# The predictive validity of peer review: a selective review of the judgmental forecasting qualities of peers, and implications for innovation in science

# The predictive validity of peer review: A selective review of the judgmental forecasting qualities of peers, and implications for innovation in science

International Journal of Forecasting, accepted 27 January 2010

Wim G. G. Benda and Tim C. E. Engels

First author:

Wim G. G. Benda

University of Antwerp

Centre for R&D Monitoring

Middelheimlaan 1

B-2020 Antwerp

Tel.: +32 3265 3011

Fax: +32 3265 3010

E-mail: wim.benda@telenet.be

Second and corresponding author:

Dr. Tim C. E. Engels

(1) University of Antwerp, Centre for R&D Monitoring, Middelheimlaan 1, B-2020 Antwerp

(2) Antwerp Maritime Academy, Department of Social Sciences and Languages, Noordkasteel Oost 6, B-2030 Antwerp

Tel.: +32 3265 3034

Fax: +32 3265 3010

E-mail: tim.engels@ua.ac.be

Abstract

In this review we investigate what the available data on the predictive validity of peer review can add to our understanding of judgmental forecasting. We found that peer review attests to the relative success of judgmental forecasting by experts. Both manuscript and group-based peer review allow, on average, for accurate decisions to be made. However, there exists tension between peer review and innovative ideas, even though the latter underlie scientific advance. This points to the dangers of biases and preconceptions in judgments. Therefore, we formulate two proposals for enhancing the likelihoods of innovative work.

Keywords:
Advice taking; cognitive bias; decision-making; expert advice; group decision making; reliability;

## 1. Introduction

Peer review is quintessential in science. As the outcome of a peer review process cannot be based solely on historical data, some form of judgmental assessment is implied. Ultimately, this assessment comes down to forecasting the impact of the work under review. In the sciences, 'impact' can be measured via citations. Therefore, knowledge about the daily practice of peer review can be informative in regard to the process of judgmental forecasting and vice versa.

In this review we investigate what the available data on the predictive validity of peer review can add to our understanding of judgmental forecasting. First, we briefly introduce peer review, the major contexts in which it is used, and the decision-making process involved in it. We then explain how this decision-making inevitably implies prediction of impact and how impact in the sciences can be measured via citations, their limitations notwithstanding. The review part of this article focuses on (1) manuscript peer review, its reliability and its predictive validity; (2) group-based peer review and its predictive validity; and (3) the tension between peer review and innovation. Next, we present two proposals to enhance the chances of innovative work. We conclude with a discussion about the potential of the study of peer review processes for judgmental forecasting research.

## 2. Peer review

Peer review is the practice of having knowledgeable colleagues judge the ideas and findings of a scientist. In this article we focus on two types of peer review processes. One is the review of manuscripts as it is generally practiced in

academia: two or more peers review a manuscript upon request from the editor or editorial team of the journal to which the manuscript was submitted. In most cases the identities of the reviewers remain unknown to the author(s) of the manuscript. The peers are invited to review the scientific quality of the work and to judge its suitability for publication in the journal. Almost no research, however, has been devoted to the mechanisms of this manuscript refereeing process (Bornmann & Daniel, in press; Suls & Martin, 2009). The second type of peer review process involves committee meetings typically aimed at allocation of research grants or fellowships. Although some qualitative studies have focused on the processes involved in such group-based peer review (Langfeldt, 2001; Langfeldt, 2004; Obrecht, Tibelius, & D'Aloisio, 2007), this type of peer review process remains seldom studied.

An analysis of both types of peer review processes reveals that they are clearly relevant for scholars studying judgment, decision-making, and forecasting. In the manuscript peer review process at least four distinct judgment and decision-making steps are identifiable. First, the (associate) editor or editorial team receiving the manuscript decides whether or not to send the manuscript for review (so-called triaging). Second, the editor decides who to invite to review the manuscript. Third, after the manuscript has been sent for review, the referees will assess its quality and importance and decide whether it is acceptable for publication immediately, after revision or not at all. Fourth, the editor integrates the advice of the referees and decides on the paper's acceptability. In theory, the peer review cycle then stops for rejected papers, unless the authors decide to protest the decision. In practice, many rejected papers will then go through all four steps again, as authors often opt to target another journal. For papers that are accepted pending revisions, the second,

third, and fourth steps will often be repeated, depending on the substance of the revisions required. Of course, this prototypical process comes in many variations, sometimes consciously introduced (e.g. when journals collaborate to pass on reviews, Saper & Maunsell, 2009) or resulting from tradition (e.g. communal decision-making among editorial board members of journals in the humanities).

Group-based peer review processes are even more diverse. We focus on committees that award research grants or fellowships. Committees will generally have an initial discussion, during which general impressions are exchanged and which may or may not involve triaging. A division of tasks is then provided, either by sending proposals for external review or by dividing the work amongst committee members. Sometimes the chair alone assumes both tasks. Third, the committee members will, regardless of whether external review has occurred, form their own judgment. Occasionally this judgment is an explicit part of the procedure, as committee members must submit judgment prior to a committee meeting; often, however, it is implicit as committees simply meet to discuss and decide. Fourth, a group-based peer review process ends with a decision, for example to award a fellowship, to offer funding, etc. As noted, variations of this prototypical process are manifold. For example, committees may meet more often so as to facilitate interviews with candidates or to comply with a two-step funding application procedure.

Clearly, both manuscript and group-based peer review processes are relevant to scholars of decision-making. Of interest are the judgment and decision-making itself, as well as how advice is gained and handled, the procedures followed and, especially, the forecasting implied. Let us explain the latter statement. In the case of manuscript peer review, the forecast concerns the potential future impact on the research community. Journals, especially the

most reputable within a particular area, require that their papers not only show high quality and address relevant and important topics, but also evidence potential for future impact on the field. Publishing influential papers is, after all, the gateway to establishing or reinforcing a journal's importance (Starbuck, 2005). Of course, editors do not ask for exact forecasting of a paper's impact or for prediction of a citation interval, but they request that reviewers indicate why they deem a paper potentially influential and how it might influence the field. In group-based peer review processes, the forecasting aspect is often more explicit, as committees are asked to judge the scientific quality, productivity, relevance and/or future impact of the work of candidates or consortia. In principle, committees working for funding agencies aim at providing opportunities to applicants with the highest potentials. In sum, one might say that peer review is all about the future.

How then can the accuracy of the forecasts that are inherent in peer review be measured? In science, citation counts are widely used to assess impact. Can citation counts shed light upon the validity of the decisions resulting from peer review processes[i]? As citation counts are, to say the least, not without controversy, many a scientist will instinctively provide a negative answer. Among sociologists of science, two basic positions are distinguishable (for a detailed discussion, see Bornmann & Daniel, 2008b). One holds that citations indicate that an author has used the works cited. Hence, in this normative view, which builds on Robert K. Merton's sociological theory of science, citations represent intellectual or cognitive influence on scientific work. The other position, however, holds that scientific knowledge is socially constructed, and thus doubts the validity of citations as a measure of scientific impact. In this view, the central problem is the lack of clear norms and conventions for using citations. As a

result, citations, rather than measuring scientific impact, may instead serve a persuasive function.

Normative and social constructionist citation theories have both been empirically tested (Bornmann & Daniel, 2008b). Adherents of normative theory argue that citations correlate with other indicators of scientists' impact, such as peer judgements, research funding, departmental prestige, awards and honors. Social constructionists, however, point to the dependence of citations on characteristics of the particular scientific field, journal, article, author, readership and/or timing of publication. Based on available evidence, it can be concluded that citations can be a valid measure of impact if the level of aggregation at which they are counted is sufficiently high (van Raan, 2004a). It has been repeatedly observed that, at the level of the individual (paper), citation counts are unreliable measures (e.g. Lehmann, Jackson, & Lautrup, 2006). At an aggregate level, however, citation counts provide insight into how a body of research impacts the field, for example, if the research has been published by a particular journal, funded by a certain agency, or pursued by a certain department. Hence, in this review of the judgmental forecasting qualities of peers, we have included studies that use citation counts.

Before we continue, we should draw attention to the possibility of contemplating alternatives to peer review. Indeed, it is far from certain that peer review is the best way to advance science. As we will see, many scientists have observed that the peer review system is anything but perfect, and some have concluded that abolishing peer review would be advisable (Abrams, 1991; Horrobin, 1996). In fact, alternatives and improvements to peer review are permanently on the agenda of scientists (e.g. Diener, 2009), sociologists of science (e.g. Moed, 2007), and funding agencies (e.g. the National Institutes of

Health, see http://enhancing-peer-review.nih.gov/). A full discussion, however, is beyond the scope of this review, as our primary interest lies with the decision-making in peer review processes.

## 3. Manuscript peer review

In this section we discuss the reliability, internal validity, and external validity of referee reports. We then present findings on the predictive validity of editorial decisions.

### *3.1. Reliability, internal validity, and external validity of referee reports*

> *The role of the editor takes on its most compelling meaning when the editor must use two clearly conflicting reviews to decide the disposition of the manuscript. At that point, the discipline begins to advance or to retreat.*
>
> Glidewell (1988, p.769)

The most common measure of the reliability of manuscript peer review at any particular moment is inter-referee agreement. Cicchetti (1991), however, in a comprehensive analyses of the reliability of journal peer review, concludes that reviewer agreement is very low. Weller (2001, p.181-200), based on a study of the relevant literature, concludes likewise: practically every study, across a range of journals, indicates that the levels of inter-referee agreement, when corrected for chance, fall in the range of 0.20 and 0.40, corresponding to a low level of reviewer agreement. For many scholars, the failure to achieve acceptable levels of agreement is the most basic and broadly supported criticism of peer

review (Jayasinghe, Marsh, & Bond, 2006). Small experiments have confirmed this lack of reliability in peer review (e.g. Ernst, Saradeth, & Resch, 1993; Peters & Ceci, 1982). Interestingly, reviewers are twice as likely to agree on which scientific documents to reject than on which to accept (Weller, 2001, p.193). This finding is reassuring, in a way, since it is the truly unsound or poor research that should not be published. However, the level of agreement among reviewers is generally not much higher than can be expected to occur based on chance alone.

But how important is inter-referee agreement? Is it a relevant criterion for judging the manuscript peer review process? In our opinion, inter-referee agreement is not all-important: it is a statistical criterion that presumes the existence of a single exact – or at least a prototypical – way to arrive at scientific knowledge and to present it in writing that referees use as the template in reviewing a manuscript. However, such a unique path to universal truth does not exist (Fara, 2009, p.363). Reviewing is a human activity. One consequence is that the small sample of referees will influence the outcome of the review process. Reviewers' personal traits are also a factor, since reviewers vary enormously in their attitudes, practices, and viewpoints (Graue, 2006). Some reviewers are naturally more lenient, whereas others are harsher (Jayasinghe et al., 2006; Siegelman, 1991). In an editorial study on correlations of ratings on four dimensions of reviewed manuscripts, Glidewell (1988) found that the correlations between ratings by two reviewers on the same dimension were consistently lower than the correlations between the ratings by the same reviewer on different dimensions, hinting at the influence of personal traits. Also notable are differences in worldview and scientific school, which inevitably influence a referee's judgment (Kostoff, 1995). According to Bedeian (2004), two referees commissioned to review the same manuscript actually read different

works because each constructs a unique interpretation of its content. For authors the implication is clear: one needs a bit of luck when reviewers are assigned.

Instead of focusing solely on reliability as measured by inter-referee agreement, we wish to draw attention to the relevance of review reports. Specifically, along the lines of constructionist science philosophy and its implications for the evaluation of findings (Engels & Kennedy, 2007), we consider the credibility (internal validity) and the applicability (external validity) of referee reports the more important criterion. The credibility of a referee report refers to the degree to which it contains information indicating that the reviewer is indeed knowledgeable about the topic and has spared no effort to offer advice. One implication is that diversity, rather than level of agreement, of advice provided by reviewers is pivotal to the credibility of such advice. Indeed, some editors willingly select reviewers with different areas of expertise because this can be helpful to the editorial team and to the authors. Also, reviewers often comment on different aspects of a paper instead of really disagreeing (Fiske & Fogg, 1990). The reviewers can offer valuable points from different backgrounds and therefore reach different conclusions about manuscript acceptance and rejection. In short, although low inter-referee agreement may diminish the reliability of peer review, it can also raise the internal validity thereof.

Likewise, the external validity of referee reports can benefit from inviting reviewers from different backgrounds. After all, peer review intends to help authors improve their manuscripts and to support editors (or editorial teams) in their decision-making. Hence, peer advice should be applicable in the sense that it is useful to authors and helpful to editors. Indeed, according to surveys of both readers and authors in medicine and psychology, the quality of manuscripts increased after peer review (Bradley, 1981; Goodman, Berlin, Fletcher, &

Fletcher, 1994; Nickerson, 2005; Pierie, Walvoort, & Overbeke, 1996). Moreover, the length of reviewers' comments appears to be positively correlated with post publication citations (Laband, 1990). At the same time, however, many authors attribute at least some of the changes requested by the editors and/or reviewers to subjective personal preferences, bias and even whim (Bradley, 1981). The latter finding points to the ambiguous relation that authors have with peer review. Evidently, authors of accepted manuscripts view the review process with somewhat more satisfaction than authors of rejected manuscripts (Nickerson, 2005; Sweitzer & Cullen, 1994). Authors prefer that reviewers provide "specifics regarding problems they see and, when feasible, concrete suggestions for fixing them and for otherwise improving the presentation" (Nickerson, 2005, p.662).

Editorial decision-making is more complex than simply counting votes, and diverse reviewer comments can facilitate and improve it. Editors frequently solicit more reviews when confronted with clearly divergent reviewer opinions. Alternatively, they may resolve the disagreement by themselves, seek input from associate editors, or discuss subsequent steps at editorial meetings (Weller, 2001, p.196). What is crucial, however, is that editors consider the content of disagreeing reviewers' reports so as to arrive at balanced decisions. If the referees' views are counted as votes for deciding whether a submission should be accepted, then lack of agreement among reviewers is merely a liability (Fletcher & Fletcher, 2003). Unfortunately, such vote counting systems appear to be in place all too often. Armstrong (1997) observes that the most prestigious journals, which have high supplies of papers, are mostly filled with papers that have only positive reviews, a result that has been confirmed for journals in different fields (Bakanic, McPhail, & Simon, 1990; Bornmann & Daniel, in press; Kupfersmid & Wonderly, 1994, p.56). Editors may also be concerned about being

fair and thereby invoke vote counting as a transparent system. However, in order to benefit from the diversity of reviewers' input, editors must resist the temptations of an ostensibly objective system and instead weigh the evidence with a view towards pointing authors in a particular direction. An editor with knowledge of reviewers' personal traits and scientific schools can make an informed decision about a paper's acceptability and quality and assess its potential impact. For an editor, votes to reject by lenient "zealots" and to accept by harsh "assassins" can be particularly decisive (Siegelman, 1991). In sum, editors, when faced with sharp disagreement, should ask themselves: "How can I explain this disagreement?" and "What kind of work must this be, if these two discerning reviewers disagree so sharply?" (Glidewell, 1988, p.766)

We conclude that reviewers' judgments of manuscripts differ rather often. But *that* reviewers hold diverse opinions should not be a problem. Of course, for both authors and editors it is important to know *why* they differ in opinion. If editors are prepared to weigh the recommendations of individual reviewers, then diversity of opinions enhances the internal and external validity of review reports. Clearly, the manuscript peer review process consists of much more than just refereeing by peers. One could argue that editorial decision-making, including triaging, choosing referees and deciding on acceptance/rejection, is in fact more influential, especially since this decision-making is generally done by only one or a few people. In the next section we investigate whether editors are indeed able to select the best articles, that is, to accurately assess the future impact of the submissions to their journals.

### *3.2. Predictive validity of editorial decisions*

*The most important question is how accurately the peer review system predicts the longer-term judgments of the scientific community. One way to address this would be through citation data; articles that stand the test of time should be highly cited relative to others in the same field, even several years after their publication.*

Charles G. Jennings (2006)

As postulated in the introduction, citation counts can be considered good indicators of the impact of a body of research. It follows that, if peer review is a good method of selecting the best articles and thus of forecasting their impact, the articles published by leading journals should on average receive more citations than the articles rejected by the same journals but subsequently published elsewhere. Such comparison is possible, as more than half of manuscripts rejected on initial submission are later published elsewhere (Weller, 2001, p.64). This allows for comparing the average and median numbers of citations of submissions accepted for publication with that of manuscripts rejected by a specific journal but published elsewhere. The few such studies that have been done all established similar results, viz. a rather high degree of predictive validity of peer review and subsequent editorial decisions.

The first study of this type concerned submissions to the *Journal of Clinical Investigation* (Wilson, 1978). Papers published in this journal were cited approximately twice as often during the first four years after publication than were papers rejected by the journal but published elsewhere, thus leading to the conclusion that the peer review system was effective in separating high- from low-impact papers. Similar results were found for submissions to *Angewandte Chemie* (Daniel, 1993), *Cardiovascular Research* (Opthof, Furstner, van Geer, &

Coronel, 2000), and *American Journal of Neuroradiology* (McDonals, Cloft, & Kallmes, 2009). In a further study of submissions to *Angewandte Chemie*, the average citation counts of accepted and rejected manuscripts were compared to subfield specific averages (i.e. baseline values, Bornmann & Daniel, 2008a). The results provided further evidence for the hypothesised high predictive validity of peer review and editorial decision-making.

A possible critique of these studies is that not only do the citations received by the initially accepted papers contribute to the relative high impact factors of the journals in which the papers appear but that the papers receive citations precisely because they have been published in reputable journals. In other words: the prestige of journals may be a confounding variable in researching the predictive validity of their editorial decision-making. Indeed, if articles receive more citations because they appear in higher-prestige journals and journals gain prestige because they publish articles that receive more citations, then feedback conditions for self-fulfilling prophecy are in place (Starbuck, 2005).

One empirical study in the biomedical field found that article citedness and a journal's impact factor are poorly correlated because the distribution of article citedness is highly skewed, even for individual authors and within defined journal impact cohorts (Seglen, 1994). The author concluded that the citation rates of articles do not seem to be detectably influenced by the status of the journals in which they are published. Leimu and Koricheva (2005) draw similar conclusions via analysing data for ecology papers and journals. But other researchers have concluded exactly the opposite: that the average citation rate of the journal that published a particular paper is in fact the best predictor of the paper's citedness (Callaham, Wears, & Weber, 2002; Judge, Cable, Colbert, & Rynes, 2007). In

other words, the likelihood of results being overlooked by the scholarly community is negligible when they are published in journals with high relative impact factors, whereas the likelihood for a relatively high citation rate increases almost six-fold for papers published in the most prestigious journals (Racki, 2009). This assertion coincides with the intuition of Eugene Garfield, the founder of the Web of Science, who believes that "the extent of a paper's "citedness" (...) is fairly predictable. If it's published in a high-impact journal, it is highly likely to be cited. If it's published in a lower-impact periodical, it may remain uncited – even if it received high marks in prepublication peer review or is frequently read." (Garfield, 1991, p.390). Based on the available evidence, it appears likely that citedness is indeed influenced by journal prestige.

Nonetheless, it is also the case that publications in prestigious journals matter greatly to authors. Hence it is just as plausible that leading journals have first choice of articles and can therefore select the best papers. Although this may be true in a general sense, prestigious journals also tend to publish many articles that are rarely cited (Starbuck, 2005). Furthermore, less prestigious journals publish excellent, highly cited articles, and it is common for highly cited articles to have been rejected from multiple journals. So although high impact papers often appear in leading journals, the journals' editorial decision-making can result in type I and type II errors (Bornmann & Daniel, 2009). In the context of editorial decision-making, a type I error involves publication of a manuscript that is later cited as frequently or less frequently than the median citation count of the manuscripts rejected by the journal; a type II error concerns rejection of a manuscript that, after publication in an other journal, is as frequently or more frequently cited than the median citation count of the manuscripts accepted for publication in the former journal. Indeed, critics of peer review offer various

examples of eyebrow-raising editorial decisions regarding later classics. There are such examples in practically every field, ranging from difficulties in having findings published and having findings temporarily neglected to facing long delays in having major findings published (Campanario, 1993; Campanario, 1998; Gans & Shepherd, 1994).

When reading about such examples, one wonders if the predictive validity of editorial decision-making can be better than mediocre. However, the case presented by critics of peer review relies mostly on anecdotal evidence (for an exception, see Gottfredson, 1978). Some of these anecdotes refer to articles that go unnoticed for years and then suddenly attract many citations. But such "Sleeping Beauties" are in fact extremely rare (van Raan, 2004b). Indeed, the strength of peer review's supporters is the statistical, rather than anecdotic, foundation of their argument.

We conclude that the predictive validity of editorial decision-making is reasonable. Although there probably exists a halo effect for articles published in prestigious journals, this effect cannot account for high or even moderate amounts of citations. Nonetheless, some initially rejected articles receive many citations, thus testifying to incorrect editorial decision-making and flawed peer judgment.

## 4. Group-based peer review

> *Every scientific institution that uses peer review has to deal with the following question: Does the peer review system implemented by my institution fulfil its declared objective to select the best scientific work?*
>
> Bornmann and Daniel (2006, p.428)

In this section we address the reliability and validity of group-based peer review. Specifically, we review studies concerned with peer review of award, grant or fellowships applications. We limit our scope to this type of committee peer review because it generally involves committee members affiliated with different universities and mainly concerns the scientific quality, productivity, relevance and future impact of the proposals or works submitted. In contrast, committees concerned with recruitment, tenure and promotion involve at least a majority of members belonging to one university or institute, and must take into account several non-scientific considerations in their decisions (e.g. institute policy or candidate demands).

A limited number of research studies qualify for discussion in this section. Few studies have experimentally addressed the issue of reliability of peer review for grant applications, and these typically lead to the conclusion that the process's reliability is low and open for improvement (Cicchetti, 1991; Hodgson, 1995; Jayasinghe et al., 2006; Marsh, Jayasinghe, & Bond, 2008). As in manuscript peer review, however, diversity of reviewers' opinions need not be a liability and may facilitate decision-making (Langfeldt, 2001). Indeed, low inter-reviewer agreement may indicate that the committee and its assessments are highly competent because the committee represents multiple views of what constitutes good research.

Most studies regarding group-based peer review investigate the correlation between committee decisions and the beneficiates' academic output, as measured by publications and citations. Nederhof and van Raan's pioneering work (1987; 1989) compared recipients of *cum laude* doctorates with recipients of ordinary doctorates in physics and chemistry, respectively. The authors

concluded that the *cum laude* recipients produced substantially more publications than did their counterparts, both before and after being singled out as exceptionally promising researchers, and their publications received more citations. Hence the doctoral committees' decisions to award *cum laude* designations appeared valid from the perspectives of past and future performance.

Other studies have worked along these lines and investigated the predictive validity of committee peer review in awarding distinctions or fellowships. For example, Mavis and Katz (2003) found that the successful applicants to the March of Dimes Birth Defects Foundation published significantly more peer-reviewed papers than did unsuccessful applicants during the study's ten-year follow-up window. The successful applicants also received more citations, were more likely to receive federal grant funding, and outnumbered unsuccessful applicants in securing positions at top-ranked institutions. Similarly, Bornmann and Daniel (2005; 2006) analysed the bibliometric performance of approved and rejected applicants for Boehringer Ingelheim Fonds (BIF) fellowships. They observed that, both prior to and after allocation of the fellowships, articles by the approved applicants were systematically cited more frequently than articles by the rejected applicants. Moreover, the articles by the approved and the rejected applicants were cited considerably more often than the average publication in the chosen journal sets.

Nevertheless, we should be cautious about jumping to general conclusions regarding the predictive validity of group-based peer review. The aforementioned studies analyse average values that can be strongly influenced by individuals (or articles) at the extreme ends of the distribution. For a more accurate picture, we must investigate the frequency of type I and type II errors. In the context of

awarding fellowships, a type I error is overestimation of an applicant, meaning that an applicant was funded but subsequently performed under the median of the rejected applicants; a type II error is underestimation of an applicant, meaning that an applicant was not funded but (s)he subsequently performed above the median of the accepted applicants. Reassuringly, at least in biomedicine, type I errors appear to be infrequent. However, a type II error occurred in decisions regarding approximately one in three rejected applicants (Bornmann, Wallon, & Ledin, 2008). Although limited funding availability may also be to blame, this relatively high level of type II errors illustrates that experts cannot always recognise and reward the best.

Other studies attest to this. Van den Besselaar and Leydesdorff (2009) examined the Dutch Research Council and found that, in the social and behavioural sciences, the bibliometric indicators for past performance of successful grant applicants were significantly higher than those of unsuccessful applicants. But when compared with the successful applicants, the best non-funded applicants had significantly higher scores on past performance than those who received funding! The authors conclude that the council succeeded in identifying and discarding the least merited applications, but not in selecting the "cream of the crop". Similarly, Hornbostel and colleagues (2009) have shown that successful and unsuccessful applicants for a highly demanding and prestigious German funding scheme have highly similar profiles in past performances and in performances subsequent to receiving the grant. On some indicators the rejected group outperformed the approved one; however, the approved applicants were more successful in securing professorships. Melin and Danell (2006) examined a Swedish programme and found no important differences between the 20 applicants that were rejected and the 20 that were

approved (all 40 applicants had been selected for further review from 500 initial applicants). In fact, the existing minor differences favoured the rejected group. But after approval, the situation changed: the approved group did significantly better, attracting more funding and generating far more patents. The latter two studies imply that peer review committees that award grants or fellowships can create differentiation among individuals who perform similarly at the time of application. Hence committees do more than merely select candidates: they also, to a certain extent, shape candidates' futures, putting some scientists at an advantage. A Matthew effect (allowing those at an advantage to gain more reputation and resources) may ensue (Langfeldt, 2006). Moreover, the apparent predictive validity of the peer review process may prove to be a self-fulfilling prophecy.

In sum, committees of peers appear to mostly succeed in "discarding the tail" and selecting groups of excellent researchers. Nevertheless, type II errors occur rather often, undermining the predictive validity of peer review. Moreover, researchers that are awarded funds begin to attract more opportunities; the outcome of the selection process actually helps them become or remain the best. This raises suspicion that the predictive validity of committee peer review is a self-fulfilling prophecy.

## 5. The tension between peer review and innovation

*The more innovative and interesting the paper, the more likely it is to be rejected, in my experience.*

Graciela Chichilnisky (cited in Gans & Shepherd, 1994, p.177)

Contemporary science and philosophy of science generally accept that the process to reach a result influences the result itself. In this sense, peer review is a socially constructed process that helps define scientific output and advancement. But within this context, most statistical studies show a generally positive correlation between peer review and bibliometric indicators. Peer review can usually select the best articles and the best researchers, or at least "discard the tail". Nevertheless, there exists a significant margin of error, especially type II errors, some of which become anecdotes that illustrate (supposedly) bias in peer review. Can these errors help us understand the small group judgemental forecasting process that creates them?

There exists abundant literature – and vivid discussion – on fairness and forms of exclusion and bias in peer review (e.g. Campanario, 1998; Marsh & Bornmann, 2009; Weller, 2001, p.207-246). Subjects of possible bias include: academic status of the scientist or institution, ethnicity, gender, ideology or scientific school. Although such biases can be pernicious to those that face them, they do not bare direct relation to scientific excellence and future citations; they only decide to what extent certain categories of people are admitted to the (core of a) discipline. One bias, however, does have a strong relation with excellence and citations: the bias against innovative ideas.

High quality and excellence are intertwined with new ideas and methods. Commenting on clinical science, Horrobin (1996) notes that quality research should be truly innovative and present ideas that will be regarded as important 50 years hence. In basic science, quality research must lead to genuine new understanding and eventually to clinical advance. In the clinical field, real innovation leads directly to improvements in patient care. Horrobin's view is easily generalised to other disciplines. Innovation is what leads to progress.

In this sense, it is discomforting to learn that many genuinely new ideas experienced difficulties in reaching wider audiences (Campanario, 1993; Campanario, 1998; Gans & Shepherd, 1994). The history of science is replete with examples of innovations that were recognised only years later. Examples include the belated recognitions of Mendel's work in genetics and Mayer's discovery of the first law of thermodynamics (Armstrong, 1982a) and the delayed acceptance of the Australopithecus as a human ancestor (Gould, 1977, p.207-213). Moreover, a common complaint is that peer review leads funding agencies to act too conservative, thus making it difficult to obtain funds for innovative projects (Campanario & Acedo, 2007).

So what is it that makes peers reject innovative findings and ideas? As innovation is a central tenet of science, it is hard to believe that there is intentional bias against it. This bias, although highly subtle, has been analysed from different angles. Horrobin (1996) refers to an interplay of competing interests of peers as scientists and scarce resources. Peers are specialists in the field and therefore compete as applicants. Moreover, in a context of scarce resources, innovative and thus risky applications are likely to arouse opposition from administrators, reviewers and committee members. Consequently, such applications often encounter more difficulties obtaining funds.

Competing interests and resource scarcity, however, are not themselves sufficient barriers to innovative ideas: they merely create the context in which decisions must be made. Within this context, certain psychological traits enter into play, namely, conservative bias towards established ideas and personal viewpoints. Several psychological studies confirm this tendency (Mahoney, 1977; Weller, 2001, p.223-224). These studies report that reviewers are strongly biased against papers which present results contrary to their beliefs. For

example, Armstrong and Hubbard (1991) found, in a survey of editors of 16 leading American Psychological Association journals, that empirical manuscripts with controversial findings were treated more harshly. And via observation of 10 meetings of grant-awarding committees, Travis and Collins (1991) revealed that committee members sometimes make decisions based upon their adherence to scientific schools of thought, thereby demonstrating what the authors label "cognitive similarity". On the basis of such observations, Travis and Collins pointedly identify that science is a social system and that there is a cognitive view of science. The social system exhibits similarities in social positions, possibly leading to bias against certain categories of people (gender, institution, etc.). The cognitive view of science is based on similar views and thus may cause bias against new ideas.

Of course, this is reminiscent of Kuhn's theory of scientific paradigms. According to Kuhn's philosophy of science, a set of views and practices guide science at a certain moment. This is a paradigm. Scientists are normally not prone to developing new theories and will even suppress new theories by other scientists because such theories undermine the foundations from which scientists commence their everyday practice. Scientists neglect disconfirming evidence and anomalies. However, when anomalies continue to accumulate, new theories suddenly have more opportunities to challenge existing paradigms. When a new theory prevails, Kuhn speaks of a scientific revolution or paradigm shift. But for this to happen, scientists will likely first encounter serious resistance from their peers.

Obviously, not every innovation initiates a paradigm shift. But what can be said at the macro level about grand paradigms also applies at the micro level. Established beliefs and practices are rather tenacious in science. For example, in

psychology and human resources, the so-called Hawthorne effect remains a major subjects in textbooks, although later research has blunted Elton Mayo's argument (Armstrong, 1982b; Levitt & List, 2009). The original idea is so persistent because it coincides with existing beliefs about human relations and productivity. Attempting to have articles published that challenge such management folklore may prove a real Calvary (Armstrong, 1996).

In sum, preconceived ideas lead all too often to disconfirming evidence being swept aside. Real, qualitative development often happens through sudden leaps; in science, through innovative ideas and technologies. But the selection process of peer review easily purveys research as merely another addition to the paradigm, that is, as the next step in a slow, gradual process of puzzle solving. Inside such a paradigm, peer review performs adequately in judging quality and forecasting success. However, predicting sudden leaps, even if they are directly in front of peer reviewers, seems to be much harder. In the next section we analyse decision-making in peer review processes with a view of addressing this hindrance to true innovative ideas.

## 6. Two proposals to enhance the chances of innovative work

Decision-making is pivotal in peer review. Although most attention is devoted to the reviewing aspect of peer review, it is the decision-making that follows the reviewing that is in fact decisive. These decisions can severely impact individual scientists, as well as journals and research groups (Lawrence, 2003). What do we know about decision-making processes in peer review? And how can these processes be improved so as to enhance the chances of innovative work?

For one thing, decision-making in manuscript peer review is not limited to the final decision to accept or reject a paper for publication. Prior to that stage the editor, or the editorial team, has already decided to consider the paper for publication (by sending it for review) and has invited referees to review it. Little is known about these and other intermediate decision-making steps, except perhaps for the observation that desk rejections appear to be on the rise (Suls & Martin, 2009) and that some editors deliberately invite referees with different backgrounds, or at least consider divergent input interesting (Glidewell, 1988; Straub, 2008). Regarding the final decision-making to accept or reject a paper for publication, one of the few things we know is that many journals, especially the leading ones, adhere to a "one negative review and you are out" policy. Strikingly, this implies that many editors do in fact pass on their decision-making power to the reviewers! We suggest that this be done in another, more positive way. Let us explain.

The system of vote counting testifies to preoccupation with avoiding type I errors. In a way, such preoccupation is understandable, as publication of sub-standard or even faulty papers can be detrimental to a journal's reputation. However, vote counting is not the best possible system, since it is not type I errors, but rather type II errors (rejecting papers that in fact surpass the standards of a journal) that are problematic for the advancement of science. And type II errors are certainly not so infrequent that one need not worry about them (Bornmann & Daniel, 2009; Starbuck, 2005). In many disciplines, a possible explanation for type II errors is preoccupation with methodology instead of intellectual novelty (Straub, 2008). Even editors rank methodological problems at the top of their lists as reason for rejection (Weller, 2001, p.54). Yet it may happen that a submission offers refreshing ways of looking at an age-old

problem, but that the scientist's method of addressing the problem need (much) improvement, as is often the case at the beginning of a new paradigm. In the current system, such a submission faces probable rejection, not least because, being a good reviewer, one wants to avoid creating an impression of not being up to the task. So for the sake of their own reputations, reviewers will highlight apparent weaknesses. In fact, reviewers, at least in cases of innovative papers, face situations of loss aversion (Tversky & Kahneman, 1991), as they stand to lose more (their reputations) from not indicating weaknesses then to gain (advancing science). Yet why would a reviewer who notes weaknesses but knows that any recommendation to reject will likely result in rejection advise thusly? Precisely because reviewers are likely to be in a screen-out mode they will attend more to a manuscript's negative features than to its positive features (Shafir, 1993). Hence at least one recommendation to reject becomes highly likely whenever an innovative paper, which typically disposes of both more negative and more positive features than does the average paper, is reviewed. The result for many journals is a status quo bias (Samuelson & Zeckhauser, 1988): editors will prefer not to publish such papers, rather than risk losses or gains whenever a reviewer recommends rejection.

Given these findings, is it possible to imagine a manuscript review process that would benefit innovative submissions and limit type II errors? Our suggestion is to provide regular reviewers with a limited number of decisive votes 'pro' instead of only votes 'contra'. In this system, regular reviewers - for example, the members of the editorial board - would have not only the right to voice opinion, but the right to decide, for a limited number of papers (e.g. one in ten reviews), in favor of publication. In this way, although another reviewer may

be critical or even explicitly negative about a submitted article, controversial and innovative findings would reach the best journals faster and more often.

In group-based peer review, decision-making mechanisms that slow or even block innovative ideas are also at work. Committees are often confronted with limited resources that force them to be severe. In many cases the approval rate for grants or fellowships hovers around 20% or less, paving the way for incorrect rejections (Bornmann et al., 2008). Moreover, committee members wish to avoid type I errors (funding people who perform sub-standard) and may therefore side with the least uncertain proposals in order to 'play safe'. The most innovative proposals may seem far riskier than mainstream research, and thus face disadvantages in trying to secure funding under circumstances of high rejection rates. The review system itself may emphasize conservative criteria, such as researchers' track records and proposals' feasibility (Langfeldt, 2006). In other words, the rather high frequency of type II errors is caused by interplay between scarce resources, preconceived ideas, and fixation on reducing risk (cf. established methodology).

But how do decisions in group-based peer review come about? Although grant schemes may have almost identical aims, their review and decision-making procedures vary widely. And procedural differences have significant implications for the outcomes of the review. Langfeldt (2001) investigated how seemingly irrelevant factors such as rating scales and peer panels' ranking methods influenced the kinds of projects funded. Langfeldt found that a rough-rating scale (e.g. with only three categories: 'fundable', 'fundable with alterations', and 'not fundable') enhances the chances of innovative research, while fine-rating scales with several categories strengthen established research. Moreover, ample budgets favour controversial projects, and tight budgets tend towards more

conservative outcome. In addition to budgets and rating scales, the decision-making process itself is crucial in deciding which projects are awarded. Langfeldt demonstrated how three different methods lead to very different outcomes. Established research is secured when panel members 'eliminate' candidates via majority voting or when top candidates are compiled via comparison of the panel members' respective ranking orders. The chances for original research are enhanced when all members propose one candidate for funding. So those review models that perform strongly on thoroughness (fine-rating scales) and reliability (agreement) underperform with regard to encouraging controversial projects, and vice versa. Again, when votes are counted, innovative ideas are at a disadvantage.

Furthermore, panels tend to develop their own rules and culture. Different panels emphasize different evaluation criteria, even when the funding organisation provides them with identical guidelines (Langfeldt, 2001). Obrecht and colleagues (2007) observed that new committee members quickly acquire the culture of their committee, for example through adapting their rating behaviour. Building on the work of Kerr, MacCoun and Kramer (1996) regarding exacerbation of individual biases in group discussions, they infer that committee culture can strongly impact the review outcome. Hodgson (1995) likewise found that scores of applications vary significantly depending on which committee they are assigned to. Furthermore, Hodgson observed that the final committee discussion contributed significantly to the final score, leading her to conclude that the dynamics of face-to-face committee meetings make considerable contribution to the peer review process.

Even when committee members provide preliminary rating of applications before they meet, the ensuing group discussion and decision-making may

decrease fairness. For example, when funds are scarce, committee members adjust their rating behaviours and tend to side with the most negative reviewer (Obrecht et al., 2007). This observation can be expected, because the task of committees facing low approval rates is the difficult one of motivating[ii] the exclusion for further consideration of many an application. This will typically leave a committee with a larger set of options than when committee members need to decide what to include for further consideration (Levin, Prosansky, Heller, & Brunick, 2001). Given the requirement to screen out the large majority of applications, a focus on negative features can be expected (Shafir, 1993), and hence the most negative reviewer will garner support. In order to counter such group dynamics, Obrecht and colleagues (2007) advocate for structured reviews by separate individuals on the basis of clearly defined criteria, for example innovation and originality. Other authors propose at-home scores to decide on most applications; such scores would free time for in-depth discussion of those applications where there is significant difference of opinion (Thornley, Spence, Taylor, & Magnan, 2002).

But would it not be possible to implement a decision-making procedure that would indeed benefit innovation and originality? Again, we wish to make a proposal. Because innovative proposals involving controversial or counter-paradigmatic ideas are unlikely to gain approval from a majority of committee members, support for such proposals will be a minority position. Therefore, if funding agencies really desire to encourage innovative research, the decision-making procedure should be such that it encourages the chances of minority positions. As Kameda and Sugimori (1995) show, this can be achieved if subgroups are allowed to make decisions before subgroup decisions are combined. So, similar to our proposal for manuscript peer review, we propose

that committees yield their decision-making power for a certain percentage of proposals to sub-committees. This, we believe, would enhance the chances of truly innovative proposals.

## 7. Concluding observations

Peer review is a good example of small group judgmental forecasting. A small group of experts, acting individually or as a group, assesses the quality and future impact of a manuscript, proposal or candidate. As is apparent from our review, peer review attests to the relative success of judgmental forecasting by experts. Both manuscript and group-based peer review allow mostly for accurate decisions to be made, as has been confirmed by several studies on the predictive validity of peer review. In general, peer review does a good job in selecting quality and judging its future impact. Of course, as with every human judgment and forecast, errors occur. Type II errors are especially problematic for science. Part of these errors can be explained by an unintentional bias of peer review against innovative work.

Several of our findings are relevant to the study of small group judgmental forecasting. In fact, the debate on peer review resembles the debate on forecasting. For example, the relative advantages of quantitative techniques (bibliometric indicators) over expert judgments (peer review), and vice versa, are a subject of intense debate in the discipline of science studies and the discipline of forecasting. Many scholars, however, consider both approaches complementary (for peer review, see Moed, 2007; van Raan, 1996 - for forecasting, see Lawrence, Godwin, O'Conner, & Önkal, 2006; Wright, Lawrence, & Collopy, 1996). We summarize our main observations below.

First, peer review attests to the observation that integrating advice from multiple and independent sources benefits accuracy (Bonaccio & Dalal, 2006). Editorial boards and funding agencies are no longer able to decide on submissions without external and multiple expert advice. By making the sources of advice more distinguishable from each other, their advice becomes more helpful (Bonaccio & Dalal, 2006). Indeed, seeking reviews of referees with differing expertise – a common practice in certain fields – should be encouraged, as advice from different perspectives is more likely to provide editors and committees with relevant information. Like in Delphi applications (Rowe & Wright, 2001), experts should be chosen carefully, based on appropriate domain knowledge and so as to represent a heterogeneity of expertise and opinions.

A second observation is that the scholarly-cognitive background of reviewers is both their reviewer qualification and their cognitive bias (Graue, 2006; Langfeldt, 2006). Editors and review committees have preconceptions and are sympathetic towards findings that confirm their ideas. This scenario is reminiscent of the observation that judges give more weight to their own positions and to advisors whose preferences are similar to their own (Bonaccio & Dalal, 2006; Kerr & Tindale, 2004). Moreover, reviewers may be biased against disconfirming evidence. Obviously, such subjectivity can have serious impact on assessment of the relevance of the work under consideration. Science is the passion of a scientist and, as with predictions of politics, the desirability of an outcome can influence an expert's judgment and its accuracy. But there is more to cognitive bias than risk of inaccurate judgment. A prediction can begin to lead its own life and influence the events predicted, for example when researchers who after receiving a highly prestigious grant become even more successful in attracting research funding. This effect is potentially strong because reviewers

are also ‘players’, often influential ones, in science. Their judgments help to shape a discipline, and their cognitive biases can endanger the progression thereof. Thus, cognitive bias can make a forecast less accurate and can even become a self-fulfilling prophecy that hinders progress and innovation.

This brings us to our third observation relevant to judgmental forecasting: unconventional ideas are at a disadvantage in peer review, despite being the core of scientific advance. Sometimes unconventional thinking may be superior to majority expert opinion, but the power of the majority is not easily countered. In Delphi procedures, for example, majorities exert a strong pull on minorities to a consensual position, even if the minority position is more accurate (Rowe, Wright, & McColl, 2005). The same applies to peer review. Therefore, if editors and review committees wish to bolster innovative research, they should implement procedures that do exactly this. As we have shown, neither vote counting nor averaging is up to the task. Voting and averaging are adequate strategies to select strong, methodologically sound work along well established lines. But positive advice should receive more weight if true innovation seems at hand. This finding may be surprising because averaging as a forecasting strategy works well across a wide range of environments (Armstrong, 2001). On the other hand, judgmental strategies perform differently under different conditions. For example, choosing may be preferable in circumstances when one expert is clearly better placed than others and there is good feedback on expertise (Soll & Larrick, 2009). Hence, we propose that if a regular reviewer, known to the editor or the committee, appraises a manuscript or proposal as truly innovative, this advice should be followed.

Closely related is our fourth observation procedures are all important. The accuracy of judgments and forecasts can and will be influenced by small changes

in peer review procedures. Committees tend to develop their own cultures (Obrecht et al., 2007), potentially making it difficult to impose procedures. Individual reviewers, however, are more inclined to follow guidelines (Langfeldt, 2001), and this opens the possibility of providing them guidelines that, for example, anticipate the fate of controversial work. Of particular concern are group meetings, which can reinforce biases (Kerr, MacCoun, & Kramer, 1996), for example in reviewers' opinions (Obrecht et al., 2007). Kerr and Tindale (2004) state that groups will more easily choose decision alternatives that fit within their shared representation. Ideas that are shared among group members will dominate, because they require no additional justification. On the other hand, new ideas presented by only one person need further elaboration and experience more difficulties in being acknowledged. In other words: groups are less-than-optimal users of information and often ignore information that is not widely shared among their members. So group meetings can exacerbate shared cognitive biases such as the adherence to a scientific paradigm (Travis & Collins, 1991). In sum, in peer review as in forecasting, group meetings, if any (Armstrong, 2006), should preferably happen only after preliminary independent judgments are available.

Fifth, our findings indicate that judgments can be seriously affected by external constraints. In peer review, budgetary and time limits play major roles in the final decision-making. Scarcity in funds or journal space is a primary cause for refusal of innovative work. So the context and its constraints impact expert opinion and prediction accuracy. But instead of looking on, authors and editors should explore alternatives to traditional peer review processes in order to advance science. We presented two proposals aimed at advancing innovative work in this paper. In the forecasting community, the system whereby

researchers are invited to write a paper and reviewers are simply asked to offer advice that can improve the paper is well known (Armstrong & Pagell, 2003). In mathematics, the journal *Rejecta Mathematica* was launched with an aim of publishing rejected papers that may nevertheless be useful for the advancement of the discipline and science in general (Wakin, Rozell, Davenport, & Laska, 2009). It is our hope that more such proposals will find their way to editorial offices and funding agencies.

Last but not least, there clearly is a dearth of research on peer review (Jefferson, Rudin, Brodney Folse, & Davidoff, 2007; Marsh et al., 2008; Suls & Martin, 2009). This is an opportunity for scholars of judgmental forecasting. For example, Delphi-like feedback is rarely part of the peer review process, and this opens the possibility of experimental manipulations to study its effect on judgment and decision-making. Also, the generalizability of observations on cueing in judge-advisor systems can be studied via examination of editorial offices' handling of manuscript submissions (Sniezek & Buckley, 1995). Particularly interesting would be a study of the extent to which there exists a lack of cueing in manuscript peer review; a phenomenon that has been coined the Oppenheim effect, in awe of an anecdote about peer review being nothing more than a formal exercise (Gorman, 2007). More generally, along the lines of the suggestions for future research presented by Lawrence and colleagues (2006), peer review can serve as a context to study the value of expertise in forecasting, the influence of heuristics and biases on forecast accuracy, the use of information by experts or the influence of variations in procedures. The fact that positive decisions by peers, i.e. to accept or to fund, become public soon after having been taken, allows for studying the accuracy of their judgments systematically. This opportunity should be seized upon.

Reference List

Abrams, P. A. (1991). The predictive ability of peer review of grant proposals: The case of ecology and the US National Science Foundation. *Social Studies of Science, 21,* 111-132.

Armstrong, J. S. (1982a). Is review by peers as fair as it appears? *Interfaces, 12,* 62-74.

Armstrong, J. S. (1982b). Research on scientific journals: Implications for editors and authors. *Journal of Forecasting, 1,* 83-104.

Armstrong, J. S. (1996). Management folklore and management science - On portfolio planning, escalation bias, and such. *Interfaces, 26,* 25-55.

Armstrong, J. S. (1997). Peer review for journals: Evidence on quality control, fairness, and innovation. *Science and Engineering Ethics, 3,* 63-84.

Armstrong, J. S. (2001). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (pp. 417-438). Norwell: Kluwer Academic Publishers.

Armstrong, J. S. (2006). Should the forecasting process eliminate face-to-face meetings? *Foresight: The International Journal of Applied Forecasting,* 3-8.

Armstrong, J. S. & Hubbard, R. (1991). Does the need for agreement among reviewers inhibit the publication of controversial findings? *Behavorial and Brain Sciences, 14,* 136-137.

Armstrong, J. S. & Pagell, R. (2003). Reaping benefits from management research: Lessons from the Forecasting Principles project. *Interfaces, 33,* 89-111.

Bakanic, V., McPhail, C., & Simon, R. (1990). If at first you don't succeed: Review procedures for revised and resubmitted manuscripts. *The American Sociologist, 21,* 373-391.

Bedeian, A. G. (2004). Peer review and the social construction of knowledge in the management discipline. *Academy of Management Learning and Education, 3,* 198-216.

Bonaccio, S. & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes, 101,* 127-151.

Bornmann, L. & Daniel, H.-D. (2005). Selection of research fellowship recipients by committee peer review. Reliability, fairness and predictive validity of Board of Trustees' decisions. *Scientometrics, 63,* 297-320.

Bornmann, L. & Daniel, H.-D. (2006). Selecting scientific excellence through committee peer review - A citation analysis of publications previously published to approval or rejection of post-doctoral research fellowship applicants. *Scientometrics, 68,* 427-440.

Bornmann, L. & Daniel, H.-D. (2008a). Selecting manuscripts for a high-impact journal through peer review: A citation analysis of communications that were accepted by *Angewandte Chemie International Edition*, or rejected but published elsewhere. *Journal of the American Society for Information Science and Technology, 59,* 1841-1852.

Bornmann, L. & Daniel, H.-D. (2008b). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation, 64,* 45-80.

Bornmann, L. & Daniel, H.-D. (2009). Extent of type I and type II errors in editorial decisions: A case study on *Angewandte Chemie International Edition*. *Journal of Informetrics, 3,* 348-352.

Bornmann, L. & Daniel, H.-D. (2009). The manuscript reviewing process - Empirical research on review requests, review sequences and decision rules in peer review. *Library & Information Science Research, in press*.

Bornmann, L., Wallon, G., & Ledin, A. (2008). Does the committee peer review select the best applicants for funding? An investigation of the selection process for two European Molecular Biology Organization programmes. *PLoS ONE, 3,* e3480.

Bradley, J. V. (1981). Pernicious publication practices. *Bulletin of the Psychonomic Society, 18,* 31-34.

Callaham, M., Wears, R. L., & Weber, E. (2002). Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals. *JAMA: The Journal of the American Medical Association, 287,* 2847-2850.

Campanario, J. M. (1993). Consolation for the scientist: Sometimes it is hard to publish papers that are later highly-cited. *Social Studies of Science, 23,* 342-362.

Campanario, J. M. (1998). Peer review for journals as it stands today - Part 1. *Science Communication, 19,* 181-211.

Campanario, J. M. & Acedo, E. (2007). Rejecting highly cited papers: The views of scientists who encounter resistance to their discoveries from other scientists. *Journal of the American Society for Information Science and Technology, 58,* 734-743.

Cicchetti, D. V. (1991). The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavorial and Brain Sciences, 14,* 119-186.

Daniel, H.-D. (1993). *Guardians of science. Fairness and reliability of peer review*. Weinheim: Wiley-VCH.

Diener, E. (2009). Improving Psychological Science. Special Issue. *Perspectives on Psychological Science, 4,* 1-111.

Engels, T. C. E. & Kennedy, H. P. (2007). Enhancing a Delphi study on family-focused prevention. *Technological Forecasting and Social Change, 74,* 433-451.

Ernst, E., Saradeth, T., & Resch, K. L. (1993). Drawbacks of peer review. *Nature, 363,* 296.

Fara, P. (2009). *Science: A four thousand year history*. Oxford: Oxford University Press.

Fiske, D. W. & Fogg, L. (1990). But the reviewers are making different criticisms of my paper! : Diversity and uniqueness in reviewer comments. *American Psychologist, 45,* 591-598.

Fletcher, R. H. & Fletcher, S. W. (2003). The effectiveness of journal peer review. In F. Godlee & T. Jefferson (Eds.), *Peer review in health sciences* (2nd ed., pp. 62-75). London: BMJ Books.

Gans, J. S. & Shepherd, G. B. (1994). How are the mighty fallen: Rejected classic articles by leading economists. *Journal of Economic Perspectives, 8,* 165-179.

Garfield, E. (1991). To be an uncited scientist is no cause for shame. In *Essays of an information scientist: Science reviews, journalism inventiveness and other essays, Vol:14* (pp. 390-391).

Glidewell, J. C. (1988). Reflections on thirteen years of editing AJCP. *American Journal of Community Psychology, 16,* 759-770.

Goodman, S. N., Berlin, J., Fletcher, S. W., & Fletcher, R. H. (1994). Manuscript quality before and after peer review and editing at *Annals of Internal Medicine*. *Annals of Internal Medicine, 121,* 11-21.

Gorman, G. E. (2007). The Oppenheim effect in scholarly journal publishing. *Online Information Review, 31,* 417-419.

Gottfredson, S. D. (1978). Evaluating psychological research reports: Dimensions, reliability, and correlates of quality judgments. *American Psychologist, 33,* 920-934.

Gould, S. J. (1977). *Ever since Darwin: Reflections in natural history*. New York: Norton.

Graue, B. (2006). The transformative power of reviewing. *Educational Researcher, 35,* 36-41.

Hodgson, C. M. (1995). Evaluation of cardiovascular grant-in-aid applications by peer review: Influence of internal and external reviewers and committees. *Canadian Journal of Cardiology, 11,* 864-868.

Hornbostel, S., Böhmer, S., Klingsporn, B., Neufeld, J., & von Ins, M. (2009). Funding of young scientist and scientific excellence. *Scientometrics, 79,* 171-190.

Horrobin, D. F. (1996). Peer review of grant applications: a harbinger for mediocrity in clinical research? *The Lancet, 348,* 1293-1295.

Jayasinghe, U. W., Marsh, H. W., & Bond, N. (2006). A new reader trial approach to peer review in funding research grants: An Australian experiment. *Scientometrics, 69,* 591-606.

Jefferson, T., Rudin, M., Brodney Folse, S., & Davidoff, F. (2007). Editorial peer review for improving the quality of reports of biomedical studies. *Cochrane Database of Systematic Reviews, Issue 1,* Art. No.: MR000016.

Jennings, C. G. (2006). *Quality and value: The true purpose of peer review*. Retrieved 13 July 2009, from: http://www.nature.com/nature/peerreview/debate/nature05032.html.

Judge, T. A., Cable, D. M., Colbert, A. E., & Rynes, S. L. (2007). What causes a management article to be cited - Article, author, or journal? *Academy of Management Journal, 50,* 491-506.

Kameda, T. & Sugimori, S. (1995). Procedural influence in 2-step group decision-making - Power of local majorities in consensus formation. *Journal of Personality and Social Psychology, 69,* 865-876.

Kerr, N. L., MacCoun, R. J., & Kramer, G. P. (1996). Bias in judgment: Comparing individuals and groups. *Psychological Review, 103,* 687-719.

Kerr, N. L. & Tindale, R. S. (2004). Group performance and decision making. *Annual Review of Psychology, 55,* 623-655.

Kostoff, R. N. (1995). Federal research impact assessment: axioms, approaches, applications. *Scientometrics, 34,* 163-206.

Kupfersmid, J. & Wonderly, D. M. (1994). *An author's guide to publishing better articles in better journals in the behavioural sciences*. Hoboken, NJ: Wiley & Sons.

Laband, D. N. (1990). Is there value-added from the review process in economics?: Preliminary evidence from authors. *The Quarterly Journal of Economics, 105,* 341-352.

Langfeldt, L. (2001). The decision-making constraints and processes of grant peer review, and their effects on the review outcome. *Social Studies of Science, 31,* 820-841.

Langfeldt, L. (2004). Expert panels evaluating research: decision-making and sources of bias. *Research Evaluation, 13,* 51-62.

Langfeldt, L. (2006). The policy challenges of peer review: managing bias, conflict of interests and interdisciplinary assessments. *Research Evaluation, 15,* 31-41.

Lawrence, M., Godwin, P., O'Conner, M., & Önkal, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting, 22,* 493-518.

Lawrence, P. A. (2003). The politics of publication. *Nature, 422,* 259-261.

Lehmann, S., Jackson, A. D., & Lautrup, B. E. (2006). Measures for measures. *Nature, 444,* 1003-1004.

Leimu, R. & Koricheva, J. (2005). What determines the citation frequency of ecological papers? *Trends in Ecology & Evolution, 20,* 28-32.

Levin, I. P., Prosansky, C. M., Heller, D., & Brunick, B. M. (2001). Prescreening of choice options in 'positive' and 'negative' decision-making tasks. *Journal of Behavioral Decision Making, 14,* 279-293.

Levitt, S. D. & List, J. A. (2009). *Was there really a Hawthorne effect at the Hawthorne plant? An analysis of the original illumination experiments*. NBER Working Papers Series: Working Paper 15016. Retrieved 1 July 2009, from http://www.nber.org/papers/w15016.pdf.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research, 1,* 161-175.

Marsh, H. W. & Bornmann, L. (2009). Do women have less success in peer review? *Nature, 459,* 602.

Marsh, H. W., Jayasinghe, U. W., & Bond, N. (2008). Improving the peer-review process for grant applications: Reliability, validity, bias, and generalizability. *American Psychologist, 63,* 160-168.

Mavis, B. & Katz, M. (2003). Evaluation of a program supporting scholarly productivity for new investigators. *Academic Medicine, 78,* 757-765.

McDonals, R. J., Cloft, H. J., & Kallmes, D. F. (2009). Fate of manuscripts previously rejected by the American Journal of Neuroradiology: a follow-up analysis. *American Journal of Neuroradiology, 30,* 253-256.

Melin, G. & Danell, R. (2006). The top eight percent: development of approved and rejected applicants for a prestigious grant in Sweden. *Science and Public Policy, 33,* 702-712.

Moed, H. F. (2007). The future of research evaluation rests with an intelligent combination of advanced metrics and transparent peer review. *Science and Public Policy, 34,* 575-584.

Nederhof, A. & van Raan, A. (1987). Peer review and bibliometric indicators of scientific performance: A comparison of cum laude doctorates with ordinary doctorates in physics. *Scientometrics, 11,* 333-350.

Nederhof, A. & van Raan, A. (1989). A validation study of bibliometric indicators: The comparative performance of cum laude doctorates in chemistry. *Scientometrics, 17,* 427-435.

Nickerson, R. S. (2005). What authors want from journal reviewers and editors. *American Psychologist, 60,* 661-662.

Obrecht, M., Tibelius, K., & D'Aloisio, G. (2007). Examining the value added by committee discussion in the review of applications for research awards. *Research Evaluation, 16,* 79-91.

Opthof, T., Furstner, F., van Geer, M., & Coronel, R. (2000). Regrets or no regrets? No regrets! The fate of rejected manuscripts. *Cardiovascular Research, 45,* 255-258.

Peters, D. P. & Ceci, S. J. (1982). Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences, 5,* 187-255.

Pierie, J. P. E., Walvoort, H. C., & Overbeke, A. J. P. (1996). Readers' evaluation of effect of peer review and editing on quality of articles in the *Nederlands Tijdschrift voor Geneeskunde*. *The Lancet, 348,* 1480-1483.

Racki, G. (2009). Rank-normalized journal impact factor as a predictive tool. *Archivum Immunologiae et Therapiae Experimentalis, 57,* 39-43.

Rowe, G. & Wright, G. (2001). Expert opinions in forecasting: The role of the Delphi technique. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (pp. 125-144). Norwell: Kluwer Academic Publishers.

Rowe, G., Wright, G., & McColl, A. (2005). Judgment change during Delphi-like procedures: The role of majority influence, expertise, and confidence. *Technological Forecasting and Social Change, 72,* 377-399.

Samuelson, W. & Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty, I,* 7-59.

Saper, C. B. & Maunsell, J. H. R. (2009). The neuroscience peer review consortium. *The Journal of Comparative Neurology, 513,* 333-334.

Seglen, P. O. (1994). Causal relationship between article citedness and journal impact. *Journal of the American Society for Information Science, 45,* 1-11.

Shafir, E. (1993). Choosing versus rejecting: Why some options are both better and worse than others. *Memory & Cognition, 21,* 546-556.

Siegelman, S. S. (1991). Assassins and Zealots: Variations in peer review. *Radiology, 178,* 637-642.

Skoda, Z. (6-11-2008). Re: The kind of email I don't need. Retrieved 22 July 2009, from http://golem.ph.utexas.edu/category/2008/11/the_kind_of_email_i_dont_need.html#c019806.

Sniezek, J. A. & Buckley, T. (1995). Cueing and cognitive conflict in judge-advisor decision making. *Organizational Behavior and Human Decision Processes, 62,* 159-174.

Soll, J. B. & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology-Learning Memory and Cognition, 35,* 780-805.

Starbuck, W. H. (2005). How much better are the most-prestigious journals? The statistics of academic publication. *Organization Science, 16,* 180-200.

Straub, D. W. (2008). Type II reviewing errors and the search for exciting papers. *MIS Quarterly, 32,* 5-10.

Suls, J. & Martin, R. (2009). The air we breathe: A critical look at practices and alternatives in the peer-review process. *Perspectives on Psychological Science, 4,* 40-50.

Sweitzer, B. J. & Cullen, D. J. (1994). How well does a journal's peer review process function? A survey of authors' opinions. *JAMA: The Journal of the American Medical Association, 272,* 152-153.

Thornley, R., Spence, M. W., Taylor, M., & Magnan, J. (2002). New decision tool to evaluate award selection process. *The Journal of Research Administration, 33,* 49-56.

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

Tversky, A. & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics, 107,* 1039-1061.

van den Besselaar, P. & Leydesdorff, L. (2009). Past performance, peer review, and project selection: A case study in the social and behavorial sciences. *Research Evaluation, 18,* 273-288.

van Raan, A. F. J. (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics, 36,* 397-420.

van Raan, A. F. J. (2004a). Measuring science. Capita selecta of current main issues. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems* (pp. 19-50). Dordrecht: Kluwer Academic Publishers.

van Raan, A. (2004b). Sleeping Beauties in science. *Scientometrics, 59,* 467-472.

Wakin, M., Rozell, C., Davenport, M., & Laska, J. (2009). Letter from the editors. *Rejecta Mathematica, 1,* 1-3.

Weller, A. C. (2001). *Editorial peer review: Its strengths and weaknesses*. Medford, NJ: American Society for Information Science and Technology.

Wilson, J. D. (1978). Peer review and publication: Presidential address before the 70th annual meeting of the American Society for Clinical Investigation. *Journal of Clinical Investigation, 61,* 1697-1701.

Wright, G., Lawrence, M. J., & Collopy, F. (1996). The role and validity of judgment in forecasting. *International Journal of Forecasting, 12,* 1-8.

[i] Of course, citation counts are not the only possible ways to scrutinize the selections by peers. Thanks to the fact that the positive decisions of peers, i.e. to publish, to grant, to fund, etc., are generally made public, the decisions themselves are open to scrutiny by other peers and the public too. As illustrated by the outrage resulting from the publicizing of misbehaviour at *Chaos, Solitions & Fractals* (Skoda, 2008), the wider academic community does take up this important task.

[ii] Motivating decisions is typically required because of the transparency rules governing funding agencies.

Wim G. G. Benda holds a Bachelor in Philosophy (1997, University of Antwerp), a Master in Antropology (1999, K.U.Leuven), and a Master in Cultures and Development Studies (2000, K.U.Leuven). He has a broad interest in the social sciences and has published extensively on developing countries (Indonesia, Venezuela). He is currently a human sciences teacher in high school and a philosophy and organization theory lecturer at the Erasmus University College Brussels.

Tim C. E. Engels holds a PhD in Psychology (2006, University of Brussels). He is a coordinator of research evaluations at the University of Antwerp and a senior researcher in the Flemish Centre for R&D Monitoring (ECOOM). He is also a psychology lecturer at the Antwerp Maritime Academy. His current research focuses on the dynamics of peer review processes and on publication activity in the social sciences and humanities.

Acknowledgements
The authors thank the Flemish government for its support of the Centre for R&D Monitoring. We thank Saskia Peersman, Dr. Eric Spruyt, Nathalie Stevens, Prof. Dr. Jean-Pierre Timmermans, and Dr. Griet Vermeesch for their advice, comments, facilitation, and help during the preparation of this manuscript.
The authors wish to express their gratitude to the editors of the special issue for their invitation and to the three reviewers (J. Scott Armstrong and two anonymous reviewers) for their helpful comments.