself had anything other than instrumental value; for them, it was important because it made the judgments more accurate. For example, a panellist made the point that having more time for discussion in the panel meeting would have improved the process by allowing the panellists to go into *more detail*. Like other panel members, she highlighted the importance of each panellist's individual work as the primary source of appropriate judgment: 'I think that the evaluation you do by reading the applications, the remote evaluation, and that several people are doing that [independently], is the most precise way of doing it' (P2).

While overlapping competence was perceived as useful for 'validating' judgments, 'blind spots' in the panel's collective expertise caused some uneasiness. The panellists I talked with were unequivocal in expressing their view that their decisions could have been improved by having expert opinions about the particular proposals that dealt with areas that were not specifically covered by any panellist: for example, remote sensing

techniques, chemical analysis and developmental biology. While they asserted that they 'felt reasonably comfortable in forming an opinion' on such proposals by drawing on their discussions with colleagues, they expressed a concern that these evaluations were less valid than the others.

As for the ENV panel, the interviewees' accounts indicated that the panel's collective expertise turned out to be more than only a cumulative stock of knowledge. The review committee was composed of so many experts with different skills and specialties that their discussion soon revealed some variance between different panellists' perspectives. While they found that this variance caused divergent judgments, they also considered it a useful basis for producing robust evaluations. For instance, a mycologist explained that he had a different take on a proposal than an ecologist, because the two panellists had 'read the application from two completely different points of view' (P9). As both views were understood to be equally legitimate appraisals of the proposal, the experts framed them as complementary parts of a comprehensive judgment:

> In this case it was really important that the two evaluators were present …. He considered it from an ecological and physiological standpoint, and also of the importance regarding global change and so on. But my standpoint was more that is it possible to successfully finish the study with these [fungal] organisms with these prerequisites. (P9)

The ENV panellists often attributed divergent opinions to the complexity of the proposals themselves. Interviewees remarked that many proposals were multidisciplinary, 'bringing in different skills' or 'involving contributions to different disciplines'. (P7) The panellists observed that they could complement and reinforce each other's views and collectively form robust judgments on proposals about which they, as individuals, were not fully competent. In this context, the strength or independence of an individual expert's opinion was not seen as necessary for making an authoritative judgment: most interviewees from this panel perceived 'the width of scientific judgment' (P6) as a more legitimate basis for evaluation than the view of a specialized expert.

For instance, one proposal dealt with a newly developed technique that was highly valued by a specialist in the given field. Some other panel members were doubtful about the proposal because the technique had not been published despite several years of development. I was not able to talk with the specialist, but another panellist had the impression that 'in the end [the specialist] sort of really changed her mind. … If so many panel members feel that we should be careful about that, then [this specialist] saw you're right and we will be careful about it' (P5). In this and other similar cases, the interviewees stressed that 'you don't feel a person is less competent if they change their mind' (P8).

In two comparable incidents, a panellist with fully developed expertise on a proposal's topic was convinced by other participants that the proposal might be better than he initially thought. This panellist explained such incidents by the reasoning that it is hard not to be critical when a proposal is 'absolutely and 100% in your field':

> If [a proposal] goes to somebody who really does this every day, then from the start, they look at it and they can tell every new problem with it, and secondly, they are less impressed by it, because they do it every day, if you know what I mean. But it was one of those things that, in the discussion it came out when they said, 'Well, if you agreed with this proposal, would you

> have done it any differently?' And you should say, 'Actually no, it's a great idea, this is a really valuable proposal.' And you realized probably that actually you are a lot more critical if you are [very close to the topic]. (P6)

These examples illustrate that the panellists clearly believed the deliberation gave them some additional insight into the proposals. Some interviewees were convinced that only through collective deliberation were they able to produce reliable evaluations:

> It's better that two minds think about something together than in isolation. We could have an emergent thing that came out of the discussion that none of us on our own would have checked out. It just made it really clear which ones were the absolute best, and which ones had weaknesses. I think you wouldn't have gotten the same ranking in the end that you've got from the discussion; the discussion made it really clear which ones shouldn't be funded. (P10)

As was the case with other committees, the SOC panellists said they had considerable collective competence, and especially that each one could 'add a distinct contribution' (P16) to the panel's range of expertise. An interviewee reported that 'whenever policy questions came about we could, you know, rely on [one panel member]; and whenever questions of technological innovation came in, we could rely on [another panel member]' (P18). Such deference to expertise was also an important means of dealing with divergent evaluations. In cases where one panellist claimed greater competence in a proposal's topic, the other panellists were inclined to withhold their own judgments. One panel member strongly supported several cultural studies proposals in her preliminary reviews, but backed down on them after the deliberation:

> I'm not really a big expert on cultural studies, so when you know [a panellist's name] or someone said, 'oh, this is already known in cultural studies, this is not new at all', I've got to say, 'well, you're probably right in that case'. So that was certainly happening with several where I disagreed with [another panellist]. (P17)

Such deference and respect for the intellectual turf of each expert was customary in this panel, and lessened potential tensions between the panel members (see Mallard et al., 2009). At the same time, it seemed to undermine the importance of transparent criteria and shared responsibility for making evaluations. The SOC panellists sometimes adopted an advocate's role for proposals emanating from their own respective areas of expertise, and they sometimes also abandoned critical appraisal of proposals that represented different specialties from their own. Their discussion of some business school proposals illustrates the latter tendency and the salience of 'cognitive contextualization' (Mallard et al., 2009) more generally. When finding that there was no expert in that field on their panel, the panellists became worried about imposing sociological criteria on those proposals. A sociologist pondered:

> Obviously we could use a general social science expertise to evaluate the proposals, but … it was quite difficult for us to place them, as it were, academically, because we don't know what the norms and values of the business school kind of proposal might be. So, for instance, from a sociological point of view, we found them lacking in many ways, but it could be that within that kind of business and critical management studies those kinds of proposals are actually great

> some time, but we didn't have anyone with that exact area of expertise to, kind of, give us the kind of key markers. (P18)

Thus, the reviewers did not worry so much about a potential gap in their knowledge of this subject area as they did about not knowing *what criteria* to use. An analogous weakness in their panel was seen to result from the predominant nationality among the members. As the majority of panellists were British, some had questions about their eligibility to evaluate the proposals; they expressed concerns about being 'too anglocentric' or having a 'British bias' (P17). This did not refer to lack of expertise on, for example, Finnish society, but to their uncertainty about 'imposing criteria that we would use in our own national context on to this situation' (P18).

Debates on how to appropriately contextualize the proposals also occurred within the limits of particular fields. For example, a proposal on a minority culture in Finland was rated low by one panellist because 'it didn't pay sufficient attention to a particular area of theory which would have completely problematized the basic assumptions of their approach' (P18). Another panellist accepted this point, but disagreed about its relevance in the given case:

> That is a kind of way in which cultural studies has moved on in Britain, but they've still got a whole tradition of using [the applicant's] approach to ethnicity. So that's where we [debated], because I thought it to be unfair to judge it by the kind of standards of British cultural studies, which is you know one country, whereas [the applicant] was coming from a different direction. (P17)

The wide scope of the ENV-SOC panel obliged panellists to take positions on topics outside their core areas of expertise, and they negotiated their judgments with colleagues with whom they had relatively little in common. The chair reported that 'it was an interdisciplinary panel. I have a feeling that we were all chosen because we were inter[disciplinary], we were broad people' (P14). Because the proposals, as well as the reviewers, were clearly interdisciplinary, it often happened that the reviewers' judgments were based on relevant, yet completely different views of a proposal. For example, panellists recalled a proposal that dealt with nature conservation and social dynamics, which had been framed quite differently by the reviewers. An ecologist saw it as addressing a highly relevant problem that had not been properly conceptualized by previous research, and praised the set of case studies that were proposed for investigating the problem. Others focused their reviews more on the social-scientific design provided for the comparison of cases, which they found inadequate. The ecologist explained the disagreement as follows:

> I acknowledged to the other people where I thought the weaknesses were, and the other people said in the synthesis, that's fine. But I think that's a [personal] viewpoint as well as a disciplinary [view], because if that was in the conservation research end people would really value it, whereas I think the people from the social [research end] would say, 'well this isn't really giving us a particular new insight'. So I think it is a different disciplinary view: it's new to the conservation mixture [of disciplines] but may not be that interesting to the social scientists. (P11)

As indicated by this debate, insights from various scholarly traditions did not always mesh together well, and the ecologist characterized their meeting as 'one of the most difficult panels [on which] I have ever sat' (P11). Even so, the panellists were not inclined to defer to disciplinary expertise in their evaluation, partly because they acknowledged being invited due to their interdisciplinary competence. Indeed, most panellists were simultaneously involved in several different epistemic communities, which often required an ability to see beyond narrow disciplinary considerations and to compare proposals from a range of disciplines. The interviewees declared that they 'tended to be not worried too much about the different disciplines', but looked for more general qualities:

> What were we looking at was not, you know, particularly disciplinary attributes of the applications. We were looking at things like research design, mostly research design in such a way that is it going to produce useful results, would the results be useful for policy-makers, were … both the methodology and … well-explained and good, was it up to the data that they were thinking of gathering. They were more generic questions, rather than was it good sociology or good economics or good this or good that. And I think we all really took that view. (P14)

Overall, the legitimacy of the evaluation process in this panel was not entirely based on the use of specialized expertise or informed arguments. The panellists explicitly acknowledged that non-experts also had an important role to play in the evaluations. While the role of experts was to judge whether 'there's a proper methodology and proper question, because only they know the literature', a wider group was needed to 'ask bigger questions … like "why are you doing this", which can often be a shock to specialists because they haven't really viewed their own topic from the outside' (P11).

Given their broad understanding of expertise, these panellists were not concerned about what areas of knowledge they collectively lacked. When I probed for potential blind spots, I received the answer: 'There was no outlier, there was *nothing* out of the frame of [the panel's] competence' (P15). More than the scope of their expertise, they worried about the potential imbalance, since the panel comprised 'just one poor ecologist against three social scientists' (P11). In this context, reviewers appeared to feel entitled to rely on their knowledge, even of topics that were less familiar to them. When I inquired about how they had gone about evaluating a network analysis proposal that was given a high rating, a panellist responded:

> The proposal was very well written, it was very easy to understand. I don't think anybody in the room really had expertise in network analysis. But because they were clear what they wanted to do, people were very happy to accept that the proposal in this area, that we didn't have expertise on, was very valid. … If they write in a very clear fashion, and you're not familiar with the area, you tend to think, 'oh well that would be okay'. (P11)

In this and other similar cases, judgment was made on the basis of a shared willingness to support a well-written proposal that explained the project clearly for non-specialists.

Agreeing on rating

Peer review panels are engaged in a search, not only for what is qualified, but also for what is valuable or meritorious. What is it that counts, and according to what

measures? The evaluative culture of panellists' own disciplines influences how they define quality, including the relative importance they attach to various values. Differences between evaluative cultures thus pose a particular challenge for reviewers, as panels are expected to produce common judgments on the quality of the proposals assigned to them. The comparison of the evaluative practices of the four panels suggests that while value conflicts are often unavoidable, they can be interpreted and settled in different ways. In fact, the analysis brought out a variety of intersubjective mechanisms used for persuasion and to reach a settlement (for a summary, see Table 2).

When scholars act as judges, they implicitly commit themselves to go beyond their personal preferences and assess quality as defined through more objective standards (Lamont, 2009). Hence, an important premise for many panellists is that quality resides in the proposals themselves, rather than being attributed by the observer or by the particular comparative setting. However, as the ECO-ENV panellists were all ecologists, the panel turned out to assume that its epistemological norms were more self-evident than did the other committees, whose broader scope and multidisciplinary structure seemed to support more mutual recognition and exposure of epistemological assumptions (see Fuchs, 2001). When I asked about their criteria for quality and how different panel members identified a proposal as meritorious, the interviewees assured me that they were 'looking for exactly the same thing' (P4). Most of these panellists also did not feel that it was problematic to 'judge about other fields within science … [because] as an experienced scientist, you *know* what good science is – *even if it's not in your own field*' (P3).

The unproblematic status the ECO-ENV panellists gave to quality criteria, together with the fact that they often claimed expertise in the same field, made it hard for them to tolerate different opinions. This became evident in a series of disagreements that had to do with the comparative weighting of different strengths and weaknesses. Proposals that posed ambitious questions but presented somewhat unfocused research strategies were praised by one panellist while being denounced by another. Conversely, projects with modest scientific goals but well-argued, feasible research plans, including clear applications for the results, were not highly prized by the first panellist while the other one gave them high ratings. The first panellist expressed his frustration with these disagreements as follows:

> These are very highly productive, internationally profiled scientists doing very interesting things. They have very bold ideas, but there may be details in the methods; it's not very well thought through, and [another panellist] got stuck on that. Especially [an applicant's name], there was one part of this proposal that [the other panellist] didn't like at all, and then he wanted to just say no to the whole application. Just because of that little part … . The other parts of it were just splendid, yeah, they were just very good. (P3)

Since neither panellist was able to convince the other, their different criteria caused serious battles over academic authority. It was not easy for these panellists to change their minds as they both felt that, along with their judgment, their identity as experts was at stake. In such incidents, collective deliberation was needed in order to reach a compromise:

> When there was a clear disagreement, then the whole panel would give their views. But these would be views about the stances taken by the two evaluators, rather than the detailed science in the application. … People would give their views on how much relative weighting should be given to the [applicant's] track record versus the application. … [It was a] discussion of general principle, rather than the specific application. (P4)

The panel had also set up a routine procedure, in accordance with the guidance of the funding officers, whereby the task of drafting the evaluation report was assigned to a third panellist, who was 'a little further remote from the respective field' so that the given proposal was 'not so close to his personal emotions'. This panellist acted 'as a kind of independent judge … [who] could look more at the formal aspects, keep things equal, and judge across different cases' (P1).

These instances indicate that the reviewers did not openly contemplate their different values, but strived to *equalize* their judgments. A key function of this process was to avoid open conflicts while guaranteeing a more or less automatic repetition of the kind of behaviour that would produce consistent judgments. The panellists thus implicitly kept their disciplinary norms 'sacred', even though differences between the norms were sometimes obvious. One interviewee commented:

> It's just the scientific cultures. He is raised in an applied area of research and wants these applied issues to be funded, whereas I am raised in a more basic science and I want … the nice ideas, the good people, the ones who have shown that they produce good science, I want them to get funded. (P3)

More often than not, a compromise was found, indicated by an average rating, somewhere in the middle of the opposing positions. Of course, such a method of closure was often disappointing for the participants, and those discussions were characterized as 'meaningless' by a panellist who had strong opinions on the proposals. Sometimes, however, a compromise took the form of 'horse-trading' (Lamont, 2009: 121–125), where one panellist enabled another panellist's objectives to be realized in the hope that he would reciprocate: 'One case I could win, and in another case, or in several other cases, he won, so to speak. I use these words just to pinpoint the situation that it's like this – it's like convincing the other one that your grade is the right one.' (P3)

While a few instances of horse-trading were evident in the evaluation reports as well, completely missing was evidence of substantial changes in the final judgments due to the deliberation. In the other three panels, the final ratings sometimes went above or below the preliminary ratings (see Figure 1).

The experts of the ENV panel also believed that there was agreement in principle as to what constitutes good quality, and they told me of being 'amazed actually, how much the joint discussion, even for five minutes, really showed you that you have given a right mark to a proposal' (P6). However, as was the case with the ECO-ENV panel, occasional conflicts arose between different preferences. Pitted against each other were 'hypothesis-driven' versus 'screening' approaches, 'creative' versus 'feasible' objectives, and 'scientific' versus 'technological' relevance. Instead of trying to achieve a balance between the different preferences through some procedural mechanism, the panellists deliberated between their various normative stances. Most disagreements were settled by mutual

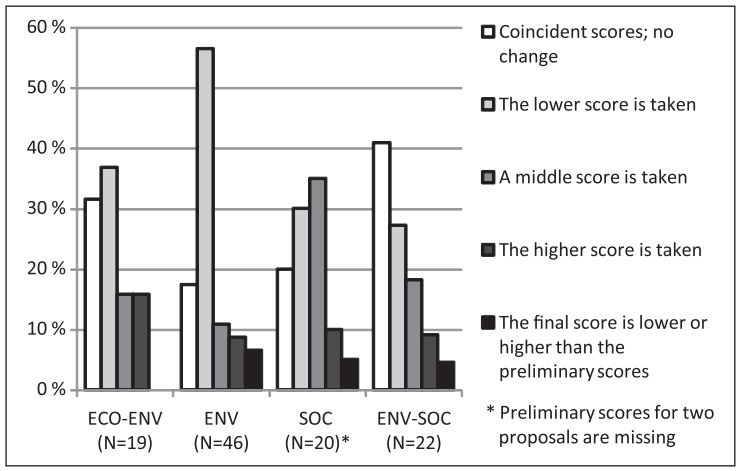

Figure 1. The effect of deliberation on scores across panels: the final score given to each proposal is compared with the two (or more) preliminary scores.

persuasion without causing anyone to 'lose face'. In the following disagreement, for example, the relative weighting of different criteria was discussed and the case was settled through consensus:

> [One reviewer] thought it was really good because of the approach to it, and [another reviewer] thought the approach was fine, but it was undoable, because [the applicant] had written every possible stress combination into the project you can think of. It was brilliant, yes, but impossible to do within the time scale. … [The former reviewer] based her review on the fact that this is really valuable science, and if she only did half of it, it would be worth doing. That was the philosophical question when looking at the proposals in the end. … But you have to review what they've written on the page. And if they've written something on the page that's impossible to do, then really and truly, they should rewrite the proposal and submit it next year, shouldn't they. That's pretty much the consensus we reached, at the end of this debate. (P6)

In cases of conflict, the ENV panellists often agreed to accommodate to the collective value frame that emerged from their deliberations. The panellists can be thought of as *calibrating* their individual senses of quality to a group standard in order to form a concerted evaluation. An important means of achieving a coherent evaluation was the reciprocal calibration of individual rating scales. At the start of their meeting, panellists had discussed which journals they regarded as most important in their fields of research, and they elaborated 'in what journals we would have published "outstanding", "excellent" and "very good" papers, or only "good" papers' (P9). Whenever they hesitated about giving the highest score, for example, they could ask: 'Can it, if we are lucky, be published in *Science* or *Nature*?' (P8). They thus established their collective criteria by setting indirect indicators of different degrees of quality, whereby they avoided quarrelling over disciplinary nuances. Interviewees highlighted the importance of such calibrating activity by referring to a few occasions where the preliminary scores coincided, but due to the calibrated rating scale, reviewers decided to lower their scores (see Figure 1).

As for the SOC panel, both the preliminary reports and the interviews suggested systematic differences in how panellists from different fields framed proposals. For example, a panellist who had expertise in empirical sociology, and conducted surveys and secondary analyses of quantitative data, was called 'technocratic' by another panel member. This 'technocrat' paid a lot of attention to the appropriateness of sampling strategies but also to the ethical dimensions of research methodology, while being less sensitive to proposals' theoretical ambitions or lack thereof. Another sociologist took the opposite position and defined her expertise in terms of certain theoretical positions. Not surprisingly, she was described by herself and others as operating under 'traditionally defined academic criteria' by looking for 'theoretical coherence, a clearly worked out relationship between method, methodology, and theory' (P18). She often assigned diminished value, or did not consider at all, whether a proposal included an elaborated research design that took into account various pragmatic constraints. Still another panellist was described by himself and others as a 'pragmatist' and explained his standards by explicitly contrasting them to the position taken by his more theoretically oriented colleagues:

> For somebody like myself, being at the very forefront of theories is a luxury that not everyone can afford and obtain, and so, if you live in very much an applied policy world … you have to do what you can, in a way. So somebody like myself, I'll be slightly more pragmatic and say, okay, there are a number of theories we can use here, I'm not going to privilege necessarily a particular author or position, I'm more concerned with what appears to be a plausible research design, and does it at least contain … some reliability, and some impact on outcomes that you could anticipate. (P16)

The SOC panellists believed that freely acknowledging their personal standpoints as an inevitable component of evaluation would help them become aware of their individual mindsets and make them more open to rethinking their evaluations. While each expert seemed to favour proposals that somehow spoke to her or his own interests (to 'technocratic,' 'pragmatic' or 'academic' criteria), the panellists were also prepared to alter their positions. An interviewee portrayed such shifts as a conscious choice: 'Am I going to kind of slightly rethink, or am I going to argue my case? One or the other, really' (P16). He explained such situations as follows:

> The panel would have to be explicit about how it understood the criteria in relation to the application, and those discussions would be explicit and substantive. One could then detect different perspectives around the criteria. … I think where positions were very different, I would say, 'This is my take on it, this is how I saw it', but, you know, 'okay, having heard what you said, and looked at some of the other applications where we had some similar discussions, I can see that I was possibly underestimating the importance of x, y, and z.' (P16)

Such judgmental openness, or awareness of how worldviews affect evaluation, made it easier for these panellists to discuss different points of view 'without people getting cemented into their position'. In strong contrast to the ECO-ENV panellists, for example, these social scientists often came back to review their positions and to re-examine a proposal, instead of 'pushing people back into their boxes' (P16). The

resulting compromises did not necessarily indicate agreement on the merits of proposals, but were often the result of conscious moves or academic politeness.[2]

In a few cases, however, disagreements clearly prevailed over politeness, and attempts to mediate between different judgments proved unsuccessful. A practical means of reaching compromise was to 'go through the [evaluation] form bit by bit' (P17). This involved leaving aside overall positions on proposals and discussing the text of the evaluation report. As explained by a reviewer who came to adjust her marks time after time: 'That kind of discussion about the different sections of the [evaluation] form was actually very fruitful for arriving at the overall consensus … . By coming to an agreement about the *text* on each section, you actually came to a compromise at the end.' (P17)

In contrast to the ENV panellists, who strove for a shared understanding on the general quality of each proposal, the members of the SOC committee found it easier to 'agree on details'. Such a strategy implied that the compromises negotiated in this panel were provisional, and that differences in framing the proposals were taken at face value rather than as discrepancies that should be explained away.

The way in which consensus was negotiated in the ENV-SOC panel produced some interesting similarities and differences when compared with the other panels. Like the ECO-ENV and ENV panellists, these experts noted a 'surprising amount of agreement' (P14) prior to any discussion. The fact that the panellists' ratings tended to coincide (see Figure 1), regardless of vast differences in their disciplinary and professional backgrounds, was interpreted as a strong sign that the intrinsic quality of the proposals was evident to every evaluator: 'When scientists with different backgrounds come to the same grade, there is at least *something* in it' (P12). At the same time, and like the SOC panellists, these interdisciplinary panellists were fully aware that their evaluative norms necessarily were influenced by their membership in particular cognitive and social networks.

In this evaluation context, the panellists avoided developing strong likes or dislikes towards particular proposals, or debating about epistemological preferences. Instead, they cultivated an appreciation for different kinds of research and tried to settle disagreements through mutual learning, compromising, or simply by trusting in each other's integrity and intuition. As one of the panellists described:

> It was quite a lot of looking at the criteria, they were up on the flip chart behind us, you know, what the grade 'one' was, what the grade 'two' was. And that was very useful, because in every slot you could put your hands on your heart and then say to each other: 'Do you really, honestly, think that it is a "good" proposal, or an "excellent" proposal? What do you think, really?' (P14)

The ENV-SOC panellists encouraged each other to downplay epistemological differences between disciplines and strengthen what was shared in their conceptions of quality. They often reached agreement through *commensuration*, a process by which the heterogeneous qualities of proposals were transformed into a standardized form in order to be compared (Espeland and Stevens, 1998). The analysis of evaluation reports and the discussions among panel members revealed that 'research design' emerged as an important criterion and also as a point of comparison between different proposals. An environmental sociologist, for instance, explained that in her evaluations, she used the same logic that guided her when teaching a research design course for incoming PhD students.

According to her view, science involves a unified methodology that is recognizable across disciplines:

> Whether they are remote sensing or feminist analysis or tree physiology or anything, science is science is science, and science is based on methodology and that's based on community … . If a research proposal can't be read and understood by another scientist from any discipline in terms of its scientific quality, then there is probably something wrong with it. [It should be clear] what the real question of the research is, what your research design is, how many samples you need and how you're going to think about the population, and how you're going to make it so that your research question and your methods and your analysis really do lead to reliable findings. It does not really matter so much whether you're talking about trees or fish or people. (P13)

However, the process of commensuration was sometimes costly and required thorough discussions of methodological questions. This became evident during a series of disagreements between two panel members whose opinions on several case study proposals were far apart. Both were experts in case-study methodology, but their theoretical backgrounds diverged. During a private discussion over breakfast, they came to an agreement concerning where their criteria of evaluation could overlap. One panellist explained:

> I had not been as critical on [particular methodological choices], because I've read [the proposals] in the context that I worked from, and I didn't have as much problem with these methodological decisions. But I concurred with his concerns when he went through them in some detail. (P13)

These practices made the ENV-SOC panellists more likely to be convinced by one another to change their initial evaluations of proposals. For example, while an economist was worried that a proposal in bioeconomics was not original or significant within the field of economics, he could be persuaded by the other panel members that the proposal should still score well on the basis of its high pragmatic value. In general, these panellists' broad understanding of expertise, as well as their belief in generalizable criteria, may have caused them to be less critical in their evaluations: the mean value of their preliminary scores, as well as the final scores they gave to the proposals overall, were higher than those given in the other three panels. Content analysis of evaluation reports suggests that the ENV-SOC panellists paid only slight attention to classical disciplinary criteria such as originality.

Discussion

Expert panel judgments have many intersubjective aspects that play a role in the evaluation of research proposals. In this paper, I have analysed only a fraction of the possible reference points that may shape judgments in an intersubjective evaluative context. The reference points I discussed have to do with panellists' disciplinary expertise and how it resonates with those of other panellists and with the proposals at hand. By comparing the four panel processes, I have illustrated variations in judgment and consensus making. In

this section, I will discuss the theoretical meaning of this variance by addressing questions such as: How does disciplinary expertise operate in the making of appraisals and communicating them to other experts? How may we take these insights into account to improve interdisciplinary evaluations?

Studies of peer review have shown that the negotiation of judgments establishes a sphere of social control and reciprocal accountability: the reviewers judge one another's standards and behaviour as much as they judge the proposals (Hirschauer, 2010; Lamont, 2009). This necessarily influences the way disciplinary expertise is used, since it makes each reviewer's disciplinary undertakings visible to the others and forces each to articulate her or his viewpoints in relation to those of the others. What is more, deliberations on individual proposals are more or less constrained by the particular set of proposals on the panel's table. This focus on local comparisons not only governs an individual panellist's appraisal of a particular proposal, but also other panellists' reactions to that appraisal (Lamont, 2009). Such collective control reinforces reviewers' perceptions of legitimacy and thus is an important part of their sense of procedural justice. On some occasions, collectively produced legitimacy may also give panellists more leeway and empower them to judge proposals more boldly or beyond their disciplinary expertise.

There appeared to be relatively tight disciplinary control among the ECO-ENV panellists. Since they occupied the same intellectual turf, they tended to view each other's appraisals in terms of their validity within the discipline and to compete for authority by 'spotting more problems' in proposals. The set of proposals they reviewed also was homogeneous enough to enable each panellist to closely monitor the consistency with others in his or her panel who used the criteria for evaluating proposals. This disciplinary competition may have influenced the panellists to hedge their remarks more than they would have done in other evaluative settings. The panel appeared to be quite selective, and possibly to filter out novel, deviant, interdisciplinary or anti-disciplinary proposals. Such a high degree of disciplinary control was not found among members of the other panels, whose accountability to one another became visible in other ways.

The SOC panellists were protective of their own disciplinary territories, but were prepared to give way to those who claimed better expertise. While their sovereignty in disciplinary issues may thus have been relatively high, they were sensitive to their co-panellists' suggestions that their own evaluative criteria might be unfair to an applicant. For example, one concern that emerged from this panel's deliberation was a potential 'British bias' in its evaluations.

The ENV panellists, in contrast, struggled to make more complete use of the various skills possessed by each panel member, in order to form majority opinions that combined these skills. This form of reciprocity often resulted in complementarity of judgments, as the value of deliberation became seamlessly intertwined with the extension and enrichment of different panellists' criteria. This complementarity of different evaluation frameworks sometimes led individual reviewers to recognize merits in proposals that they had not previously seen. Thus, in addition to urging criticism of proposals, panellists also encouraged each other's enthusiasm about, and support for, proposals, even in cases that they deemed to have 'undetermined' merit.

There was relatively little disciplinary control in the ENV-SOC panel, and it was sometimes impossible for its members to assess the expertise of other participants or the validity of their arguments. Moreover, as the proposals also tended to be interdisciplinary, their evaluation was more controversial from the outset. However, as with the other panels, deliberation established strong intersubjective ties that played a crucial role in creating trust. Rather than monitoring others' appraisals for appropriateness, the panellists often based their own judgments on the deliberation itself and on the shared standards that emerged from it. In the absence of an a priori epistemological framework, they believed that deliberation would lead to optimal decisions, as only it could allow for flexibility and for individual participants to develop a shared sense of merit in each case (Lamont, 2009).

Attributes of a successful proposal are likely to be somewhat different in evaluation settings where the authority to judge rests more on panel-wide dialogue, rather than on specialized expertise. A proposal has to speak to different audiences in order to receive a high rating after a discussion among diverse experts. In practice, such a proposal has to be written in a way that is easy for a supportive reviewer to present it in a compelling way to other panel members. To some extent, this appeared to be true in the ENV and ENV-SOC panels, but less so in the ECO-ENV and SOC panels. The proposals that received the highest scores in the ENV panel, for instance, typically dealt with interdisciplinary issues that had broad environmental significance, rather than with specialized questions designed to advance the state of the art in particular fields of environmental research. Moreover, only in the ENV and ENV-SOC panels did I find evidence that participants could convince others of the strengths of proposals that they had not yet recognized (in contrast, in all panels it was easier to make a persuasive argument about weak points in proposals).

How, then, may these findings be taken into consideration to make peer review work more effectively? First, some choices have to be made. As highlighted by Chubin and Hackett (1990; Hackett and Chubin, 2003), an optimal peer review procedure can hardly be put in place without some trade-offs between the various values the system is asked to serve. As stated at the outset, a particular concern of this paper is to enhance complementary judgments in peer review and thereby the validity of interdisciplinary evaluations. To meet these goals, drawing from my findings, I will highlight the priority of the intersubjective dynamics that encourage individual panellists to stretch their disciplinary standards in the service of dialog and mutual understanding. The comparison of panel deliberations suggests that a panel that develops good interdisciplinary dynamics does not, at the same time, allow much unidisciplinary discretion for individual panel members. As reviewers adapt their behaviour to take account of the views and arguments of other participants, some features of their own expert judgments gain strength while others are left aside. It seems that 'good' interdisciplinary and unidisciplinary judgments are not entirely consistent with one another.

When interdisciplinary considerations are given priority, at least two important choices have to be made:

(1) One important choice in organizing peer review panels concerns the selection of panellists in terms of their degree of specialized expertise. There is a continuum

between 'specialist' and 'interdisciplinary' (or 'generalist') panellists. While some aspects of proposals can be successfully assessed only by specialists who really know the subject (such as the adequacy and completeness of the applicant's account of the state of research, as well as the originality and the methodological correctness of the proposal), for other aspects specialist knowledge is unnecessary or even prejudicial. Significance and impact, for example, as well as pragmatic and societal utility, are often assessed on the basis of a general or interdisciplinary understanding of the field. The choice of panel members thus partly determines which aspects of proposals become decisive. Recruiting generalist panellists' from a wider pool of specialties is likely to improve the chances of interdisciplinary proposals, because such proposals typically are stronger in the aspects that non-specialists tend to focus on (such as relevance and pragmatic value), whereas they may fall behind in aspects that specialists examine (methodological correctness, stringency or solidity). The findings also suggest that interdisciplinary or generalist panellists use each others' experience and views as sources of their own judgments and usually have less difficulty with operating within multiple epistemological regimes.

(2) Another choice involves the mix of experts in a panel. There is a continuum between 'unidisciplinary' and 'multidisciplinary' panels, with varying degrees of overlap in competencies. High overlap in reviewers' competencies in unidisciplinary panels offers greater 'reliability' of evaluations, in the sense that panellists may easily calibrate their standards and validate one another. At the same time, a shared value framework may bring about 'consensual bias' by filtering out proposals that do not 'fit in' to a disciplinary frame (Langfeldt, 2004; Olbrecht and Bornmann, 2010). A multidisciplinary panel design, in contrast, ensures a breadth of expertise and creates a shared sense among panellists that they are accountable to multiple disciplinary communities. This encourages panellists to present their views by drawing on sources other than their own particular fields of knowledge. With an optimal amount of overlap, panellists can debate the strengths and weaknesses of different research approaches in relation to the proposal at hand. Too little overlap, however, may cause panellists to divide responsibilities between panel members and forgo such interdisciplinary deliberation.

The above choices also depend on other matters, of course. For example, the grouping of proposals, which also implicates the design of panels, needs to be tailored to the size of research fields and overall scientific activity in a country. In a small country such as Finland, even 'unidisciplinary' panels generally are broader than in large countries. On the whole, the disciplinary structure of science is much stronger in countries such as the US, where the sheer volume of scholarship within a field is many times higher than in Finland. Structures, however, can be changed in the long term.

In addition to the design of panels, other ill-defined factors are likely to play a role as well, ranging from a reviewer's personal wisdom and the norms that prevail in her field, to the procedural rules set by the funding organization (see Lamont and Huutoniemi, 2011). Many differences identified between, for example, the SOC and

ENV panels, very likely pertain to the different institutionalized practices of social scientists versus environmental scientists (Whitley, 1984). The evaluation rules of the Academy of Finland probably give rise to somewhat different consensual practices than, say, a more unstructured procedure where no preliminary reviewers are nominated (see Thorngate et al., 2009: 107–122). Various properties of the group, such as sex and age distribution and the number of participants, probably influence the deliberation rules, too.

Conclusions

'Peer consensus' is often believed to be an indicator of 'inter-rater reliability', and is typically regarded as the most valuable collective product of panel deliberation (see Brenneis, 1994; Cicchetti, 1991; Cole et al., 1981; Hemlin, 2009; Marsh et al., 2008). It indeed results in a clear signal for funding decisions. However, as demonstrated by the present analysis and several other studies on the topic, some variance in reviewers' judgments is inevitable. This hardly means that the outcome of evaluation depends mainly on chance, as some have suggested (Cole et al., 1981). The present study has been an attempt to make the variance in panel judgments more understandable. It has considered the reasons offered by panellists for disagreement, how the disagreements are negotiated and how such negotiations influence the evaluation outcome.

This study has also demonstrated that most disagreements can be settled through deliberation. Because reviewers understand proposals through their particular scholarly and professional lenses, communication about epistemological differences is a customary practice and can often lead to an agreement. This does not necessarily mean, however, that reviewers agree in the sense of reaching a shared understanding of a proposal's merits. Instead, the process is as much about agreeing, more or less tacitly, on the conventions for tolerating divergent views, adjusting initial views, and resolving disagreements. Such intersubjective rules are crucial, as they lead panellists to believe that the process is fair. Participants' faith in the evaluation process, in turn, has a tremendous influence on how well the process works (Lamont, 2009).

As is often the case with peer review of grant proposals, the panels analysed in this study were required by the funding agency to produce consensual decisions, and they evidently were able to do so. However, their consensual practices varied. The above comparison of deliberative processes suggests that an important, yet understudied, variable for explaining this variation is the mix of disciplinary specialties in a review panel. The requirement to negotiate one's own judgments with other panel members establishes a local sphere of reciprocal accountability, which necessarily influences the way in which panellists make their evaluations. Depending on other disciplinary standards and other panellists' behaviour, a panel member may acknowledge the strengths and weaknesses of a proposal differently when negotiating with colleagues from her or his own field than when doing so with colleagues from other fields.

These findings resonate with broader considerations of the competing functions of peer review and the way it relates to scientific knowledge production. It is often acknowledged that there are two different perspectives on peer review that are at odds with each

other. On the one hand, those who are concerned with upholding high standards of technical merit prefer review panels that are composed of more narrowly defined established experts. On the other hand, those who wish to promote innovative approaches and socially relevant research prefer review panels that represent more broadly defined constituencies (Eisenhart, 2002; Fuller, 2002; Hackett and Chubin, 2003; Weinberg, 1962).

The present study concurs with this two-pole view, and suggests that interdisciplinary goals are better served by adjusting the review process towards the latter pole. However, the study also illuminates why the selection of experts is so important. By highlighting the intersubjective dynamics that emerge during panel deliberations, the study suggests that the relationships between panel members create a temporary accountability environment, which, in turn, plays a major role in the kinds of proposals a panel tends to reward. While there is obviously no static relationship that would give judgments a predictable tendency (Hirschauer, 2010), the intersubjective context of evaluation could be designed in a way that facilitates interdisciplinary dialog.

The requirement of consensus is itself a strategic choice that does not come without consequences. As indicated in this paper, the consequences for the deliberative process and the decisions that follow from it are not self-evident but depend on context. Settling on an average between two extreme scores, for example, has different implications from, say, deferring to the judgment of the technically best expert or relying on a majority rule. We may need to be more conscious about what kind of consensuses our evaluations produce, and to be responsible about the distributive outcomes that follow. At the same time, considered moves could be made to promote consensual practices that lead to complementary judgments more often than they do to stand-offs between incommensurable viewpoints.

Funding

This work was supported by the Finnish Post-Graduate School in Science, Technology and Innovation Studies; the Academy of Finland [decision number 120577]; and the Emil Aaltonen Foundation.

Notes

1. The year is not given here in order to better guarantee the anonymity of both applicants and evaluators.
2. See Pomerantz (1984) for an analysis of how agreements in assessments are calibrated in ordinary conversations.

References

Boix Mansilla V, Feller I and Gardner H (2006) Quality assessment in interdisciplinary research and education. *Research Evaluation* 15(1): 69–74.

Boltanski L and Thévenot L (1999) The sociology of critical capacity. *European Journal of Social Theory* 2(3): 359–377.

Brenneis D (1994) Discourse and discipline at the National Research Council: A bureaucratic *Bildungsroman*. *Cultural Anthropology* 9(1): 23–36.

Bruun H, Hukkinen J, Huutoniemi K and Klein JT (2005) Promoting interdisciplinary research: The case of the Academy of Finland. *Publications of the Academy of Finland* 8/05. The Academy of Finland, Helsinki.

Chubin DE and Hackett EJ (1990) *Peerless Science: Peer Review and U.S. Science Policy.* Albany, NY: State University of New York Press.

Cicchetti DV (1991) The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioral and Brain Sciences* 14(1): 119–135.

Cole JR and Cole S (1981) *Peer Review in the National Science Foundation: Phase Two of a Study.* Washington, DC: National Academy Press.

Cole S (1992) *Making Science: Between Nature and Society*. Cambridge, MA: Harvard University Press.

Cole S, Cole JR and Simon GA (1981) Chance and consensus in peer review. *Science* 214(4523): 881–886.

Cole S, Rubin L and Cole JR (1978) *Peer Review in the National Science Foundation: Phase One of a Study.* Washington, DC: National Academy Press.

Eisenhart M (2002) The paradox of peer review: Admitting too much or allowing too little? *Research in Science Education* 32(2): 241–255.

Espeland WN and Stevens ML (1998) Commensuration as a social process. *Annual Review of Sociology* 24: 313–343.

Feller I (2006) Multiple actors, multiple settings, multiple criteria: Issues in assessing interdisciplinary research. *Research Evaluation* 15(1): 5–15.

Fuchs S (2001) *Against Essentialism. A Theory of Culture and Society.* Cambridge, MA: Harvard University Press.

Fuller S (2002) *Knowledge Management Foundations.* Boston, MA and Oxford: Butterworth-Heinemann.

General Accounting Office (1994) *Peer Review Reforms Needed to Ensure Fairness in Federal Agency Grant Selection*. Report to the Chairman, Committee on Governmental Activities, US Senate. Washington, DC: General Accounting Office.

Grigg L (1999) *Cross-Disciplinary Research.* A Discussion Paper. Commissioned Report No. 61. Canberra: Australian Research Council.

Gulbrandsen JM (2000) Research quality and organisational factors: An investigation of the relationship. Doctoral dissertation, Department of Industrial Economics and Technology Management, Norwegian University of Science and Technology, Trondheim.

Hackett EJ and Chubin DE (2003) *Peer Review for the 21st Century: Applications to Education Research*. Prepared for a National Research Council Workshop, Washington, DC.

Hemlin S (1991) Quality in science: Researchers' conceptions and judgments. Doctoral Dissertation, Department of Psychology, University of Göteborg, Göteborg, Sweden.

Hemlin S (2009) Peer review agreement or peer review disagreement: Which is better? *Journal of Psychology of Science and Technology* 2(1): 5–12.

Hirschauer S (2010) Editorial judgments: A praxeology of 'voting' in peer review. *Social Studies of Science* 40(1): 71–103.

Huutoniemi K (2010) Evaluating interdisciplinary research. In: Frodeman R, Klein JT and Mitcham C (eds) *Oxford Handbook of Interdisciplinarity.* Oxford: Oxford University Press, 309–319.

Huutoniemi K, Klein JT, Bruun H and Hukkinen J (2010) Analyzing interdisciplinarity: Typology and indicators. *Research Policy* 39: 79–88.

Klein JT (2008) Evaluation of interdisciplinary and transdisciplinary research: A literature review. *American Journal of Preventive Medicine* 35: S116–S123.

Lamont M (2009) *How Professors Think: Inside the Curious World of Academic Judgment.* Cambridge, MA: Harvard University Press.
Lamont M, Fournier M, Guetzkow J, Mallard G and Bernier R (2007) Evaluating creative minds: The assessment of originality in peer review. In: Sales A and Fournier M (eds) *Knowledge, Communication, and Creativity*. London: Sage, 166–181.
Lamont M and Huutoniemi K (2011) Comparing customary rules of fairness: Evaluative practices in various types of peer review panels. In: Camic C, Gross N and Lamont M (eds) *Social Knowledge in the Making.* Chicago: University of Chicago Press, 209–232.
Lamont M, Mallard G and Guetzkow J (2006) Beyond blind faith: Overcoming the obstacles to interdisciplinary evaluation. *Research Evaluation* 15(1): 43–55.
Langfeldt L (2004) Expert panels evaluating research: Decision-making and sources of bias. *Research Evaluation* 13(1): 52–62.
Langfeldt L (2006) The policy challenges of peer review: Managing bias, conflict of interests and interdisciplinary assessments. *Research Evaluation* 15(1): 31–41.
Laudel G (2006) Conclave in the Tower of Babel: How peers review interdisciplinary research proposals. *Research Evaluation* 15(1): 57–68.
Laudel G and Origgi G (2006) Introduction to a special issue on the assessment of interdisciplinary research. *Research Evaluation* 15(1): 2–4.
Mallard G, Lamont M and Guetzkow J (2009) Fairness as appropriateness: Negotiating epistemological differences in peer review. *Science, Technology, & Human Values* 34(5): 573–606.
Marsh HW, Jayasinghe UW and Bond NW (2008) Improving the peer review process for grant applications: Reliability, validity, bias, and generalizability. *American Psychologist* 63(3): 160–168.
Merton RK (1996) *On Social Structure and Science*. Chicago: University of Chicago Press.
Muller Mirza N, Perret-Clermont A-N, Tartas V and Iannaccone A (2009) Psychosocial processes in argumentation. In: Muller Mirza N and Perret-Clermont A-N (eds) *Argumentation and Education*: *Theoretical Foundations and Practices*. New York: Springer, 67–90.
Olbrecht M and Bornmann L (2010) Panel peer review of grant applications: What do we know from research in social psychology on judgment and decision-making in groups? *Research Evaluation* 19(4): 293–304.
Pomerantz A (1984) Agreeing and disagreeing with assessments: Some features of preferred/dispreferred turn shapes. In: Atkinson JM and Heritage J (eds) *Structures of Social Action: Studies in Conversation Analysis.* Cambridge: Cambridge University Press, 57–101.
Porter AL and Rossini FA (1985) Peer review of interdisciplinary research proposals. *Science, Technology, & Human Values*10(3): 33–38.
Research Evaluation (2006) Special issue on interdisciplinary research assessment. *Research Evaluation* 15(1).
Roy R (1985) Funding science: The real defects of peer review and an alternative to it. *Science, Technology, & Human Values* 10(3):73–81.
Russell MG (1983) Peer review in interdisciplinary research: Flexibility and responsiveness. In: Epton SR, Payne RL and Pearson AW (eds) *Managing Interdisciplinary Research.* New York: John Wiley & Sons, 184–202.
Shimada K, Akagi M, Kazamaki T and Kobayashi S (2007) Designing a proposal review process to facilitate interdisciplinary research. *Research Evaluation* 16(1): 13–21.
Thorngate W, Dawes RM and Foddy M (2009) *Judging Merit.* New York: Psychology Press.
Travis GDL and Collins HM (1991) New light on old boys: Cognitive and institutional particularism in the peer review system. *Science, Technology, & Human Values* 16(3): 322–341.
Weinberg AM (1962) Criteria for scientific choice. *Minerva* 1(2): 158–171.
Whitley R (1984) *The Intellectual and Social Organization of the Sciences.* Oxford: Clarendon Press.

Biographical note

Katri Huutoniemi is a doctoral candidate in environmental policy at the University of Helsinki, Finland. Her dissertation title is *Interdisciplinary Accountability in Research Evaluation: Prospects for Quality Control across Disciplinary Boundaries*. Her publications include: 'Analyzing interdisciplinarity: typology and indicators', *Research Policy* 39 (2010); 'Evaluating interdisciplinary research', in *The Oxford Handbook of Interdisciplinarity* (Oxford University Press, 2010); and 'Comparing customary rules of fairness: Evaluative practices in various types of peer review panels', in *Social Knowledge in the Making* (University of Chicago Press, 2011).