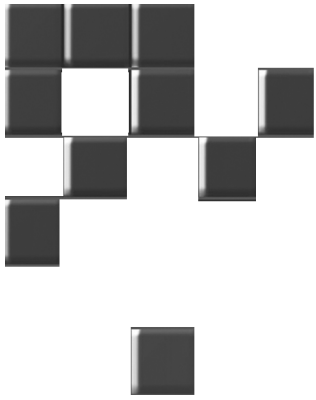

SECTION III

# Scholarly Communication

CHAPTER 5

# Scientific Peer Review

*Lutz Bornmann*
*Eidgenössische Technische Hochschule, Zurich, Switzerland*

## Introduction

Peer review is the principal mechanism for quality control in most scientific disciplines. By assessing the quality of research, peer review determines what scientific research receives funding and what research results are published. This review of the literature published on the topic of peer review describes the state of research on journal, fellowship, and grant peer review. The emphasis is on empirical research dealing with the reliability, fairness, and predictive validity of the process—the three quality criteria for professional evaluations.

Peer review, the instrument for ensuring trustworthiness (Cronin, 2005), grounds all scholarship (Ziman, 2000). Quality control undertaken by experts in the traditional peer review of manuscripts for scientific journals is essential in most scientific disciplines in order to create valid and reliable knowledge (Hemlin & Rasmussen, 2006). According to Lamont (2009), peers monitor the flow of ideas through the various gates of the academic community. But journal peer review influences not only scholarship: "The Intergovernmental Panel on Climate Change and other similar advisory groups base their judgments on peer-reviewed literature, and this is part of their success. Many legal decisions and regulations also depend on peer-reviewed science" (Alberts, Hanson, & Kelner, 2008, p. 15). The mid-18th century is cited as the beginning of reviewing. At that time the Royal Society in London took over fiscal responsibility for the journal *Philosophical Transactions* and established what they called a Committee on Papers (Kronick, 1990). Nowadays, innovative review possibilities have emerged on the World Wide Web, where peers can comment on internet-based materials in a review process sometimes called sky-writing (Harnad, 1990).

Almost all aspects of the contemporary scientific enterprise rely on quality evaluations by peers. Such evaluations determine, among other things, who gets which job, who gets tenure, and who gets which awards and honors (Feist, 2006). Research evaluation systems in various countries (e.g., the British research assessment exercise) are normally based on peer review. Whitley and Gläser's (2007) edited book shows how these systems are changing the organization of scientific knowledge production

and universities in the countries involved (see also Moed, 2008). Aside from the selection of manuscripts for publication in journals, the most common contemporary application of peer review in scientific research is for the selection of fellowship and grant applications. Following World War II, and at first mostly just in the U.S., peer review became the process for allocating research funds (Biagioli, 2002). Today researchers rely less and less on regular research funds from their universities and more on external research grants that are allocated on the basis of peer review (Guston, 2003). For Wessely and Wood (1999) the peer review of grant proposals may be more relevant than publication practices to the health of science. Good papers will get published somewhere, as will bad ones, whereas applications for grants that do not succeed represent research that no one conducts.

Peers or colleagues asked to evaluate fellowship or grant applications or manuscripts in a peer review process take on the responsibility for ensuring high standards in their disciplines. Although peers active in the same field might be unaware of other perspectives, they "are said to be in the best position to know whether quality standards have been met and a contribution to knowledge made" (Eisenhart, 2002, p. 241). Peer evaluation in research thus entails a process by which a jury of equals active in a given scientific field convenes to evaluate the undertaking of scientific activity or its outcomes (i.e., applications for research fellowships and grants and manuscripts for publication). Such a jury of equals may be consulted as a group or individually, without the need for personal contacts among the evaluators. The peer review process lets the active producers of science, the experts, become the gatekeepers of science (McClellan, 2003). "By using the judgment and opinions of peers and members of the 'community of science,' the process of peer review is aimed at keeping the review 'in the family'" (Geisler, 2000, pp. 218–219). Nevertheless, keeping it "in the family" can result in intellectual closed-mindedness.

### *The Formation of a Peer Review Process*

Social psychology conceptualizes the peer review process as a social judgment process of individuals in a small group (for example, one or more reviewers and one or more editors of a disciplinary in-group in manuscript reviewing) (Krampen & Montada, 2002). In the review of manuscripts intended for publication and grant proposals for research funding, it is the reviewers' task to recommend for selection the "best" scientific research under the condition of scarce resources (such as limited space in journals, limited funds) (Hackett & Chubin, 2003). "With grants, an applicant submits a proposal, which is then reviewed by peers who make a judgement on its merits and eligibility for funding. With publications, an author submits a paper to a journal or a book proposal to a publisher, and peers are asked to offer a judgement as to whether it should be published" (British Academy, 2007, p. 2).

In journal peer review, reviewers sought by the editor normally provide a written review and an overall publication recommendation. "The editor, on the basis of the reviews and his or her own evaluation, decides to reject the submission, seek further review, ask the author to revise the manuscript in response to suggestions by the reviewers and the editor, or accept the manuscript" (Jayasinghe, Marsh, & Bond, 2001, p. 344). Most of the studies examining the relation of reviewers' (overall) ratings and editors' decisions on submissions at single journals have found that the reviewers' ratings are highly correlated with the editors' final decisions (Bakanic, McPhail, & Simon, 1987; Bornmann & Daniel, 2008a; Fogg & Fiske, 1993; Lock, 1985; Petty & Fleming, 1999; Sternberg, Hojjat, Brigockas, & Grigorenko, 1997; Zuckerman & Merton, 1971a). That means editors' decisions on manuscripts depend on the judgments of the reviewers. Peer review for fellowships shows similar associations between reviewers' ratings and the decisions of a selection committee (e.g., Bornmann, Mutz, & Daniel, 2007b).

Many aspects of the peer review process vary case by case and this variation largely depends on the type of application (see Hansson, 2002; Marsh & Ball, 1991; Shashok, 2005). Several publications describe the differing peer review processes at different research funding agencies, such as those by Geisler (2000), Kostoff (1997), and the U.S. General Accounting Office (1999). Weller (2002) specifically describes the journal peer review processes, which may proceed by highly formalized protocols or leave the choice of selection criteria to the peers in the committee (Hornbostel, 1997). Reviewers may work anonymously or openly. The persons reviewed may or may not be anonymous (double-blind versus single-blind). Reviewers may be assigned permanently or ad hoc (Geisler, 2000). A reviewer may represent one scientific discipline or a variety of disciplines (U.S. Office of Management and Budget, 2004). Funding agencies may select reviewers from academia, private industry, and/or government (U.S. General Accounting Office, 1999). A single reviewer or a committee may provide a peer review (Marsh & Ball, 1991). After a group of reviewers is assigned, the members may review either collectively or independently (Geisler, 2000). The peer review process can make its results public (Pöschl, 2004) or reveal them only to those directly involved. Accordingly, peer review practices can be characterized as heterogeneous processes across and among different disciplines, journal editors, funding agencies, rating schemas, and so on.

### *The Content of a Review*

A review normally consists of a summarizing judgment regarding suitability for publication or funding, followed by comments, sometimes numbered by point or page, which may track criticisms in a sequential manner (Gosden, 2003). Negative comments predominate in review documents according to the findings of Bakanic, McPhail, and Simon (1989) in the analysis of manuscripts submitted to the *American Sociological*

*Review*; positive comments occur far less frequently. This result is in accordance with critical rationalism, the epistemological philosophy advanced by Popper (1961): critical rationalists hold that claims to knowledge can and should be rationally criticized so that scholarship moves forward. Bornmann, Nast, and Daniel (2008) examined the criteria employed in peer review. They conducted a quantitative content analysis of 46 research studies on editors' and reviewers' criteria for the assessment of manuscripts and their grounds for accepting or rejecting manuscripts. The 572 differing criteria and reasons from the 46 studies could be assigned to nine areas: (1) relevance of contribution, (2) writing/presentation, (3) design/conception, (4) method/statistics, (5) discussion of results, (6) reference to the literature and documentation, (7) theory, (8) author's reputation/institutional affiliation, and (9) ethics. The most significant criteria for editors and reviewers in manuscript assessment are those that relate to the quality of the research underlying a manuscript: theory, design/conception, and discussion of results.

Bornmann and Daniel (2005b) investigated the fellowship peer review process of the Boehringer Ingelheim Fonds (BIF, Heidesheim, Germany, an international foundation for the promotion of basic research in biomedicine). They found that fellowships were awarded to post-graduate researchers according to the following main criteria: (1) scientific quality as demonstrated by the applicant's achievements to date, (2) the originality of the proposed research project, and (3) the scientific standing of the laboratory where the research will be conducted. Even if the applicant's track record is a predictor of success in the grant peer review process of the National Science Foundation (NSF) (Abrams, 1991), classic studies by Jonathan R. Cole and Stephen Cole found that the characteristics of the proposal were more important than attributes of the applicant (see an overview in Cole, 1992).

## *Advantages of Peer Review*

Proponents of the peer review system argue that it is more effective than any other known instrument for self-regulation in promoting the critical selection that is so crucial to the evolution of scientific knowledge. Putting it into a wider context: According to Popper's (1961) critical rationalism, intellectual life and institutions should be subjected to "maximum criticism, in order to counteract and eliminate as much intellectual error as possible" (Bartley, 1984, p. 113). If, for example, the editors of the journal *Social Text* had sent Alan D. Sokal's manuscript "Transgressing the boundaries: Toward a transformative hermeneutics of quantum gravity" to external peers, the manuscript would probably have been rejected and the well-known Sokal affair (Sokal, 2008) might not have happened. It can also be assumed that the fake paper Philip Davis submitted to the open-access journal *The Open Information Science Journal*, which the editor accepted for publication (Editor to quit

over hoax open-access paper, 2009), would have been recommended for rejection if peers had been involved.

Evidence supports the view that peer review improves the quality of the reporting of research results (Goodman, Berlin, Fletcher, & Fletcher, 1994; Pierie, Walvoort, & Overbeke, 1996). Although alternatives to peer review have been suggested, they have not been implemented. For example, Roy (1985) developed the Peer-Reviewed Formula System, in which research money is allocated proportional to prior research productivity. This system certainly disadvantages younger scientists. Abelson (1980, p. 62), a proponent of peer review, writes: "The most important and effective mechanism for attaining good standards of quality in journals is the peer review system." According to Shatz (2004, p. 30) journal peer review "motivates scholars to produce their best, provides feedback that substantially improves work which is submitted, and enables scholars to identify products they will find worth reading."

The proponents of peer review are not the only people well disposed to the process. A series of surveys on grant and journal peer review have already reported on scientists' wide satisfaction with it. In a global survey on the attitudes and behavior of 3,040 academics, a large majority (85 percent) agreed that journal peer review greatly helps scientific communication, and 83 percent believed that without peer review researchers would have no control over scientific communication (Publishing Research Consortium, 2008). The majority of corresponding authors of papers published in *Academy of Management Journal* and *Academy of Management Review* agreed that the reviewers' recommended revisions improved their papers (Bedeian, 2003). Similar results obtained in surveys of authors for *Nature* (Overview: *Nature*'s peer review trial, 2006) and *Obstetrics & Gynecology* papers (Gibson, Spong, Simonsen, Martin, & Scott, 2008). Ninety-seven percent of more than a thousand *Astronomy & Astrophysics* authors reported that reviewers had dealt competently with their manuscripts (Bertout & Schneider, 2004). The Swiss National Science Foundation (SNSF) surveyed researchers at Swiss universities about its grant reviews and found that the evaluation process was regarded as good and its administration efficient (Hoffmann, Joye, Kuhn, & Métral, 2002). The German Research Foundation's (DFG) survey of subject reviewers yielded a similar positive result (Hornbostel & Olbrecht, 2007).

## *Critics of Peer Review*

Critics of peer review argue that: (1) reviewers rarely agree on whether to recommend that a manuscript be published or a research grant be awarded, thus making for poor *reliability* of the peer review process; (2) reviewers' recommendations are frequently biased, that is, judgments are not based solely on scientific merit, but are also influenced by personal attributes of the authors, applicants, or the reviewers themselves (where the *fairness* of the process is not a given); (3) the process lacks *predictive*

*validity* because there is little or no relationship between the reviewers' judgments and the subsequent usefulness of the work to the scientific community, as indicated by the frequency of citations of the work in later scientific papers; (4) reviewing is inefficient because it delays publications; inhibits the publication of new, innovative, and unconventional ideas; and is time consuming and costly; and (5) reviewing can be personally damaging, an experience that is particularly painful and distressing for new authors (for criticism on peer review see Eysenck & Eysenck, 1992; Ross, 1980).

Frey (2003, p. 206) holds that peer review constitutes a form of intellectual prostitution: The authors are forced to follow reviewers' demands "slavishly." Critics of peer review often cite the main results of the highly influential study on grant peer review at the NSF conducted by Cole, Cole, and Simon (1981, p. 885): "The fate of a particular application is roughly half determined by the characteristics of the proposal and the principal investigator, and about half by apparently random elements which might be characterized as 'the luck of the reviewer draw.'" The only reason for the further implementation of the peer review process—according to its skeptics—is the lack of any clear consensus on a better alternative (Young, 2003).

Nevertheless, if a workable synthesis could evolve between the diametrically opposed positions of proponents and opponents of peer review, perhaps an improved system could be developed.

## *Research on Peer Review*

Because peer review is so central to what is published and funded, and because so much hinges on peer review in and outside of science, it is essential that it be carried out well and professionally (Hames, 2007). The research on peer review, which in recent years has addressed criticisms of the process, deals for the most part with journal peer review (see overviews in Armstrong, 1997; Campanario, 1998a, 1998b; Overbeke & Wager, 2003; Speck, 1993; Stieg Dalton, 1995) and somewhat less frequently with peer review for fellowship and grant proposals (see overviews in Demicheli & Pietrantonj, 2007; Kostoff, 2004; Wessely, 1998). Weller (2002), a former member of the editorial staff of the *Journal of the American Medical Association* (*JAMA*), has provided the most comprehensive review of research on journal peer review. Her book, *Editorial Peer Review: Its Strengths and Weaknesses*, covers 1,439 studies published in English between 1945 and 1997. Garfield's (2004) historiography shows that as of September 2006 some 3,720 publications in 1,228 journals by 6,708 different authors have focused on the peer review process. In addition to these publications, a multitude of monographs, compilations, and gray literature was not included in the historiography (which visualizes the results of literature searches in the *Web of Science*, provided by Thomson Reuters; this database covers only journal publications).

Over the past dozen years scientific research on journal peer review has been stimulated by *JAMA*'s conferences on the topic of journal peer review. The first *JAMA* conference on peer review and biomedical publication was held in 1989, and five more have been held in 1993, 1997, 2001, 2005, and 2009. *JAMA* published research articles from the first through fourth peer review congresses in peer review theme issues: March 6, 1990; July 13, 1994; July 15, 1998; June 5, 2002 (*Journal of the American Medical Association*, 1990, 1994, 1998, 2002). Weller (2002, p. 12) comments: "Overall, although isolated research endeavors were certainly taking place, the congress failed to uncover a large, coordinated research effort in the field of editorial peer review in biomedical publications."

Because comprehensive review articles of research on peer review have been published prior to August 2000, this review concentrates on research from the last decade. It particularly focuses on recent methodological developments, which have already been suggested or used in the investigation of peer review processes. According to Stieg Dalton (1995, p. 215) many peer review studies have methodological weaknesses, and "most of the publications on journal peer review are more opinion than research, often the ruminations of a former editor." This review focuses on studies concerning journal, fellowship, and grant peer review, where most of the research has focused. Peer review of completed research (e.g., research assessment exercises in various countries) and the review of conference papers or abstracts are dealt with only marginally.

The literature research for this review was conducted at the beginning of 2009. In a first step, studies were located using the reference lists provided by narrative reviews of research on journal and grant peer review and using tables of contents of certain journals (e.g., *JAMA*, *Nature*, *Research Policy*, *Scientometrics*). In a second step, in order to obtain keywords for searching databases, a bibliogram (White, 2005) was prepared for the studies located in the first step. The bibliogram ranks by frequency the words included in the abstracts of the studies located. Words at the top of the ranking list (e.g., peer review, quality, research, and evaluation) were used for searches in literature databases (e.g., Web of Science and Scopus) and internet search engines (e.g., Google). In the final step of the literature search, all of the citing publications for a series of articles (found in the first and second steps) which had a substantially large number of citations were examined.

In the following sections, empirical research is investigated with regard to the three assessment criteria for professional evaluations of peer review processes: reliability, fairness, and predictive validity. As an assessment tool, peer review is asked to be reliable (is the selection of scientific contributions reliable or is the result purely incidental?), fair (are certain groups of applicants or authors favored or at a disadvantage?), and predictively valid (do the selection decisions correlate with scientific performance measures subsequent to decision?). These three quantitative social science research tools used by psychologists are not

only connected to the criticism on peer review, they are also uniquely appropriate for evaluating the peer review process (Marsh, Bonds, & Jayasinghe, 2007).

## Reliability, Fairness, and Predictive Validity of Peer Review

### *Reliability of Peer Review*

"In everyday life, intersubjectivity is equated with realism" (Ziman, 2000, p. 106). Therefore, scientific discourse is also distinguished by its striving for consensus. Scientific activity would clearly be impossible unless scientists could come to similar conclusions. According to Wiley (2008, p. 31) "just as results from lab experiments provide clues to an underlying biological process, reviewer comments are also clues to an underlying reality (they did not like your grant for some reason). For example, if all reviewers mention the same point, then it is a good bet that it is important and real." An established consensus among scientists must of course be voluntary and achieved under conditions of free and open criticism (Ziman, 2000). The norms of science make these conditions possible and regulate them (Merton, 1942): The norms of communalism (scientific knowledge should be made public knowledge) and universalism (knowledge claims should be judged impersonally, independently of their source) envisage eventual agreement. "But the norm of 'organized scepticism', which energizes critical debates, rules out any official procedure for closing them. Consensus and dissensus are thus promoted simultaneously" (Ziman, 2000, p. 255).

A group of peers normally reviews journal manuscripts, grant applications, and fellowship proposals and recommends acceptance or rejection. Cicchetti (1991, p. 120) defines inter-reviewer reliability "as the extent to which two or more independent reviews of the same scientific document agree." Manuscripts and applications are rated reliably when there is a high level of agreement between independent reviewers. In many studies of peer review the intraclass correlation coefficient measures the extent of agreement within peer review groups. "The intraclass correlation coefficient (ICC) … is a variance decomposition method to assess the portion of overall variance attributable to between-subject variability. … Raters are assumed to share common metric and homogeneous variance (i.e., *intra*class variance)" (von Eye & Mun, 2005, p. 116). The ICC can vary between -1.0 and +1.0. However, high agreement alone with low between-reviewer variability cannot result in high reliability because a certain level of agreement can be expected to occur on the basis of chance alone. Therefore the Kappa coefficient figures in many studies on peer review as a measure of between-reviewer variability. Kappa ($\kappa$) statistically indicates the level of agreement between two or more raters. If the raters are in complete agreement then $\kappa = 1$; if $\kappa$ is

near 0, the observed level of agreement is not much higher than by chance (von Eye & Mun, 2005).

If a submission (manuscript or application) meets scientific standards and contributes to the advancement of science, it can be expected that two or more reviewers will agree on its value. This is frequently not the case. Ernst, Saradeth, and Resch (1993) offer a dramatic demonstration of the unreliability of the journal peer review process. Copies of a paper submitted to a medical journal were sent simultaneously to 45 experts. They were asked to express their opinion of the paper with the journal's standard questionnaire containing eight quality criteria on a numerical scale from 5 (excellent) to 1 (unacceptable). The 31 completed forms demonstrate poor reproducibility with extreme judgments ranging from "unacceptable" to "excellent" for most criteria. Table 5.1 shows an overview of the results of some studies on reliability in the areas of journal (submission of manuscripts), meeting (submission of abstracts), and grant (submission of applications) peer review. The results indicate that the levels of inter-reviewer agreement, when corrected for chance, generally fall in the range from 0.2 to 0.4. A meta-analysis by Bornmann, Mutz, and Daniel (2009) of 48 studies on the reliability of agreement between reviewers' ratings in journal peer review reports overall agreement coefficients of ~.23/ ~.34 (ICC and Pearson product-moment correlation) and

**Table 5.1 Reliability: agreement among reviewers**

| | **Kappa coefficient/ intraclass correlation** |
|---|---|
| **Journal (submission of manuscripts)** | |
| *Social Problems* (Smigel & Ross, 1970) | .40 |
| *Journal of Educational Psychology* (Marsh & Ball, 1981) | .34 |
| *British Medical Journal* (Lock, 1985) | .31 |
| *American Sociological Review* (Hargens & Herting, 1990) | .28 |
| *Physiological Zoology* (Hargens & Herting, 1990) | .28 |
| *Journal of Personality and Social Psychology* (Scott, 1974) | .26 |
| *New England Journal of Medicine* (Ingelfinger, 1974) | .26 |
| *Law & Society Review* (Hargens & Herting, 1990) | .17 |
| *Angewandte Chemie International Edition* (Bornmann & Daniel, 2008a) | .15 |
| *Angewandte Chemie* (Daniel, 1993/2004) | .14 |
| *Physical Therapy* (Bohannon, 1986) | .12 |
| **Meeting (submission of abstracts)** | |
| National Meeting of the American Association for the Study of Liver Disease (Cicchetti & Conn, 1976) | .24 |
| Annual Meeting of the Orthopaedic Trauma Association (Bhandari, Templeman, & Tornetta, 2004) | .23 |
| **Research funding organization (submission of applications)** | |
| American Heart Association (Wiener, Urivetsky, Bregman, Cohen, Eich, Gootman, et al., 1977) | .37 |
| National Science Foundation (solid states physics) (Cicchetti, 1991) | .32 |
| Heart and Stroke Foundation, Medical Research Council of Canada (Hodgson, 1997) | .29 |

.17 (κ). According to Fleiss's (1981) guidelines, Kappa coefficients between 0 and 0.2 indicate a slight level of reviewer agreement.

Given these results, the norm of organized skepticism, which ensures that "much of the action in science is systematic controversy over the credibility of 'facts' and theories" (Ziman, 2000, p. 225), appears to be functioning effectively. The norm of organized skepticism is stronger than the norms of communalism and universalism, both of which envisage eventual agreement. Further examination of the data from various studies on the reliability of journal peer review by Cicchetti (1991, 1997) and Weller (2002, Chapter 6) shows that agreement is substantially higher in recommendations for rejection than for acceptance; reviewers are twice as likely to agree on rejection than on acceptance (Weller, 2002). "Although there seems to be satisfactory qualitative agreement as to what constitutes a manuscript of poor quality (a list of don'ts), there is little consistency among reviewers as to what constitutes good quality" (Kupfersmid, 1988, p. 639). With regard to grant reviewing, Opthof and Wilde (2009, p. 153) state that "the peer-review system is solid for recognising work that should not be granted, but is poor, if not unsuitable, for making a distinction between what is very good or excellent. This renders the whole process subject to personal bias." It can be surmised that the peer review process is better at encumbering the progress of bad research than at identifying the "best."

For Marsh and colleagues (2007), the lack of reliability constitutes the most important weakness of the peer review process. Bedeian (2004, p. 202) excludes the possibility of a "universal and articulable latent dimension of 'merit' or 'publishability,' against which manuscripts can be judged, as the definition of any such standard will inevitably vary from one reader to the next" (see also Campanario, 1998a; LaFollette, 1992). "The 'point faible' of the peer review system is not so much the referee and his human judgment (though that certainly is one of its weaknesses); it is the SELECTION of the referee, a function performed by the Editor" (Harnad, 1996, p. 111). Evidently the question becomes one of who constitutes a peer for a certain manuscript; journal editors answer that very differently. For one it constitutes a researcher in the same discipline, for another it could be someone who uses similar methods, and for a third, someone who conducts the same kind of research (Suls & Martin, 2009).

The fate of a manuscript depends on which small sample of reviewers underpins the editorial decision, as research such as that of Bornmann and Daniel (2009) for the AC-IE (Angewandte Chemie-International Edition) indicates. In AC-IE's peer review process, a manuscript is generally published only if two reviewers rate the results of the study as important and also recommend publication in the journal (what the editors have called the clear-cut rule). Even though the clear-cut rule is based on two reviewer reports, submitted manuscripts generally go out to three reviewers in total. An editor explains this process in a letter to an author as follows: "Many papers are sent initially to three referees (as

in this case), but in today's increasingly busy climate there are many referees unable to review papers because of other commitments. On the other hand we have a responsibility to authors to make a rapid fair decision on the outcome of papers" (Bornmann & Daniel, in press). For 23 percent of those manuscripts, for which a third reviewer report arrived after the editorial decision was made (37 of 162), this rule would have led to a different decision if the third report had replaced either of the others. Thus, if the editor had been able to consider the recommendations from all three reviewers whom the editor judged to be suitable to review the manuscript, the editor would presumably have made a different decision.

According to J. R. Cole (2000, p. 115), a low level of agreement among reviewers reflects the lack of consensus that is prevalent in all scientific disciplines at the "research frontier." Cole says that usually no one reliably assesses scientific work occurring at the frontiers of research. Eckberg (1991) and Kostoff (1995) point out that differing judgments in peer review are not necessarily a sign of disagreement about the quality of a manuscript but may instead reveal differing positions, judgment criteria, and areas of competency among the reviewers. In addition, reviewers tend to be either more critical or more lenient in their judgments (Siegelman, 1991) if they direct their attention to "different points, and may draw different conclusions about 'worth'" (Eckberg, 1991, p. 146). The question of whether the comments of reviewers are in fact based on different perspectives, positions, and so forth has been examined by only a few empirical studies (Weller, 2002).

Fiske and Fogg (1990, p. 591), for example, found that reviewers of the same submission simply commented on different aspects of the manuscript: "In the typical case, two reviews of the same paper had no critical point in common. … [T]hey wrote about different topics, each making points that were appropriate and accurate. As a consequence, their recommendations about editorial decisions showed hardly any agreement." Bedeian (2004) brings the perspective of social constructivism to bear on the issue. The argument runs as follows: "Knowledge-claims are socially constructed, being subject to the inevitable author-editor-referee tensions operating throughout the publication process. The impact of this social component, and the influence it has on referee judgments and, in turn, upon claims entering the list of science, offers an (as yet) overlooked explanation for the seeming randomness with which manuscripts are either accepted or rejected" (Bedeian, 2004, p. 204; see also Cole, 1992).

Although a high level of agreement among the reviewers is usually seen as an indicator of the high quality of the process, some scientists see high agreement as problematic: "Too much agreement is in fact a sign that the review process is *not* working well, that reviewers are not properly selected for diversity, and that some are redundant" (Bailar, 1991, p. 138). Many scientists see disagreement as a way of evaluating a contribution from a number of different perspectives. Although selecting

reviewers according to the principle of complementarity (for example, choosing a generalist and a specialist) will lower inter-reviewer agreement, validity can be increased considerably because the decision makers (such as journal editors or grant program managers) can base their decisions about a manuscript or proposal on much broader information. "This may reflect a deliberate strategy by editors to have the paper reviewed by specialists with different strengths, although both are usually experts in the substantive topic of the paper" (Rossiter, 2003, p. 86).

Researchers who see low reliability as helpful for the decision-making process believe that reviewers should be selected precisely because of their different perspectives, judgment criteria, and so on (Stricker, 1991). Scientific merit is multifaceted; Hackett and Chubin (2003) therefore regard it as crucial that a valid review of a grant application attend to each of the many elements that make up a good proposal. "The combined assessments of several diverse experts may be needed to achieve a rounded evaluation of a proposal. With a multifaceted proposal evaluated from several divergent perspectives, it is not surprising that inter-rater agreement may be low. Different experts might properly reach different judgments about the quality of the proposal when their particular area of concern is given central importance and evaluated through their particular set of epistemic lenses" (Hackett & Chubin, 2003, p. 20).

Weller (2002) notes that experts have engaged in considerable debate on the subject of proper statistical tests for reviewer agreement studies. Although widely used, the Cohen's Kappa coefficient is prone to numerous methodological problems. Feinstein and Cicchetti (1990, p. 543) found that for Kappa the observed proportion of agreement can be "drastically lowered by a substantial imbalance in the table's marginal totals either vertically or horizontally." Furthermore, the true level of existing agreement may be systematically underestimated because diverging recommendations generally reflect not only the degree of discord, the factor of interest, but also inter-individual differences, called the dislocational component (Eckes, 2004; Lienert, 1987). For example, differences in strictness or leniency in reviewers' judgments (reviewer A may be inclined to rate all manuscripts one level lower than reviewer B) cannot be taken into account (Jayasinghe, Marsh, & Bond, 2003; Siegelman, 1991). In addition to the application of more appropriate coefficients (Cicchetti & Feinstein, 1990) such as Brennan and Prediger's (1981) Kappa, modeling approaches for the analysis of rater agreement have been discussed (Schuster, 2002). The modeling strategy aims not only to summarize the overall judgment concordance but also to model the assignment process of articles to the judgment categories.

Hargens and Herting (1990) took up the idea of latent class agreement models for journal peer review in the 1990s. They are critical of the fact that reviewers' assessments implicitly assume that judgments vary along one latent scientific quality dimension; the authors note that this assumption can hardly be tested by calculating $\kappa$ or ICC. For this reason Hargens and Herting calculate row-column (RC) association models

(Goodman, 1984). Their investigation of five journals shows that in four cases one quality dimension accounts for the association in the reviewers' judgments. In contrast to the findings of the other studies on the reliability of peer review, Hargens and Herting report substantial statistical association among reviewers' judgments.

Bornmann, Mutz, and Daniel (2007b) stated the same hypothesis as Hargens and Herting (1990) in respect of fellowship peer review: For such reviews, does one quality dimension appear among reviewers? Using the data on applications for doctoral and postdoctoral fellowships that were assessed by means of a three-stage evaluation process, they tested the extent of the association between reviewers' recommendations (internal and external evaluation) and final decisions on fellowship applications by the BIF Board of Trustees. Using the RC association model, they showed that a single latent dimension was sufficient to account for the association between (internal and external) reviewers' recommendations and the fellowship award decision by the Board. Favorable ratings by the (internal and external) reviewers corresponded with favorable decisions by the Board (and vice versa). This indicates that the latent dimension underlying reviewers' recommendations and the Board's decisions reflects the merit of an application under evaluation. These results support Hargens and Herting's (1990) findings about the journal peer review system, although they contradict many other studies relating to the reliability of peer review.

Many general approaches are available to evaluate reliability, identify multiple sampling strategies, and choose which of many statistical tests to apply. The identification of appropriate testing assumptions and of appropriate tests appears to be critical to a sound reliability analysis.

### *Fairness of Peer Review*

According to Merton (1942), the functional goal of science is the expansion of potentially true and secure knowledge. The norm of universalism prescribes that the evaluation of scientific contributions should be based upon objective scientific criteria. Journal submissions or grant applications are not supposed to be judged on the attributes of the author/applicant or on personal biases of the reviewer, editor, or program manager (Ziman, 2000). "First, universalism requires that when a scientist offers a contribution to scientific knowledge, the community's assessment of the validity of that claim should not be influenced by personal or social attributes of the scientist. … Second, universalism requires that a scientist be fairly rewarded for contributions to the body of scientific knowledge. … Particularism, in contrast, involves the use of functionally irrelevant characteristics, such as sex and race, as a basis for making claims and gaining rewards in science" (Long & Fox, 1995, p. 46). To the degree that particularism influences how claims are made and rewards are gained, the fairness of the peer review process is at risk (Godlee & Dickersin, 2003). Ever since Kuhn (1962) discussed the significance of

different scientific or paradigmatic views for the evaluation of scientific contributions in his seminal work *The Structure of Scientific Revolutions* (see also Mallard, Lamont, & Guetzkow, 2009), researchers, in particular proponents of social constructivism, have expressed increasing doubts about the norm-governed, objective evaluation of scientific work (Hemlin, 1996). For Cole (1992), the constructivists' research supports a new view of science, which casts doubt on the existence of a set of rational criteria. According to Sismondo (1993, p. 548), social constructivist research has brought about the recognition that "social objects in science exist and act as causes of, and constraints on, scientists' actions." Because reviewers are human, factors influence their writing that cannot be predicted, controlled, or standardized (Shashok, 2005). Hames (2007), Hemlin (1996), and Ziman (2000) have indicated that reviewers' minds cannot be completely cleansed of individual interests and emotional factors and that their actions can lapse into bad behavior. Ziman (2000) finds the influence of non-cognitive factors on academic cycles of credibility and credit quite substantial in the more highly structured social world of post-academic science.

Proponents of Mertonian sociology of science have responded to the claim of social constructivists that Merton (1942) overestimates the effects of norms on research with the counterclaim that Merton's work "was later misinterpreted by members of the 'constructivist' school, who thought that Merton was arguing that this is how science actually is" (Cole, 2004, p. 839). Norms only affirm ideals; they do not describe realities. The moral order in any social system always involves a tense balance between its norms and the corresponding counter-norms. It would make no sense to insist on universalism if particularism were not a tempting alternative (Ziman, 2000).

Explicitly or implicitly, many researchers have studied biases in peer review against the backdrop of the norm of universalism (Mertonian sociology of science) or the existence of social objects in science as causes of, and constraints on, scientists' actions (social constructivism). Biases enter peer review when factors that are independent of the quality of a submission (manuscript or application) and are functionally irrelevant to the research correlate statistically with the judgment of reviewers or the decision of journals or funding organizations (Marsh, Jayasinghe, & Bond, 2008; Weller, 2002), For example, consideration of features such as:

- Authors' or applicants' academic status
- Research reported in a submission (e.g., the sub-field or the statistical significance of results)
- The reviewers' gender
- The peer review process, such as the practice of reviewing the applications in alphabetic order (Bornmann & Daniel, 2005a) or (as a recent study by Johnson [2008] on the system used

> by National Institutes of Health [NIH] has shown) the undue weight of the few reviewers who have read a proposal on the full committee's decision (see also Russo, 2008)

Grayson (2002, p. 8) holds the most fundamental charge of bias in peer review is that of "conservatism, or discrimination against innovatory, heretical, or dissenting opinions. This is in line with a Kuhnian [1962] view of scientific progress in which knowledge is shaped in line with conventional wisdom until the discrepancies between 'normal science' and observed reality become too glaring to ignore" (see also Stehbens, 1999).

Overviews of the peer review research literature (e.g., Hojat, Gonnella, & Caelleigh, 2003; Martin, 2000; Owen, 1982; Pruthi, Jain, Wahid, Mehra, & Nabi, 1997; Ross, 1980; Sharp, 1990; Wood & Wessely, 2003) have named up to 25 different sources of bias that can potentially compromise the fairness of the peer review process. These can be divided into sources closely related to the research (e.g., the reputation of the scientific institution to which an applicant belongs) and irrelevant to the research (e.g., author's gender or nationality). However, source of bias closely related to research may not in fact be functionally irrelevant in peer review. Such sources can contain important information for the reviewer. Biases can also be divided by valence into positive or negative: "that is, a bias may lead to a more negative evaluation of an article than the referee would give were it not for the biasing factor, or it may lead to a more positive evaluation, in which case we may speak of a 'halo effect,' whereby the quality of a work is exaggerated upward by the appraiser. Or the bias may make no difference: a biased evaluation might be identical with what an unbiased evaluation would yield" (Shatz, 2004, p. 36).

Godlee and Dickersin (2003, pp. 91–92) designate whichever biases lead to unfairness in peer review as bad biases, which they distinguish from good biases:

> By good biases we mean, for example, those in favour of important, original, well designed, and well reported science. Most editors and peer reviewers take these good biases for granted, as part of their responsibility to advance a particular field and to meet readers' needs. By bad biases we mean those that reflect a person's pre-existing views about the source of a manuscript (its authors, their institutions, or countries of origin) or about the ideas or findings it presents. Whether held by editors or peer reviewers, bad biases mean that decisions may be systematically influenced by factors other than a manuscript's methodological quality or relevance to the journal's readers.

Empirical research on peer review has dedicated itself to the investigation of potential sources of bias principally in two ways. On one hand,

surveys ask scientists to estimate the fairness of particular peer review processes. As in the following examples, surveys often reveal that many grant applicants and authors have concerns about a lack of fairness in peer review. For the National Institute of Handicapped Research (now National Institute on Disability and Rehabilitation Research, NIDRR), a survey of applicants' opinions found that 41 percent of respondents did not agree with the following statement: "I think that as a whole, the peer reviewer comments were fair" (Fuhrer & Grabois, 1985, p. 319). In a survey of applicants for grants from the National Cancer Institute (NCI), 40 percent of applicants found that reviewers were biased against researchers in minor universities or institutions in certain regions of the U.S. (Gillespie, Chubin, & Kurzon, 1985).

Applicants to the Australian Research Council (ARC) claimed that "there is a strong element of luck or chance in being awarded a Large Grant, the outcome of applications depends too much on whom the ARC panel selects as assessors, panels favour applicants from some universities, there is an element of cronyism in the ARC evaluation processes, knowing panel members is a significant advantage in being funded under the Large Grants Scheme, and panel members are themselves advantaged when applying for a Large Grant" (Over, 1996, p. 25). A survey of the *Academy of Management Journal* and *Academy of Management Review* found that about one in ten respondents was dissatisfied with reviewers' objectivity (Bedeian, 2003). Further surveys showed similar results regarding the perceived fairness of peer review; these were published by Chubin and Hackett (1990), McCullough (1989, 1994), and Resnik, Gutierrez-Ford, and Peddada (2008). Predictably there is a close connection between satisfaction with the peer review process and both one's own success in a review and one's participation in a satisfaction survey: Successful applicants and authors are more satisfied and more often take part in surveys than unsuccessful applicants (Gillespie et al., 1985).

In addition to the surveys of grant applicants and authors, the research on the fairness of peer review has used process-generated data to examine the influence of bias variables on judgments. The total group of applicants, authors, or reviewers is divided into subgroups to see whether they consistently and systematically differ from each other in their judgments (e.g., differences in the mean ratings of male and female reviewers). This kind of research is as a rule very elaborate and costly because: (1) research has identified a large number of attributes (of applicants, authors, reviewers, etc.) that can represent potential sources of bias in the peer review process (as has been discussed); (2) the study design should meet the highest standards in order to establish unambiguously that the work from a particular group of applicants or authors has a higher rejection rate due to biases in the peer review process and not simply as a consequence of the lesser scientific merit of the group of applications or manuscripts; and (3) the peer review

process is a secret activity (Tight, 2003) and reviews are secured with a guarantee of confidentiality.

Studies that have investigated systematic differences in the reviews of various subgroups primarily concern potential biases, which are attributed to specific features of applicants or authors. Features of reviewers, the peer review processes, or the research reported in a submission receive substantially less attention. Two studies are considered classics in the research on fairness: in a highly cited study of the reviewing process as practiced by scholarly journals, Zuckerman and Merton (1971a, 1971b) showed that the academic status of the author influences the probability that a manuscript will be accepted for publication. The results of the study by Wennerås and Wold (1997) on the fellowship peer review of the Swedish Medical Research Council, which appeared in the journal *Nature*, strongly suggest that reviewers cannot judge scientific merit independently of an applicant's gender. For Cole (1992) the results of such peer review studies suggest that the evaluation of scientific work is influenced by a complex interaction between universalistic factors, such as scientific merit, and scientific and non-scientific particularistic factors, such as gender. Given these findings, Cole (1992) assumes that there is no way to evaluate new scientific work objectively.

Even though numerous studies have reported a lack of fairness in the peer review process, a fundamental problem makes the findings difficult to generalize. Empirical studies have produced inconsistent findings. For example, some studies report that women scientists are at a disadvantage. However, a similar number of studies report only moderate gender effects, no effects, or mixed results (see Table 5.2). Inconsistent results also arise in the research on bias relating to an author's institutional affiliation, the second most frequently researched bias in the context of journal peer review. Some studies provide statistical evidence of discrimination against researchers coming from less prestigious institutions (e.g., Epstein, 1990). Other studies, however, have found little statistical evidence of systematic institutional bias (Lock, 1985; McIntosh & Ross, 1987) or describe contradictory results (Garfunkel, Ulshen, Hamrick, & Lawson, 1994; Pfeffer, Leong, & Strehl, 1977). Weller (2002) recommends more research on this particular topic in the future.

Meta-analysis could be a promising avenue for peer review research. The term meta-analysis (Glass, 1976) refers to a statistical approach that combines evidence from different studies to obtain an overall estimate of treatment effects (Shadish, Cook, & Campbell, 2002). Meta-analysis allows generalized statements on the strength of the effects, regardless of the specificity of individual studies (Matt & Navarro, 1997). It is necessary, of course, that each study be designed similarly with regard to certain properties (e.g., methods, sampling). Bornmann, Mutz, and Daniel (2007a) have presented the first meta-analysis in peer review research on one of the most frequently researched features, the gender of applicants. They were able to include in the meta-analysis data on proportions of women and men for 66 different sets of results

**Table 5.2 Results of empirical studies on gender effects in journal peer review**

| Study results indicate a gender effect | Study results indicate no gender effect | Study results on gender effect are mixed |
|---|---|---|
| **Author's gender** | | |
| Petty & Fleming (1999) | Caelleigh, Hojat, Steinecke, & Gonnella (2003) | Tregenza (2002) |
| Lloyd (1990) | Gilbert, Williams, & Lundberg (1994) | Levenson, Burford, Bonno, & Davis (1975) |
| Paludi & Bauer (1983) | Blank (1991) | |
| Paludi & Strayer (1985) | Patterson, Bailey, Martinez, & Angel (1987) | |
| Sahner (1982) | Bernard (1980) | |
| Ward (1981) | Lee, Boyd, Holroyd-Leduc, Bacchetti, & Bero (2006) | |
| Ferber & Teiman (1980) | | |
| Goldberg (1968) | | |
| **Reviewer's gender** | | |
| Blank (1991) | Davo, Vives, & Alvarez-Dardet (2003) | |
| Lloyd (1990) | Gilbert, Williams, & Lundberg (1994) | |
| Paludi & Bauer (1983) | Nylenna, Riis, & Karlsson (1994) | |
| Paludi & Strayer (1985) | | |
| Ward (1981) | | |
| Levenson, Burford, Bonno, & Davis (1975) | | |

from 21 studies since the 1980s. The studies included findings relating to research funding organizations in Australia, North America, and Europe in the life sciences, exact sciences, social sciences, and humanities fields. The data were analyzed using a generalized linear mixed model (Skrondal & Rabe-Hesketh, 2004).

The results show a statistically significant gender effect on the distribution of fellowship and grant proposals: "Even though the estimates of the gender effect vary substantially from study to study, the model estimation shows that all in all, among grant applicants men have statistically significant greater odds of receiving grants than women by about 7 percent" (Bornmann et al., 2007a, p. 226). What importance can be attributed to this finding on the use of peer review for selection of applicants to receive fellowships or grants? Assume for a moment that worldwide in a certain time period decisions are made to approve or reject 100,000 applications (50,000 submitted by women scientists and 50,000 by men scientists). Half of these applications are approved, and half are rejected. Based on the results of the meta-analysis, an approval rate of 52 percent (26,000 approvals) for men and 48 percent (24,000 approvals) for women can be expected. This makes a difference of 2,000 approvals due to the gender of the applicants (Bornmann et al., 2007a).

Bornmann and colleagues (2007a) reported substantial heterogeneity in effect sizes for the different peer review processes considered in the meta-analysis that compromised the robustness of their results, but models incorporating variables to explain this heterogeneity failed to converge. In an extension and reanalysis of their data, Marsh, Bornmann, Mutz, Daniel, and O'Mara (in press) juxtapose traditional (fixed and random effects) and multilevel models, demonstrating important advantages to the multilevel approach and its appropriate application. (Marsh and Bornmann [2009] reported on the results of the reanalysis in *Nature*.) The results show that gender differences in peer review varied significantly in relation to the type of application (grant proposals or pre- and post-doctoral fellowships), but not in relation to the other covariates (e.g., publication year of the studies included in the meta-analysis). Gender differences in favor of men are larger for fellowship applications than for grant applications. Indeed, there are no gender differences for the 40 (of 66) sets of results based on grant proposals, but statistically significant gender differences in favor of men for the 26 sets of results that were for fellowship applications. These findings support Cole's (1979, p. 75) conclusion that "functionally irrelevant characteristics such as sex will be more quickly activated when there are no or few functionally relevant criteria on which to judge individual performance." The younger the applicants are, the shorter their track records. And the less information there exists on which to judge an application, the greater the risk of bias.

According to Marsh and colleagues (2008, p. 166) most peer review research is correlational; "it provides a weak basis for any causal inferences particularly in evaluating potential biases." These authors thus address a second principal problem, in addition to the heterogeneity of results, that affects bias research in general: the lack of experimental studies in which (1) an independent variable is varied or manipulated by an experimenter (e.g., applicant's gender), so that its effect on a dependent variable (e.g., selection decisions) may be measured and (2) the effect of the variable is monitored. Only very few researchers have attempted to study reviewer bias experimentally in the natural setting of actual reviewer evaluations (see Abramowitz, Gomes, & Abramowitz, 1975; Baxt, Waeckerle, Berlin, & Callaham, 1998; Mahoney, 1977; Nylenna, Riis, & Karlsson, 1994). Peters and Ceci (1982), for example, examined reviewers' evaluations of manuscripts submitted to American psychology journals in a natural setting (Duncan & Magnuson, 2003). They looked for reviewer bias that could be attributed to reviewers' knowledge of the authors' institutions or names. As test materials they selected already published research articles by investigators from prestigious and highly productive American psychology departments. With fictitious names and institutions substituted for the original ones, the altered manuscripts were formally resubmitted to the journals that had originally reviewed and published them. Eight of the nine altered articles that were peer reviewed a second time were rejected. Peters and

Ceci's bias study was criticized, however, for having violated ethical principles (Chubin, 1982; Fleiss, 1982; Honig, 1982; Weller, 2002).

Experimental research has been undertaken with regard to fellowship peer review as well as journal peer review. Noting that female applicants have had a consistently lower success rate when applying for the European Molecular Biology Organization's (EMBO) Long-Term Fellowship (LTF) and Young Investigator (YI) programmes, Ledin, Bornmann, Gannon, and Wallon (2007) conducted an experiment to test whether unconscious gender bias influenced the decisions made by the selection committee. First, they gender-blinded the committee for two rounds of applications in 2006. They eliminated all references to gender from the applications, letters of recommendation, and interview reports that were sent to the committee for scoring. Nevertheless, the committee reached the same conclusions when gender-blinded. Although the studies of Ledin and colleagues (2007) on fellowship peer review and Peters and Ceci (1982) on journal peer review used an experimental approach, they found different results regarding the effect of bias variables.

The lack of experimental studies makes it impossible to establish unambiguously whether work from a particular group of scientists (e.g., junior or senior researchers) receives better reviews (and thus has a higher approval or acceptance rate) due to biases in the review and decision-making process, or if favorable review and greater success in the selection process are simply consequences of the scientific merit of the corresponding group of proposals or manuscripts (Daniel, 1993/2004). As with non-experimental studies, it is almost impossible to demonstrate certain biases in the decision-making process of peer review. Marsh and colleagues (2008, p. 166) note that if "it may be possible to construct artificial laboratory studies with true random assignment in which potential biases are experimentally manipulated, researchers need to be careful that experimental manipulations reflect the actual bias being tested and that results generalize to actual peer-review practice."

### *Predictive Validity of Peer Review*

The goal for peer review of grant/fellowship applications and manuscripts is usually to select the "best" from among the work submitted (Smith, 2006). The validity of judgments in peer review is often questioned. For example, the former editor of the journal *Lancet*, Sir Theodore Fox (1965, p. 8), wrote: "When I divide the week's contributions into two piles—one that we are going to publish and the other that we are going to return—I wonder whether it would make any real difference to the journal or its readers if I exchanged one pile for another." The selection function is a difficult research topic to investigate. According to Jayasinghe and colleagues (2001) and also Figueredo (2006) there exists no mathematical formula or uniform definition of what makes a manuscript worthy of publication, or of what makes a research proposal worthy of funding (see also Smith, 2006).

As a result of the increasing reliance on soft money in the financing of research and the use of publications as a measure of scientific performance (whether by individual researchers or groups of scientists) in nearly all scientific disciplines, the peer review system is faced with an ever-growing number of submissions (e.g., Gölitz, 2008). This trend creates new challenges. Previously, reviewers filtered out work that did not meet a certain minimum standard (negative selection), whereas today they need to select the "best" from a mass of high quality work (positive selection). Instead of minimum standards, today excellence is required (Koch, 2006). According to Yalow (1982, p. 401) the question for today's peer review is "how to identify the few, those who make the breakthroughs which permit new horizons to open, from the many who attempt to build on the breakthroughs—often without imagination and innovation."

One approach to testing the predictive validity of peer review is to investigate the fate of rejected scientific contributions. In journal peer review "a rejection usually does not kill a paper … a rejected paper usually finds life at another journal" (Gans & Shepherd, 1994, p. 177). In the 1980s Abelson (1980) reported that almost all of the manuscripts rejected by the journal *Science* were published later in other journals. For manuscripts rejected by the journal AC-IE in 1984, Daniel (1993/2004) determined that 71 percent were subsequently published elsewhere; Bornmann and Daniel (2008a, 2008b) found 95 percent of manuscripts the journal rejected at the beginning of 2000 were published later elsewhere. (The follow-up study determined that no alterations or only minor alterations had been made to approximately three-quarters of the rejected manuscripts published elsewhere). Other studies on the fate of manuscripts rejected by a journal report subsequent publication figures ranging from 28 percent to 85 percent (Weller, 2002). Taken together, these studies demonstrate that in the peer review process of one manuscript submitted at various times, reviewers and editors arrive at different judgments: manuscripts that are rejected by a journal (after peer review) are then accepted by another journal (after peer review). This suggests that manuscript review is not based only on *generally* valid quality criteria; the review (or the outcome of the review) seems also to be dependent upon the local conditions under which the peer review process takes place.

*Rejecta Mathematica* (Wakin, Rozell, Davenport, & Laska, 2009)—a peer-reviewed open access journal—was launched in 2009 and is the first journal to publish only manuscripts that have been previously rejected from peer-reviewed journals in the mathematical sciences. As studies concerning the fate of rejected manuscripts show, this journal is not alone in publishing rejected manuscripts. The novelty is that rejections are made public and openly discussed. All research papers appearing in *Rejecta Mathematica* include an open letter from the authors discussing the manuscript's original review process, disclosing any

known flaws in the manuscript and stating the case for the manuscript's value to the community.

Several studies have investigated the fate of rejected contributions in both grant and journal peer review. In a sample of projects rejected by the NIH at the beginning of the 1970s, 22 percent of proposals were subsequently carried out without substantial changes and 43 percent were not conducted after being rejected (Carter, Cooper, Lai, & March, 1978). Similarly, in a survey of applicants for grants from the NSF, 48 percent of researchers whose applications failed reported that they stopped that line of research (Chubin & Hackett, 1990). At the NCI two-thirds of unsuccessful applicants did pursue their research in spite of rejection and most were eventually published (Gillespie et al., 1985). In light of the paucity of studies on the fate of rejected applications,"further cohort studies of unfunded proposals are needed. Such studies will, however, always be difficult to interpret—do they show how peer review prevents resources from being wasted on bad science, or do they reveal the blinkered conservative preferences of senior reviewers who stifle innovation and destroy the morale of promising younger scientists? We cannot say" (Wessely, 1998, p. 303).

Following recommendations, such as those of Harnad (2008, p. 103), that "peer review … [has] to be evaluated objectively (i.e., via metrics)," the second step in the research on the predictive validity of journal peer review consists of gauging the quality of journals that accepted previously rejected manuscripts. According to Jennings (2006, online) "there is a hierarchy of journals. At the apex of the (power law-shaped?) pyramid stand the most prestigious multidisciplinary journals; below them is a middle tier of good discipline-specific journals with varying degrees of selectivity and specialization; and propping up the base lies a large and heterogeneous collection of journals whose purviews are narrow, regional, or merely unselective." In her literature review covering research on journal peer review, Weller (2002) cites five studies (Chew, 1991; Cronin & McKenzie, 1992; Gordon, 1984; Weller, 1996; Whitman & Eyre, 1985) that have ranked the quality of the rejecting and the later accepting journals mostly by means of the Journal Impact Factors (JIF, provided by Thomson Reuters in the Journal Citation Reports, JCR). The JIF is the average number of times papers from the journal published in the past two years (e.g., 2005 and 2006) have been cited in the JCR year (e.g., 2007) (Bensman, 2007).

Seven further studies, which are not included in the literature review by Weller (2002), have been published by Armstrong, Idriss, Kimball, and Bernhard (2008); Bornmann and Daniel (2008a, 2008b); Daniel (1993/2004); Lock (1985); McDonald, Cloft, and Kallmes (2007); Opthof, Furstner, van Geer, and Coronel (2000); and Ray, Berkwits, and Davidoff (2000). In the total of twelve studies, between 0 percent (Daniel, 1993/2004) and 70 percent (Gordon, 1984) of the rejected manuscripts in a higher quality journal could be tracked. The results of these studies show "that authors do not necessarily move from 'leading' journals to

less prestigious journals after a rejection" (Weller, 2002, p. 68). Authors seem to select a journal for a rejected manuscript based on the quality of the rejecting journal and the availability of additional high(er)-impact journals: the higher the quality ranking of the rejecting journal, the lower the chance that a rejected manuscript will appear in another journal ranked as higher quality.

Neither grant nor fellowship peer review has yet undergone similar analysis. At least no studies have compared the reputations of funding organizations, which rejected an application and/or funded a previously rejected application. This may have to do with the fact that the reputation of a grant-giving organization cannot be gauged as easily as the reputation of a journal (through its JIF or through circulation data).

A third, important step for the investigation of the predictive validity of peer review consists of comparing the impact of papers accepted or rejected (but published elsewhere) in journal peer review or the impact of papers that were published by applicants whose proposals were either accepted or rejected in grant or fellowship peer review. As the number of citations to a publication reflects its international impact (Borgman & Furner, 2002; Nicolaisen, 2007), and given the lack of other operationalizable indicators, it is common in peer review research to evaluate the success of the process on the basis of citation counts. Citation counts are attractive raw data for the evaluation of research output: They are "unobtrusive measures that do not require the cooperation of a respondent and do not themselves contaminate the response (i.e., they are nonreactive)" (Smith, 1981, p. 84). Although citations have been a controversial measure of both quality and scientific progress (e.g., scholars might cite because the cited source corroborated their own views or preferred methods, rather than because of the significance and relevance of the works cited), they are still accepted as a measure of scientific impact and thus as a partial aspect of scientific quality (Martin & Irvine, 1983).

Van Raan (1996, p. 404), holds that citations provide "a good to even very good quantitative impression of at least one important aspect of quality, namely international impact." For Lindsey (1989, p. 201), citations are "our most *reliable* convenient measure of quality in science—a measure that will continue to be widely used." For Pendlebury (2008, p. 1) "tracking citations and understanding their trends in context is a key to evaluating the impact and influence of research." Against the backdrop of these and similar statements (Borgman, 2007; British Academy, 2007; Evidence Ltd., 2007; Jennings, 2006), scientific judgments on submissions (manuscripts or applications) are said to show predictive validity in peer review research if the citation counts of manuscripts accepted for publication (or manuscripts published by accepted applicants) and manuscripts rejected by a journal but then published elsewhere (or manuscripts published by rejected applicants) differ statistically significantly.

Until now only a few studies have analyzed citation counts from individual papers as the basis for assessing predictive validity in peer

review. A literature search found only five empirical studies into the level of predictive validity associated with the journal peer review process. Research in this area is extremely labor-intensive because a validity test requires information regarding the fate of rejected manuscripts and their citation counts (Bornstein, 1991). The editor of the *Journal of Clinical Investigation* (Wilson, 1978) undertook his own investigation into the matter of predictive validity. Daniel (1993/2004) and Bornmann and Daniel (2008a, 2008b) investigated the peer review process of AC-IE, and Opthof and colleagues (2000) did the same for *Cardiovascular Research*. McDonald, Cloft, and Kallmes (2009) examined the predictive validity of editorial decisions at the *American Journal of Neuroradiology*. All five studies confirmed that the editorial decisions (acceptance or rejection) for the various journals indicated a rather high degree of predictive validity, using citation counts as validity criteria.

Wilson (1978) was able to show that the 306 manuscripts accepted for publication in the *Journal of Clinical Investigation* during the 1970s were cited more frequently in the four years after their appearance than the 149 rejected manuscripts that subsequently appeared in other journals. Daniel (1993/2004) and Bornmann and Daniel (2008a, 2008b) reported similar results for manuscripts that were submitted to AC-IE, as did Opthof and colleagues (2000) and McDonald and colleagues (2009) for manuscripts that were submitted to *Cardiovascular Research* and the *American Journal of Neuroradiology*, respectively. Bornmann and Daniel (2008b) not only conducted a comparison of average citation counts of accepted and rejected (but published elsewhere) manuscripts, they also compared the average citation counts of both groups with international scientific reference standards using (1) mean citation rates for the journal set provided by Thomson Reuters corresponding to the field "chemistry" and (2) specific reference standards that refer to the subject areas of *Chemical Abstracts* (Bornmann, Mutz, Neuhaus, & Daniel, 2008; Neuhaus & Daniel, 2009). The comparisons reveal that mean citation rates below baseline values were significantly less frequent for accepted manuscripts than for rejected manuscripts.

Only six studies on the assessment of citation counts as a basis of predictive validity in selection decisions in fellowship or grant peer review have been published in recent years, according to a literature search. These studies tested whether papers by applicants whose proposals for funding were approved were cited more frequently than papers by applicants whose proposals were rejected. Armstrong, Caverson, Adams, Taylor, and Olley's (1997) study of the Heart and Stroke Foundation of Canada (HSFC); Bornmann and Daniel's (2005c, 2006) research on the BIF; and Bornmann, Wallon, and Ledin's (2008) work on the EMBO confirmed the predictive validity of the selection decisions. However, Hornbostel, Böhmer, Klingsporn, Neufeld, and von Ins (2009) investigated the Emmy Noether Programme of the DFG and Melin and Danell (2006) studied the Swedish Foundation for Strategic Research; neither

group showed significant differences between the performance of accepted and rejected applicants. Van den Besselaar and Leydesdorff (2007) reported contradictory results regarding the Council for Social Scientific Research of the Netherlands Organisation for Scientific Research. Carter (1982) investigated the association between (1) assessments given by the reviewers for the NIH regarding applicants for research funding and (2) the number of citations obtained by journal articles produced with grant funding. This study showed that better votes in fact correlate with more frequent citations; however, the correlation coefficient was low. Unlike the clearer results for journal peer review, contradictory results emerge in research on fellowship or grant peer review: some studies confirm the predictive validity of peer review but others leave room for doubt.

Melin and Danell (2006) examined the publication histories of the top 8 percent of all applicants to the Swedish Foundation for Strategic Research and found only slight mean differences in scientific productivity between approved and rejected applicants. Similar results were reported by van den Besselaar and Leydesdorff (2007) when they compared in a second step of analysis the 275 successful applicants and the top performing 275 non-successful applicants (of all 911 non-successful applicants). Rejection decisions by the selection committees can be categorized as type I error (falsely drawn approval) or type II error (falsely drawn rejection). With type I error, the selection committee concluded that an applicant had the scientific potential for promotion (and was approved), when he or she actually did not (as reflected in an applicant's low scientific performance). With type II error, the selection committee concluded that an applicant did *not* have the scientific potential for promotion (and was rejected), when he or she actually did (as reflected in a high scientific performance).

Bornmann and Daniel (2007a) expanded on such approaches by determining the extent of type I and type II errors in the selection decisions of the BIF committee peer review. Approximately one third of the decisions to award a fellowship to an applicant showed a type I error, and about one third of the decisions not to award a fellowship to an applicant showed a type II error. In a similar manner, Bornmann, Mutz, and Daniel (2008) examined EMBO selection decisions (for two programs) to determine the extent of errors due to over-estimation (type I errors) and under-estimation (type II errors) of future scientific performance. The statistical analyses showed that between 26 and 48 percent of the decisions made to award or reject an application exhibited one or the other error type. It should be emphasized, of course, that even though the selection committees did not correctly estimate the applicants' performance, a statistically significant association emerged for both institutions between selection decisions and the applicants' subsequent scientific achievements.

Most studies on the predictive validity of journal, fellowship, and grant peer review rely on statistical methods, which strictly speaking

should not be applied to bibliometric data. (Future peer review studies should adopt sound statistical methods.) For example, citation impact differences between accepted manuscripts and manuscripts that were rejected but published elsewhere were determined on the basis of arithmetic means. As a rule, the distribution of citation counts for a larger number of publications is skewed to the right according to a power law (Joint Committee on Quantitative Assessment of Research, 2008). In the face of non-normal distributed citation data, the arithmetic mean value is not appropriate because it can give a distorted picture of the kind of distribution and "it is a rather crude statistic" (p. 2). Arithmetic mean graphics show primarily where publications with high citation counts can be found. Evidence Ltd. (2007, p. 10), a British firm that has specialized in the evaluation of research, therefore recommends "where bibliometric data must stand alone, they should be treated as distributions and not as averages."

Comparisons drawn between groups of papers (e.g., of accepted or rejected applicants) in terms of reserach impact are, according to Bornmann, Mutz, Neuhaus, and colleagues (2008), valid only if (1) the scientific impact of the groups are looked at by using box plots, Lorenz curves, and Gini coefficients to represent the distribution characteristics of data (in other words, going beyond the usual arithmetic mean value); (2) different reference standards are used to assess groups' impact and the appropriateness of the reference standards undergoes critical examination; and (3) the comparative analysis of the citation counts for publications takes into consideration that, in statistical analysis, citations are a function of many factors in addition to scientific quality, including number of co-authors; location of the authors; the prestige, language, and availability of the publishing journal; and the size of the citation window. Including these factors in the statistical analysis makes it possible to establish the adjusted covariation between selection decisions and citation counts.

Hornbostel and colleagues (2009) incorporated not only bibliometric data into their study investigating predictive validity of the DFG peer review process, but also additional indicators of career success. Their results thus show that "a gratifyingly high number of the approved individuals were appointed to a professorship or had found a career position in research that they themselves described as satisfactory, either during the programme itself or after the funding had ended" (p. 188). Especially in fellowship peer review the evaluation of career data provides a good complement to the bibliometric analyses and should thus be applied in studies on predictive validity for determining the effectiveness of the peer review process (e.g., Wellcome Trust, 2001). Yet further performance measures in addition to citation counts can be employed not only for research on grant and fellowship peer review, but also for journal peer review. For example, publishers record the number of times that a paper in electronic form is accessed, which means the information on how many readers have downloaded it can be used. Many publishers are

already members and vendors of COUNTER and provide the usage report "Number of Successful Full-Text Article Requests by Month and Journal" (see www.projectcounter.org). According to Perneger (2004) and Brody, Harnad, and Carr (2006), usage statistics can be used as early performance indicators for papers and authors.

## Conclusions

For many years the peer review process has been a target for criticism. It has been criticized especially "in relation to traditional psychological research criteria of reliability, validity, generalizability, and potential biases" (Marsh et al., 2008, p. 160). On the other hand, defenders of the system argue that only qualified specialists can properly judge cutting-edge research and that peer review is necessary to maintain and improve the quality of submissions. This review has offered an overview of research on peer review processes and the arguments used by proponents and opponents in recent years. The most important research results are summarily described in the next section, thematic areas are pointed out, and recommendations for the conduct of future research are given.

### *Reliability of Peer Review*

Most studies report a low level of agreement between reviewers' judgments on the basis of $\kappa$ and/or ICC. Daniel (1993/2004) can certainly prove, at least for the AC-IE manuscript reviewing, that broad discrepancies between reviewers' judgments are rare, even if the chance-corrected agreement (measured on the basis of $\kappa$) between the judgments appears negligible. Agreement between reviewers appears less when measured with $\kappa$ or ICC than it is in fact. Studies applying Hargens and Herting's (1990) RC association model to the analysis of reviewer agreement in journal and fellowship peer review report similar results. These studies show substantial association among reviewers' judgments. The dependence of results upon the statistical procedure raises an important issue for future research (Weller, 2002): What are the correct statistical procedures for reliability studies?

After reflecting on the abundance of studies reporting low inter-reviewer agreement, Weller (2002) considers a second issue significant: the meaning or importance of results of reliability studies. In Kostoff's (1995, p. 180) view, low agreement among reviewers is simply a response to the review of mediocre academic work:

> While a peer review can gain consensus on the projects and proposals that are either outstanding or poor, there will be differences of opinion on the projects and proposals that cover the much wider middle range. For projects or proposals in this middle range, their fate is somewhat more sensitive to

> the reviewers selected. If a key purpose of a peer review is to insure that the outstanding projects and proposals are funded or continued, and the poor projects are either terminated or modified strongly, then the capabilities of the peer review instrument are well matched to its requirements.

As has been thoroughly discussed, a low reviewer agreement is seen as beneficial to the review process (Bailar, 1991). According to Marsh and Ball (1991), when varied perspectives are represented in the selection of reviewers, *reliability* declines, whereas the *validity* of the process is significantly increased.

However, very few studies have investigated reviewer agreement with the aim of identifying the actual reasons behind reviewer disagreement (e.g., by carrying out comparative content analyses of reviewers' comment sheets). LaFollette (1992), for example, noted the scarcity of research on such questions as how reviewers apply standards and the specific criteria established for making a decision on a manuscript. In-depth studies that address these issues might prove to be fruitful avenues for future investigation (Weller, 2002). Such research should dedicate itself primarily to the dislocational component in reviewers' judgments as well as differences in the strictness or leniency of the judgments (Eckes, 2004; Lienert, 1987).

### *Fairness of Peer Review*

Although reviewers like to believe that they choose the "best" based on objective criteria, "decisions are influenced by factors—including biases about race, sex, geographic location of a university, and age—that have nothing to do with the quality of the person or work being evaluated" (National Academy of Sciences, 2006). Given that peers are not prophets, but ordinary human beings with their own opinions, strengths, and weaknesses (Ehses, 2004), a number of studies have examined potential sources of bias in peer review. Numerous studies have already shown an association between potential sources of bias and judgments in peer review and thus called into question the fairness of the process itself; however, research on bias faces two fundamental problems that make generalization of the findings difficult.

On one hand, the various studies have yielded heterogeneous results. Some have demonstrated the indisputable effects of potential sources of bias; others report moderate or slight effects. The heterogeneity of the results may possibly be explained by disparate definitions of the phenomena under investigation, or, alternatively, the application of different research designs and statistical procedures. With the application of meta-analyses in peer review research, however, Bornmann and colleagues (2007a) and Marsh and colleagues (in press) have introduced a promising possibility for arriving at *generalized* statements with a heterogeneous set of results on the effect of a potential source of bias,

*regardless* of the specificity of individual studies. Therefore, future research should rely more heavily on meta-analyses in order to find a definitive answer to the question regarding the effect of specific sources of bias in peer review processes.

A second principal problem that affects bias research in general is the lack of experimental studies. This makes it impossible to establish unambiguously whether work from a particular group of scientists receives better reviews due to biases in the review and decision-making process, or if favorable reviews and greater success in the selection process are simply a consequence of the scientific merit of the corresponding group of proposals or manuscripts. Therefore, according to Wessely (1998, p. 304), "randomised controlled trials are needed to assess the role of … sex and institutional bias. The absence of controlled trials in this area of scientific decision making is ironic." The conduct of experimental studies on peer review remains problematic and runs up against ethical limits.

In order to be able to arrive at results with greater accuracy and validity in the analysis of potential sources of bias in peer review, several suggestions for statistical analysis and/or research design are given. These suggestions should lead to significantly more robust results than heretofore (see also Weller, 2002).

- Cross-validation: The statistical analyses of fairness should be conducted not only for the group of first reviewers but also for the group of second (or further) reviewers of manuscripts or applications; the quality of the results can be tested using cross-validation (Efron & Gong, 1983): In statistics, cross-validation is the practice of partitioning a sample of data into subsamples, in this case the first and the second reviewers, such that analysis is initially performed on the first subsample and the second subsamples are retained "blind" for subsequent use in confirming and validating the initial analysis.

- Within-proposal analysis: To test the influence of potential sources of bias regarding the reviewer (e.g., his or her nationality) on judgments in peer review, a so-called within-proposal analysis can be undertaken. Jayasinghe and colleagues (2001, p. 353) analyzed reviewers' gender as a potential source of bias in the ARC peer review and conducted "a within-proposal analysis based on those proposals with at least one male external reviewer and at least one female external reviewer." The advantage of this statistical approach is that it controls for the many characteristics associated with a proposal (e.g., the applicant's gender). Differences in the mean ratings between both reviewer groups can be investigated with a statistical test for two related samples (e.g., the paired sample *t*-test).

- Multiple regression analysis: With multiple regression analysis, one can establish the association between a bias variable (which is included in the regression model as an independent variable) and the judgments in peer review (which constitute the dependent variable in the model), such that the scientific quality of the proposal or manuscript is controlled for (measured ex-post, for example, using bibliometric indicators). The indicator of scientific quality is also included in the model as an independent variable. This means the association between the bias variable and judgments is investigated, when manuscript quality is *statistically controlled*. In statistical bias analysis this procedure is called the control variable approach (Cole & Fiorentine, 1991).
- Interaction effects: Because attributes of the authors or applicants *and* attributes of the reviewers are potential sources of bias in peer review, both should always be included in the statistical analysis; the peer review process should be examined with regard to so-called interaction effects. An example of an interaction effect is when the regional origin of a submission influences the judgment of a U.S. reviewer but does not influence the judgment of a reviewer from another country (Bornmann & Daniel, 2007b; Jayasinghe et al., 2003).
- Multilevel regression analysis: Reviewers' ratings on proposals or manuscripts in a data set on peer review of a grant organization or a journal normally take the following form: (1) it has a multilevel structure, which means two or more reviewers for a journal are nested within manuscripts or two or more reviewers for a grant organization are nested within grant proposals. (2) the data structure is often cross-classified, which means the same reviewer reviews many submissions and one submission is reviewed by many reviewers. Calculating single-level regression models for these data is usually inappropriate (Jayasinghe et al., 2003). For Jayasinghe (2003, p. 341), an important finding from his investigation of the ARC peer review process is that "future peer review data need to be analysed using multilevel models with either categorical (e.g., accept, reject, fund, no fund for grant proposals) or continuous (e.g., assessor ratings for the quality of grant proposals) response variables."
- Case study: Biases in peer review might assume different configurations and exhibit different dynamics across different disciplines, professions, regions, time spans, funding organizations, journal editorial teams, and so forth. Accordingly, the case study method might be more heavily utilized to detect the possibly unique dynamics and

> configurations of bias in different settings, times, circumstances. Broad surveys and grand analyses appear to show too much inconsistency and to yield weak conclusions.

For research on journal peer review, Geisler (2001) has proposed that a journal's process should be studied continuously and that any evidence of bias in judgment should be brought to the attention of the editor for correction and modification. This naturally goes for research on both grant and fellowship peer review. According to Hojat, Gonnella, and Caelleigh (2003, p. 75), the controversy surrounding the peer review process demands "that the journal editors conduct periodic internal and external evaluations of their journals' peer review process and outcomes, with participation of reviewers, contributors, readers, and owners" to assure its integrity and fairness. In the most comprehensive review of research on editorial biases, Godlee and Dickersin (2003, p. 112) also conclude that "journals should continue to take steps to minimize the scope for unacceptable biases, and researchers should continue to look for them." In Weller's (2002, p. 100) view, the area of *editorial* bias is "ripe for more research." The investigations on potential sources of bias can help not only to minimize the scope of unacceptable biases, but also to prevent scientific misbehavior; Martinson, Anderson, Crain, and de Vries (2006, p. 51) report that "when scientists believe they are being treated unfairly they are more likely to behave in ways that compromise the integrity of science."

Investigations into the influence of bias variables on the peer review process are as a rule very elaborate and costly. Before a funding organization or a journal conducts an extensive evaluation study, it should therefore seek indications of the influence of potential sources of bias both in order to determine the necessity of an evaluation study and, if a necessity appears, then to identify the sources of bias that should be examined more closely. Bornmann, Mutz, and Daniel (2008) present a statistical method that program managers at a research funding organization or journal editors can use to obtain *initial indications* of potential sources of bias in their peer review process. To implement the method in grant peer review, the data required are the number of approved and the number of rejected applicants for grants among different groups (for example, women and men or natural scientists and social scientists). Editors of a journal require only data on specific manuscript groups for a number of years or a number of research fields. If these evaluations show an influence of bias variables in a peer review process, an in-depth evaluation study should be conducted.

Focusing on classical bias variables like an author's gender or an applicant's academic status in peer review studies can lead to the exclusion of other variables when such other variables are really more likely to produce unfair outcomes. The following merit further research:

- A high number of cited references in a submission appears to impress reviewers, perhaps more than a substantive or significant submission with a low number of cited references.
- Reviewers might tend to reward mathematical and statistical pyrotechnics irrespective of the significance of the contribution. Furthermore, do advanced statistical tests tend to be employed more than simple statistical tests when the simple test might be more appropriate?
- Is there a tendency for many reviewers to forget older contributions (particularly when their originators die—as in the case of Piaget's theories of child cognitive development) or to reward citations to current theories (or even academically fashionable theories) more than to old theories? Does knowledge (particularly inter-generational knowledge) tend to accumulate or to be cumulative (see Kuhn, 1962)?
- Is there a tendency for disciplines in traditional peer review processes to form intellectual silos (see Mallard et al., 2009) or editorial cliques?
- Does there exist a so called Oppenheim effect in scholarly publishing and if so, to what extent? Is external review just a formal exercise initiated by an editor because the publication decision is already clear to the editor? Gorman (2007) calls this phenomenon the Oppenheim effect.

### *Predictive Validity of Peer Review*

The few studies that have examined the predictive validity of journal peer review on the basis of citation counts confirm that peer review represents a quality filter and works as an instrument for the self-regulation of science. Seven studies have investigated the predictive validity of selection decisions on the basis of citation counts for fellowship or grant peer review. Unlike with journal peer review, these studies have provided mixed results: Some confirm the predictive validity of peer review; others raise doubts.

The variable results on fellowship and grant peer review reflect the fact that "funding decisions are inherently speculative because the work has not yet been done" (Stamps, 1997, p. 4). Whereas in journal peer review the *results* of the research are assessed, grant and fellowship peer review is principally an evaluation of the *potential* of the proposed research (Bornmann & Daniel, 2005b). Evaluating the application involves deciding whether the proposed research is significant, determining whether the specific plans for investigation are feasible, and evaluating the competence of the applicant (Cole, 1992). Fellowship or grant peer review—when compared to journal peer review—is perceived

as entailing a heightened risk for judgments and decisions with low predictive validity. Accordingly it is likely that studies on grant or fellowship peer review will have more difficulty confirming the predictive validity than studies on journal peer review.

The heterogeneous results can also be attributed to the widely differing designs and statistical procedures used. Some studies include all applicants in one cohort (or several cohorts) (e.g., Bornmann, Wallon, et al., 2008), whereas Melin and Danell (2006) examine only a highly select group of the "best" applicants among those accepted and rejected. The study by Carter (1982) includes only applicants with an approved proposal. In only two studies (Bornmann & Daniel, 2005c, 2006; van den Besselaar & Leydesdorff, 2007) did the researchers test for statistical significance. Only one study (Bornmann & Daniel, 2005c, 2006) applied discipline-specific reference standards (van Raan, 1999). This comparison with international scientific reference values revealed, for example, that (1) articles published by successful and non-successful applicants were cited considerably more often than the "average" publication in a certain (sub-)field, and (2) excellent research performance can be expected more from successful than non-successful applicants.

Such a comparison with international scientific reference values is an important step in the analysis of the predictive validity of peer review processes. If a study compares the impact of publications—not only by successful but also by non-successful applicants—with reference standards, the scientific performance of the total group of applicants for grants or fellowships can be determined. At an organization that claims to have a highly regarded grant program, even non-successful grant applicants should demonstrate research performance at a higher level—otherwise one cannot speak validly of this program's renown. (This certainly implies that applicants are familiar with a program's renown.) Highly regarded programs should have the ability to stimulate the "best" researchers in a certain discipline to submit their "best" applications.

On the other hand, as has been discussed, five studies on journal peer review have found a high degree of predictive validity through the comparison of mean citation rates for accepted manuscripts and rejected manuscripts published elsewhere. Cicchetti (1999) argued against this form of validity test, pointing out that papers accepted by journals may have been cited on average more frequently than those published elsewhere simply because they appeared in journals with a high impact factor. Higher citation rates are not necessarily the result of a paper's superior scientific quality; instead, they may just show the higher impact or higher visibility of a journal. Seglen (1994), however, reports that articles' citation counts do not seem to be detectably influenced by the status of the journals in which they are published. In spite of Cicchetti's (1999) criticism, this form of validity test for journal peer review decisions should produce useful results.

The limited comparability of accepted and rejected contributions in studies on the predictive validity of grant and fellowship peer review has

been noted. Chapman and McCauley (1994, p. 428) write: "Criterion data for rejected applicants are difficult to obtain and difficult to interpret, even when available; those accepted are no longer comparable to those rejected because the two groups have had different experiences." According to Sonnert (quoted in Bornmann & Daniel, 2005c), fellowships clearly have a dual function. They reward prior excellence (i.e., they are given to the "best" applicants, who are selected according to merit criteria), but they also afford the successful applicants resources that might enable them to do excellent scientific work in their future careers. If a study establishes significant performance advantages for accepted applicants but not rejected ones, one can argue—with reference to Merton's (1948) concept of self-fulfilling prophecy—that the funding organization gives the fellows such an advantage in training, prestige, self-confidence, and so on that they later become superior scientists because of the fellowship, not because they were particularly promising at the point of application. Rather than picking the "best" scientists, the selection committee might, in this view, create them (see also Cole & Cole, 1967; Hagstrom, 1965; Merton, 1968).

There is accordingly a circularity in the study designs for the analysis of predictive validity that should be considered in future studies investigating grant or fellowship peer review. To control, for example, in statistical analysis of the influence of funding on subsequent publication and citation numbers, information is needed on the funding of rejected research by investigating in particular the fate of the rejected applicants and their research projects (Bornmann, Wallon, et al., 2008).

A general weakness of research on the predictive validity of peer review is the lack of studies; significantly more have been published on the reliability and fairness of peer review than on predictive validity. Weller (2002) has addressed this lack in journal peer review, and both Wessely (1998) and Jayasinghe and colleagues (2001) have done so for fellowship and grant peer review. Although the number of studies published in recent years has increased, particularly on fellowship peer review, comprehensive research is still lacking. Nonprofit and for-profit funding organizations and journals might be studied to provide an enhanced view of the peer review process.

## Acknowledgments

This review of studies on journal, grant, and fellowship peer review is dedicated to my longstanding mentor Hans-Dieter Daniel, Professor at the ETH Zurich and Director of the Evaluation Office of the University of Zurich. The author wishes to express his gratitude to three anonymous reviewers for their helpful comments.

## References

Abelson, P. H. (1980). Scientific communication. *Science*, *209*(4452), 60–62.

Abramowitz, S. I., Gomes, B., & Abramowitz, C. V. (1975). Publish or politic: Referee bias in manuscript review. *Journal of Applied Social Psychology*, *3*(5), 187–200.

Abrams, P. A. (1991). The predictive ability of peer review of grant proposals: The case of ecology and the United States National Science Foundation. *Social Studies of Science*, *21*(1), 111–132.

Alberts, B., Hanson, B., & Kelner, K. L. (2008). Reviewing peer review. *Science*, *321*(5885), 15.

Armstrong, A. W., Idriss, S. Z., Kimball, A. B., & Bernhard, J. D. (2008). Fate of manuscripts declined by the *Journal of the American Academy of Dermatology*. *Journal of the American Academy of Dermatology*, *58*(4), 632–635.

Armstrong, J. S. (1997). Peer review for journals: Evidence on quality control, fairness, and innovation. *Science and Engineering Ethics*, *3*(1), 63–84.

Armstrong, P. W., Caverson, M. M., Adams, L., Taylor, M., & Olley, P. M. (1997). Evaluation of the Heart and Stroke Foundation of Canada Research Scholarship Program: Research productivity and impact. *Canadian Journal of Cardiology*, *13*(5), 507–516.

Bailar, J. C. (1991). Reliability, fairness, objectivity, and other inappropriate goals in peer review. *Behavioral and Brain Sciences*, *14*(1), 137–138.

Bakanic, V., McPhail, C., & Simon, R. J. (1987). The manuscript review and decision-making process. *American Sociological Review*, *52*(5), 631–642.

Bakanic, V., McPhail, C., & Simon, R. J. (1989). Mixed messages: Referees' comments on the manuscripts they review. *Sociological Quarterly*, *30*(4), 639–654.

Bartley, W. W. (1984). *The retreat to commitment* (2nd ed.). La Salle, IL: Open Court.

Baxt, W. G., Waeckerle, J. F., Berlin, J. A., & Callaham, M. L. (1998). Who reviews the reviewers? Feasibility of using a fictitious manuscript to evaluate peer reviewer performance. *Annals of Emergency Medicine*, *32*(3), 310–317.

Bedeian, A. G. (2003). The manuscript review process: The proper roles of authors, referees, and editors. *Journal of Management Inquiry*, *12*(4), 331–338.

Bedeian, A. G. (2004). Peer review and the social construction of knowledge in the management discipline. *Academy of Management Learning and Education*, *3*(2), 198–216.

Bensman, S. J. (2007). Garfield and the impact factor. *Annual Review of Information Science and Technology*, *41*, 93–155.

Bernard, H. R. (1980). Report from the editor. *Human Organization*, *39*(4), 366–369.

Bertout, C., & Schneider, P. (2004). Editorship and peer-review at A&A. *Astronomy & Astrophysics*, *420*, E1–E14.

Bhandari, M., Templeman, D., & Tornetta, P. (2004). Interrater reliability in grading abstracts for the Orthopaedic Trauma Association. *Clinical Orthopaedics and Related Research*, *423*, 217–221.

Biagioli, M. (2002). From book censorship to academic peer review. *Emergences*, *12*(1), 11–45.

Blank, R. M. (1991). The effects of double-blind versus single-blind reviewing: Experimental evidence from the *American Economic Review*. *American Economic Review*, *81*(5), 1041–1067.

Bohannon, R. W. (1986). Agreement among reviewers. *Physical Therapy*, *66*(9), 1431–1432.

Borgman, C. L. (2007). *Scholarship in the digital age: Information, infrastructure, and the internet*. Cambridge, MA: MIT Press.

Borgman, C. L., & Furner, J. (2002). Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology*, *36*, 3–72.

Bornmann, L., & Daniel, H.-D. (2005a). Committee peer review at an international research foundation: Predictive validity and fairness of selection decisions on post-graduate fellowship applications. *Research Evaluation*, *14*(1), 15–20.

Bornmann, L., & Daniel, H.-D. (2005b). Criteria used by a peer review committee for selection of research fellows: A Boolean probit analysis. *International Journal of Selection and Assessment*, *13*(4), 296–303.

Bornmann, L., & Daniel, H.-D. (2005c). Selection of research fellowship recipients by committee peer review: Analysis of reliability, fairness and predictive validity of Board of Trustees' decisions. *Scientometrics*, *63*(2), 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*(3), 427–440.

Bornmann, L., & Daniel, H.-D. (2007a). Convergent validation of peer review decisions using the *h* index: Extent of and reasons for type I and type II errors. *Journal of Informetrics*, *1*(3), 204–213.

Bornmann, L., & Daniel, H.-D. (2007b). Gatekeepers of science: Effects of external reviewers' attributes on the assessments of fellowship applications. *Journal of Informetrics*, *1*(1), 83–91.

Bornmann, L., & Daniel, H.-D. (2008a). The effectiveness of the peer review process: Inter-referee agreement and predictive validity of manuscript refereeing at *Angewandte Chemie*. *Angewandte Chemie International Edition*, *47*(38), 7173–7178.

Bornmann, L., & Daniel, H.-D. (2008b). 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*(11), 1841–1852.

Bornmann, L., & Daniel, H.-D. (2009). The luck of the referee draw: The effect of exchanging reviews. *Learned Publishing*, *22*(2), 117–125.

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

Bornmann, L., Mutz, R., & Daniel, H.-D. (2007a). Gender differences in grant peer review: A meta-analysis. *Journal of Informetrics*, *1*(3), 226–238.

Bornmann, L., Mutz, R., & Daniel, H.-D. (2007b). Row-column (RC) association model applied to grant peer review. *Scientometrics*, *73*(2), 139–147.

Bornmann, L., Mutz, R., & Daniel, H.-D. (2008). How to detect indications of potential sources of bias in peer review: A generalized latent variable modeling approach exemplified by a gender study. *Journal of Informetrics*, *2*(4), 280–287.

Bornmann, L., Mutz, R., & Daniel, H.-D. (2009). *A reliability-generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants*. (Submitted for publication)

Bornmann, L., Mutz, R., Neuhaus, C., & Daniel, H.-D. (2008). Use of citation counts for research evaluation: Standards of good practice for analyzing bibliometric data and presenting and interpreting results. *Ethics in Science and Environmental Politics*, *8*, 93–102.

Bornmann, L., Nast, I., & Daniel, H.-D. (2008). Do editors and referees look for signs of scientific misconduct when reviewing manuscripts? A quantitative content analysis of studies that examined review criteria and reasons for accepting and rejecting manuscripts for publication. *Scientometrics*, *77*(3), 415–432.

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*(10), e3480.

Bornstein, R. F. (1991). The predictive validity of peer-review: A neglected issue. *Behavioral and Brain Sciences*, *14*(1), 138–139.

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, *41*(3), 687–699.

British Academy (2007). *Peer review: The challenges for the humanities and social sciences*. London: The Academy.

Brody, T., Harnad, S., & Carr, L. (2006). Earlier web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science and Technology*, *57*(8), 1060–1072.

Caelleigh, A. S., Hojat, M., Steinecke, A., & Gonnella, J. S. (2003). Effects of reviewers' gender on assessments of a gender-related standardized manuscript. *Teaching and Learning in Medicine*, *15*(3), 163–167.

Campanario, J. M. (1998a). Peer review for journals as it stands today: Part 1. *Science Communication*, *19*(3), 181–211.

Campanario, J. M. (1998b). Peer review for journals as it stands today: Part 2. *Science Communication*, *19*(4), 277–306.

Carter, G. (1982). *What we know and do not know about the peer review system* (Rand Report N-1878-RC/NIH). Santa Monica, CA: RAND Corporation.

Carter, G., Cooper, W., Lai, C., & March, D. (1978). *The consequences of unfunded NIH applications for the investigator and his research* (Rand Report R-2229-NIH). Santa Monica, CA: RAND Corporation.

Chapman, G. B., & McCauley, C. (1994). Predictive validity of quality ratings of National Science Foundation graduate fellows. *Educational and Psychological Measurement*, *54*(2), 428–438.

Chew, F. S. (1991). Fate of manuscripts rejected for publication in the AJR. *American Journal of Roentgenology*, *156*(3), 627–632.

Chubin, D., & Hackett, E. (1990). *Peerless science: Peer review and U.S. science policy*. Albany, NY: State University of New York Press.

Chubin, D. E. (1982). Reforming peer-review: From recycling to reflexivity. *Behavioral and Brain Sciences*, *5*(2), 204.

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

Cicchetti, D. V. (1997). Referees, editors, and publication practices: Improving the reliability and usefulness of the peer review system. *Science and Engineering Ethics*, *3*(1), 51–62.

Cicchetti, D. V. (1999). Guardians of science: Fairness and reliability of peer review. *Journal of Clinical and Experimental Neuropsychology*, *21*(3), 412–421.

Cicchetti, D. V., & Conn, H. O. (1976). A statistical analysis of reviewer agreement and bias in evaluating medical abstracts. *Yale Journal of Biology and Medicine*, *49*(4), 373–383.

Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low Kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 551–558.

Cole, J. R. (1979). *Fair science: Women in the scientific community*. New York: The Free Press.

Cole, J. R. (2000). The role of journals in the growth of scientific knowledge. In B. Cronin & H. B. Atkins (Eds.), *The web of knowledge: A Festschrift in honor of Eugene Garfield* (pp. 109–142). Medford, NJ: Information Today, Inc.

Cole, S. (1992). *Making science. Between nature and society*. Cambridge, MA: Harvard University Press.

Cole, S. (2004). Merton's contribution to the sociology of science. *Social Studies of Science*, *34*(6), 829–844.

Cole, S., & Cole, J. R. (1967). Scientific output and recognition: A study in operation of reward system in science. *American Sociological Review*, *32*(3), 377–390.

Cole, S., Cole, J. R., & Simon, G. A. (1981). Chance and consensus in peer-review. *Science*, *214*(4523), 881–886.

Cole, S., & Fiorentine, R. (1991). Discrimination against women in science: The confusion of outcome with process. In H. Zuckerman, J. R. Cole, & J. T. Bruer (Eds.), *The outer circle: Women in the scientific community* (pp. 205–226). London: W. W. Norton.

Cronin, B. (2005). *The hand of science: Academic writing and its rewards*. Lanham, MD: Scarecrow Press.

Cronin, B., & McKenzie, G. (1992). The trajectory of rejection. *Journal of Documentation*, *48*(3), 310–317.

Daniel, H.-D. (1993/2004). *Guardians of science: Fairness and reliability of peer review*. Weinheim, Germany: Wiley-VCH. Retrieved July 16, 2004, from DOI: 10.1002/3527602208

Davo, M. D., Vives, C., & Alvarez-Dardet, C. (2003). Why are women underused in the JECH peer review process? *Journal of Epidemiology and Community Health*, *57*(12), 936–937.

Demicheli, V., & Pietrantonj, C. (2007). Peer review for improving the quality of grant applications. *Cochrane Database of Systematic Reviews*, Issue 2. Art. No.: MR000003. DOI: 10.1002/14651858.MR000003.pub2

Duncan, G. J., & Magnuson, K. A. (2003). The promise of random-assignment social experiments for understanding well-being and behavior. *Current Sociology*, *51*(5), 529–541.

Eckberg, D. L. (1991). When nonreliability of reviews indicates solid science. *Behavioral and Brain Sciences*, *14*(1), 145–146.

Eckes, T. (2004). Rater agreement and rater severity: A many-faceted Rasch analysis of performance assessments in the "Test Deutsch als Fremdsprache" (TestDaF). *Diagnostica*, *50*(2), 65–77.

Editor to quit over hoax open-access paper. (2009). *Nature*, *459*, 901.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician*, *37*(1), 36–48.

Ehses, I. (2004). By scientists, for scientists. The Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) and how it functions. *B.I.F. Futura*, *19*, 170–177.

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

Epstein, W. M. (1990). Confirmational response bias among social-work journals. *Science, Technology, & Human Values*, *15*(1), 9–38.

Ernst, E., Saradeth, T., & Resch, K. L. (1993). Drawbacks of peer review. *Nature*, *363*(6427), 296.

Evidence Ltd. (2007). *The use of bibliometrics to measure research quality in UK higher education institutions*. London: Universities UK.

Eysenck, H. J., & Eysenck, S. B. G. (1992). Peer review: Advice to referees and contributors. *Personality and Individual Differences*, *13*(4), 393–399.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low Kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 543–549.

Feist, G. J. (2006). *The psychology of science and the origins of the scientific mind*. New Haven, CT: Yale University Press.

Ferber, M. A., & Teiman, M. (1980). Are women economists at a disadvantage in publishing journal articles? *Eastern Economic Journal*, *6*(3–4), 189–193.

Figueredo, E. (2006). The numerical equivalence between the impact factor of journals and the quality of the articles. *Journal of the American Society for Information Science and Technology*, *57*(11), 1561.

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*(5), 591–598.

Fleiss, J. (1981). *Statistical methods for rates and proportions*. New York: Wiley VCH.

Fleiss, J. L. (1982). Deception in the study of the peer-review process. *Behavioral and Brain Sciences*, *5*(2), 210–211.

Fogg, L., & Fiske, D. W. (1993). Foretelling the judgments of reviewers and editors. *American Psychologist*, *48*(3), 293–294.

Fox, T. (1965). *Crisis in communication: The functions and future of medical publication*. London: Athlone Press.

Frey, B. S. (2003). Publishing as prostitution? Choosing between one's own ideas and academic success. *Public Choice*, *116*(1–2), 205–223.

Fuhrer, M. J., & Grabois, M. (1985). Grant application and review procedures of the National Institute of Handicapped Research: Survey of applicant and peer reviewer opinions. *Archives of Physical Medicine and Rehabilitation*, *66*(5), 318–321.

Gans, J. S., & Shepherd, G. B. (1994). How are the mighty fallen: Rejected classic articles by leading economists. *Journal of Economic Perspectives*, *8*(1), 165–179.

Garfield, E. (2004). Historiographic mapping of knowledge domains literature. *Journal of Information Science*, *30*(2), 119–145.

Garfunkel, J. M., Ulshen, M. H., Hamrick, H. J., & Lawson, E. E. (1994). Effect of institutional prestige on reviewers' recommendations and editorial decisions. *Journal of the American Medical Association*, *272*(2), 137–138.

Geisler, E. (2000). *The metrics of science and technology*. Westport, CT: Quorum Books.

Geisler, E. (2001). The mires of research evaluation. *The Scientist*, *15*(10), 39.

Gibson, M., Spong, C. Y., Simonsen, S. E., Martin, S., & Scott, J. R. (2008). Author perception of peer review. *Obstetrics and Gynecology*, *112*(3), 646–651.

Gilbert, J. R., Williams, E. S., & Lundberg, G. D. (1994). Is there gender bias in *JAMA*'s peer review process? *Journal of the American Medical Association*, *272*(2), 139–142.

Gillespie, G. W., Chubin, D. E., & Kurzon, G. M. (1985). Experience with NIH peer review: Researchers' cynicism and desire for change. *Science, Technology, & Human Values*, *52*, 44–54.

Glass, G. V. (1976). Primary, secondary, and meta-analysis. *Educational Researcher*, *5*, 3–8.

Godlee, F., & Dickersin, K. (2003). Bias, subjectivity, chance, and conflict of interest. In F. Godlee & J. Jefferson (Eds.), *Peer review in health sciences* (2nd ed., pp. 91–117). London: BMJ Publishing Group.

Goldberg, P. (1968). Are women prejudiced against women? *Transactions*, *5*(5), 28–30.

Gölitz, P. (2008). Appeals. *Angewandte Chemie International Edition*, *47*(38), 7144–7145.

Goodman, L. A. (1984). *The analysis of cross-classified data having ordered categories*. Cambridge, MA: Harvard University Press.

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*(1), 11–21.

Gordon, M. D. (1984). How authors select journals: A test of the reward maximization model of submission behavior. *Social Studies of Science*, *14*(1), 27–43.

Gorman, G. E. (2007). The Oppenheim effect in scholarly journal publishing. *Online Information Review*, *31*(4), 417–419.

Gosden, H. (2003). "Why not give us the full story?": Functions of referees' comments in peer reviews of scientific research papers. *Journal of English for Academic Purposes*, *2*(2), 87–101.

Grayson, L. (2002). *Evidence based policy and the quality of evidence: Rethinking peer review*. London: Centre for Evidence Based Policy and Practice (ESRC).

Guston, D. H. (2003). The expanding role of peer review processes in the United States. In P. Shapira & S. Kuhlmann (Eds.), *Learning from science and technology policy evaluation: Experiences from the United States and Europe* (pp. 81–97). Cheltenham, UK: Edward Elgar.

Hackett, E. J., & Chubin, D. E. (2003, February). *Peer review for the 21st century: Applications to education research*. Paper presented at the conference Peer Review of Education Research Grant Applications: Implications, Considerations, and Future Directions, Washington, DC.

Hagstrom, W. O. (1965). *The scientific community*. New York: Basic Books.

Hames, I. (2007). *Peer review and manuscript management of scientific journals: Guidelines for good practice*. Oxford, UK: Blackwell.

Hansson, F. (2002). How to evaluate and select new scientific knowledge? Taking the social dimension seriously in the evaluation of research quality. *Vest*, *15*(2–3), 27–52.

Hargens, L. L., & Herting, J. R. (1990). A new approach to referees assessments of manuscripts. *Social Science Research*, *19*(1), 1–16.

Harnad, S. (1990). Scholarly skywriting and the prepublication continuum of scientific inquiry. *Psychological Science*, *1*(6), 342–344.

Harnad, S. (1996). Implementing peer review on the net: Scientific quality control in scholarly electronic journals. In P. R. Peek & G. B. Newby (Eds.), *Scholarly publishing: The electronic frontier* (pp. 103–118). Cambridge, MA: MIT Press.

Harnad, S. (2008). Validating research performance metrics against peer rankings. *Ethics in Science and Environmental Politics*, *8*, 103–107.

Hemlin, S. (1996). Research on research evaluations. *Social Epistemology*, *10*(2), 209–250.

Hemlin, S., & Rasmussen, S. B. (2006). The shift in academic quality control. *Science, Technology, & Human Values*, *31*(2), 173–198.

Hodgson, C. (1997). How reliable is peer review? An examination of operating grant proposals simultaneously submitted to two similar peer review systems. *Journal of Clinical Epidemiology*, *50*(11), 1189–1195.

Hoffmann, H., Joye, D., Kuhn, F., & Métral, G. (2002). *Der SNF im Spiegel der Forschenden: Synthesebericht*. Neuchâtel, Switzerland: Schweizerischer Informations und Datenarchivdienst für die Sozialwissenschaften (SIDOS).

Hojat, M., Gonnella, J. S., & Caelleigh, A. S. (2003). Impartial judgment by the "gatekeepers" of science: Fallibility and accountability in the peer review process. *Advances in Health Sciences Education*, *8*(1), 75–96.

Honig, W. M. (1982). Peer review in the physical sciences: An editor's view. *Behavioral and Brain Sciences*, *5*(2), 216–217.

Hornbostel, S. (1997). *Wissenschaftsindikatoren: Bewertungen in der Wissenschaft*. Opladen, Germany: Westdeutscher Verlag.

Hornbostel, S., Böhmer, S., Klingsporn, B., Neufeld, J., & von Ins, M. (2009). Funding of young scientist and scientific excellence. *Scientometrics*, *79*(1), 171–190.

Hornbostel, S., & Olbrecht, M. (2007). *Peer Review in der DFG: Die Fachkollegiaten* (iFQ-Working Paper No. 2). Bonn, Germany: Institut für Forschungsinformation und Qualitätssicherung.

Ingelfinger, F. J. (1974). Peer review in biomedical publication. *American Journal of Medicine*, *56*(5), 686–692.

Jayasinghe, U. W. (2003). *Peer review in the assessment and funding of research by the Australian Research Council.* Greater Western Sydney, Australia: University of Western Sydney.

Jayasinghe, U. W., Marsh, H. W., & Bond, N. (2001). Peer review in the funding of research in higher education: The Australian experience. *Educational Evaluation and Policy Analysis*, *23*(4), 343–364.

Jayasinghe, U. W., Marsh, H. W., & Bond, N. (2003). A multilevel cross-classified modelling approach to peer review of grant proposals: The effects of assessor and researcher attributes on assessor ratings. *Journal of the Royal Statistical Society Series a-Statistics in Society*, *166*, 279–300.

Jennings, C. G. (2006). Quality and value: The true purpose of peer review. What you can't measure, you can't manage: The need for quantitative indicators in peer review. *Nature*. Retrieved July 6, 2006, from www.nature.com/nature/peerreview/debate/nature05032.html

Johnson, V. E. (2008). Statistical analysis of the National Institutes of Health peer review system. *Proceedings of the National Academy of Sciences*, *105*(32), 11076–11080.

Joint Committee on Quantitative Assessment of Research (2008). *Citation statistics. A report from the International Mathematical Union (IMU) in cooperation with the International Council of Industrial and Applied Mathematics (ICIAM) and the Institute of Mathematical Statistics (IMS)*. Berlin, Germany: International Mathematical Union (IMU).

Journal of the American Medical Association. (1990). Guarding the guardians: Research on editorial peer review. Selected proceedings from the First International Congress on Peer Review in Biomedical Publication. *Journal of the American Medical Association*, *263*(10), 1317–1441.

Journal of the American Medical Association. (1994). The 2nd International Congress on Peer Review in Biomedical Publication. Proceedings. *Journal of the American Medical Association*, *272*(2), 79–174.

Journal of the American Medical Association. (1998). The International Congress on Biomedical Peer Review. *Journal of the American Medical Association*, *280*(3), 203–306.

Journal of the American Medical Association. (2002). The International Congress on Biomedical Peer Review. *Journal of the American Medical Association*, *287*(21), 2759–2871.

Koch, S. (2006). Die Begutachtungsverfahren der Deutschen Forschungsgemeinschaft nach Einführung der Fachkollegien. In S. Hornbostel & D. Simon (Eds.), *Wie viel (In-) Transparenz ist notwendig? Peer Review revisited* (Vol. 1, pp. 15–26). Bonn, Germany: Institut für Forschungsinformation und Qualitätssicherung.

Kostoff, R. N. (1995). Federal research impact assessment: Axioms, approaches, applications. *Scientometrics*, *34*(2), 163–206.

Kostoff, R. N. (1997). The principles and practices of peer review. *Science and Engineering Ethics*, *3*(1), 19–34.

Kostoff, R. N. (2004). *Research program peer review: Purposes, principles, practices, protocols* (DTIC Technical Report Number ADA424141). Arlington, VA: Office of Naval Research.

Krampen, G., & Montada, L. (2002). *Wissenschaftsforschung in der Psychologie*. Göttingen, Germany: Hogrefe.

Kronick, D. A. (1990). Peer review in 18th century scientific journalism. *Journal of the American Medical Association*, *263*(10), 1321–1322.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. *American Psychologist*, *43*(8), 635–642.

LaFollette, M. C. (1992). *Stealing into print: Fraud, plagiarism and misconduct in scientific publishing*. Berkeley: University of California Press.

Lamont, M. (2009). *How professors think: Inside the curious world of academic judgment*. Cambridge, MA: Harvard University Press.

Ledin, A., Bornmann, L., Gannon, F., & Wallon, G. (2007). A persistent problem: Traditional gender roles hold back female scientists. *EMBO Reports*, *8*(11), 982–987.

Lee, K. P., Boyd, E. A., Holroyd-Leduc, J. M., Bacchetti, P., & Bero, L. A. (2006). Predictors of publication: Characteristics of submitted manuscripts associated with acceptance at major biomedical journals. *Medical Journal of Australia*, *184*(12), 621–626.

Levenson, H., Burford, B., Bonno, B., & Davis, L. (1975). Are women still prejudiced against women? A replication and extension of Goldberg's Study. *Journal of Psychology*, *89*(1), 67–71.

Lienert, G. A. (1987). *Schulnoten-Evaluation*. Frankfurt am Main, Germany: Athenäum.

Lindsey, D. (1989). Using citation counts as a measure of quality in science: Measuring what's measurable rather than what's valid. *Scientometrics*, *15*(3–4), 189–203.

Lloyd, M. E. (1990). Gender factors in reviewer recommendations for manuscript publication. *Journal of Applied Behavioral Analysis*, *23*(4), 539–543.

Lock, S. (1985). *A difficult balance: Editorial peer review in medicine*. Philadelphia, PA: ISI Press.

Long, J. S., & Fox, M. F. (1995). Scientific careers: Universalism and particularism. *Annual Review of Sociology*, *21*, 45–71.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research*, *2*, 161–175.

Mallard, G., Lamont, M., & Guetzkow, J. (2009). Fairness as appropriateness: Negotiating epistemological differences in peer review. *Science*, *Technology*, *& Human Values*, *34*(5), 573–606.

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

Marsh, H. W., & Ball, S. (1981). Interjudgmental reliability of reviews for the *Journal of Educational Psychology*. *Journal of Educational Psychology*, *73*(6), 872–880.

Marsh, H. W., & Ball, S. (1991). Reflections on the peer review process. *Behavioral and Brain Sciences*, *14*(1), 157–158.

Marsh, H. W., Bonds, N. W., & Jayasinghe, U. W. (2007). Peer review process: Assessments by applicant-nominated referees are biased, inflated, unreliable and invalid. *Australian Psychologist*, *42*(1), 33–38.

Marsh, H. W., Bornmann, L., Mutz, R., Daniel, H.-D., & O'Mara, A. (in press). Gender effects in the peer reviews of grant proposals: A comprehensive meta-analysis comparing traditional and multilevel approaches. *Review of Educational Research*.

Marsh, H. W., Jayasinghe, U. W., & Bond, N. W. (2008). Improving the peer-review process for grant applications: Reliability, validity, bias, and generalizability. *American Psychologist*, *63*(3), 160–168.

Martin, B. (2000). Research grants: Problems and options. *Australian Universities' Review*, *43*(2), 17–22.

Martin, B. R., & Irvine, J. (1983). Assessing basic research: Some partial indicators of scientific progress in radio astronomy. *Research Policy*, *12*(2), 61–90.

Martinson, B. C., Anderson, M. S., Crain, A. L., & de Vries, R. (2006). Scientists' perceptions of organizational justice and self-reported misbehaviors. *Journal of Empirical Research on Human Research Ethics*, *1*(1), 51–66.

Matt, G. E., & Navarro, A. M. (1997). What meta-analyses have and have not taught us about psychotherapy effects: A review and future directions. *Clinical Psychology Review*, *17*(1), 1–32.

McClellan, J. E. (2003). *Specialist control: The publications committee of the Academie Royal des Sciences (Paris) 1700–1793* (Transactions of the American Philosophical Society, v. 93). Philadelphia, PA: American Philosophical Society.

McCullough, J. (1989). First comprehensive survey of NSF applicants focuses on their concerns about proposal review. *Science*, *Technology*, *& Human Values*, *14*(1), 78–88.

McCullough, J. (1994). The role and influence of the US National Science Foundation's program officers in reviewing and awarding grants. *Higher Education*, *28*(1), 85–94.

McDonald, R. J., Cloft, H. J., & Kallmes, D. F. (2007). Fate of submitted manuscripts rejected from the *American Journal of Neuroradiology*: Outcomes and commentary. *American Journal of Neuroradiology*, *28*(8), 1430–1434.

McDonald, 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*(2), 253–256.

McIntosh, E. G., & Ross, S. (1987). Peer review in psychology: Institutional ranking as a factor. *Psychological Reports*, *60*(3), 1049–1050.

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*(10), 702–712.

Merton, R. K. (1942). Science and technology in a democratic order. *Journal of Legal and Political Sociology*, *1*, 115–126.

Merton, R. K. (1948). The self-fulfilling prophecy. *Antioch Review*, *8*, 193–210.

Merton, R. K. (1968). The Matthew effect in science. *Science*, *159*(3810), 56–63.

Moed, H. (2008). UK Research Assessment Exercises: Informed judgments on research quality or quantity? *Scientometrics*, *74*(1), 153–161.

National Academy of Sciences. (2006). *Beyond bias and barriers: Fulfilling the potential of women in academic science and engineering*. Washington, DC: The National Academies Press.

Neuhaus, C., & Daniel, H.-D. (2009). A new reference standard for citation analysis in chemistry and related fields based on the sections of Chemical Abstracts. *Scientometrics*, *78*(2), 219–229.

Nicolaisen, J. (2007). Citation analysis. *Annual Review of Information Science and Technology*, *41*, 609–641.

Nylenna, M., Riis, P., & Karlsson, Y. (1994). Multiple blinded reviews of the same two manuscripts: Effects of referee characteristics and publication language. *Journal of the American Medical Association*, *272*(2), 149–151.

Opthof, T., Furstner, F., van Geer, M., & Coronel, R. (2000). Regrets or no regrets? No regrets! The fate of rejected manuscripts. *Cardiovascular Research*, *45*(1), 255–258.

Opthof, T., & Wilde, A. A. M. (2009). The Hirsch-index: A simple, new tool for the assessment of scientific output of individual scientists: The case of Dutch professors in clinical cardiology. *Netherlands Heart Journal*, *17*(4), 145–154.

Over, R. (1996). Perceptions of the Australian Research Council Large Grants Scheme: Differences between successful and unsuccessful applicants. *Australian Educational Researcher*, *23*(2), 17–36.

Overbeke, J., & Wager, E. (2003). The state of the evidence: What we know and what we don't know about journal peer review. In F. Godlee & T. Jefferson (Eds.), *Peer review in health sciences* (2nd ed., pp. 45–61). London: BMJ Books.

Overview: *Nature*'s peer review trial. (2006). *Nature*. Retrieved May 18, from www.nature.com/nature/peerreview/debate/nature05535.html

Owen, R. (1982). Reader bias. *Journal of the American Medical Association*, *247*(18), 2533–2534.

Paludi, M. A., & Bauer, W. D. (1983). Goldberg revisited: What's in an author's name. *Sex Roles*, *9*(3), 387–390.

Paludi, M. A., & Strayer, L. A. (1985). What's in an author's name? Differential evaluations of performance as a function of author's name. *Sex Roles*, *12*(3–4), 353–361.

Patterson, S. C., Bailey, M. S., Martinez, V. J., & Angel, S. C. (1987). Report of the managing editor of the *American Political Science Review*, 1986–1987. *PS*, *20*, 1006–1016.

Pendlebury, D. A. (2008). *Using bibliometrics in evaluating research*. Philadelphia, PA: Research Department, Thomson Scientific.

Perneger, T. V. (2004). Relation between online "hit counts" and subsequent citations: Prospective study of research papers in the *BMJ*. *British Medical Journal*, *329*, 546–547.

Peters, D. P., & Ceci, S. J. (1982). Peer-review practices of psychological journals: The fate of accepted, published articles, submitted again. *Behavioral and Brain Sciences*, 5(2), 187–195.

Petty, R. E., & Fleming, M. A. (1999). The review process at *PSPB*: Correlates of interreviewer agreement and manuscript acceptance. *Personality and Social Psychology Bulletin*, *25*(2), 188–203.

Pfeffer, J., Leong, A., & Strehl, K. (1977). Paradigm development and particularism: Journal publication in three scientific disciplines. *Social Forces*, *55*(4), 938–951.

Pierie, J. P. E. N., Walvoort, H. C., & Overbeke, A. J. P. M. (1996). Readers' evaluation of effect of peer review and editing on quality of articles in the *Nederlands Tijdschrift voor Geneeskunde*. *Lancet*, *348*(9040), 1480–1483.

Popper, K. R. (1961). *The logic of scientific discovery* (2nd ed.). New York: Basic Books.

Pöschl, U. (2004). Interactive journal concept for improved scientific publishing and quality assurance. *Learned Publishing*, *17*(2), 105–113.

Pruthi, S., Jain, A., Wahid, A., Mehra, K., & Nabi, S. A. (1997). Scientific community and peer review system: A case study of a central government funding scheme in India. *Journal of Scientific & Industrial Research*, *56*(7), 398–407.

Publishing Research Consortium. (2008). *Peer review in scholarly journals: Perspective of the scholarly community: An international study*. Bristol, UK: The Consortium.

Ray, J., Berkwits, M., & Davidoff, F. (2000). The fate of manuscripts rejected by a general medical journal. *American Journal of Medicine*, *109*(2), 131–135.

Resnik, D. B., Gutierrez-Ford, C., & Peddada, S. (2008). Perceptions of ethical problems with scientific journal peer review: An exploratory study. *Science and Engineering Ethics*, *14*(3), 305–310.

Ross, P. F. (1980). *The sciences' self-management: Manuscript refereeing, peer review, and goals in science*. Lincoln, MA: The Ross Company.

Rossiter, J. R. (2003). Qualifying the importance of findings. *Journal of Business Research*, *56*(1), 85–88.

Roy, R. (1985). Funding science: The real defects of peer-review and an alternative to it. *Science*, *Technology*, *& Human Values*, *52*, 73–81.

Russo, G. (2008). Statistical analyses raise questions over NIH grant reviews. *Nature*, *454*, 801.

Sahner, H. (1982). On the selectivity of editors: An input-output analysis of the *Zeitschrift für Soziologie*. *Zeitschrift für Soziologie*, *11*(1), 82–98.

Schuster, C. (2002). A mixture model approach to indexing rater agreement. *British Journal of Mathematical & Statistical Psychology*, *55*, 289–303.

Scott, W. A. (1974). Interreferee agreement on some characteristics of manuscripts submitted to the *Journal of Personality and Social Psychology*. *American Psychologist*, *29*(9), 698–702.

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

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin Company.

Sharp, D. W. (1990). What can and should be done to reduce publication bias: The perspective of an editor. *Journal of the American Medical Association*, *263*(10), 1390–1391.

Shashok, K. (2005). Standardization vs diversity: How can we push peer review research forward? *Medscape General Medicine*, *7*(1), 11.

Shatz, D. (2004). *Peer review: A critical inquiry*. Lanham, MD: Rowman & Littlefield.

Siegelman, S. S. (1991). Assassins and zealots: Variations in peer review: Special report. *Radiology*, *178*(3), 637–642.

Sismondo, S. (1993). Some social constructions. *Social Studies of Science*, *23*(3), 515–553.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. London: Chapman & Hall/CRC.

Smigel, E. O., & Ross, H. L. (1970). Factors in editorial decision. *American Sociologist*, *5*(1), 19–21.

Smith, L. C. (1981). Citation analysis. *Library Trends*, *30*(1), 83–106.

Smith, R. (2006). Peer review: A flawed process at the heart of science and journals. *Journal of the Royal Society of Medicine*, *99*(4), 178–182.

Sokal, A. D. (2008). *Beyond the hoax: Science, philosophy and culture*. Oxford, UK: Oxford University Press.

Speck, B. W. (1993). *Publication peer review: An annotated bibliography*. Westport, CT: Greenwood Press.

Stamps, A. E. (1997). Advances in peer review research: An introduction. *Science and Engineering Ethics*, *3*(1), 3–10.

Stehbens, W. E. (1999). Basic philosophy and concepts underlying scientific peer review. *Medical Hypotheses*, *52*(1), 31–36.

Sternberg, R. J., Hojjat, M., Brigockas, M. G., & Grigorenko, E. L. (1997). Getting in: Criteria for acceptance of manuscripts in *Psychological Bulletin*, 1993–1996. *Psychological Bulletin*, *121*(2), 321–323.

Stieg Dalton, M. F. (1995). Refereeing of scholarly works for primary publishing. *Annual Review of Information Science and Technology*, *30*, 213–250.

Stricker, L. J. (1991). Disagreement among journal reviewers: No cause for undue alarm. *Behavioral and Brain Sciences*, *14*(1), 163–164.

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*(1), 40–50.

Tight, M. (2003). Reviewing the reviewers. *Quality in Higher Education*, *9*(3), 295–303.

Tregenza, T. (2002). Gender bias in the refereeing process? *Trends in Ecology & Evolution*, *17*(8), 349–350.

U.S. General Accounting Office (1999). *Peer review practices at federal science agencies vary*. Washington, DC: The Office.

U.S. Office of Management and Budget. (2004). *Revised information quality bulletin for peer review*. Washington, DC: The Office.

van den Besselaar, P., & Leydesdorff, L. (2007). *Past performance as predictor of successful grant applications. A case study*. Den Haag, The Netherlands: Rathenau Instituut.

van Raan, A. F. J. (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics*, *36*(3), 397–420.

van Raan, A. F. J. (1999). Advanced bibliometric methods for the evaluation of universities. *Scientometrics*, *45*(3), 417–423.

von Eye, A., & Mun, E. Y. (2005). *Analyzing rater agreement: Manifest variable methods*. Mahwah, NJ: Erlbaum.

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

Ward, C. (1981). Prejudice against women: Who, when, and why? *Sex Roles*, *7*(2), 163–171.

Wellcome Trust. (2001). *Review of Wellcome Trust PhD research training: Career paths of a 1988 – 1990 prize student cohort*. London: The Trust.

Weller, A. C. (1996). Editorial peer review: A comparison of authors publishing in two groups of US medical journals. *Bulletin of the Medical Library Association*, *84*(3), 359–366.

Weller, A. C. (2002). *Editorial peer review: Its strengths and weaknesses*. Medford, NJ: Information Today, Inc.

Wennerås, C., & Wold, A. (1997). Nepotism and sexism in peer-review. *Nature*, *387*(6631), 341–343.

Wessely, S. (1998). Peer review of grant applications: What do we know? *Lancet*, *352*(9124), 301–305.

Wessely, S., & Wood, F. (1999). Peer review of grant applications: A systematic review. In F. Godlee & T. Jefferson (Eds.), *Peer Review in Health Sciences* (pp. 14–31). London: BMJ Books.

White, H. D. (2005). On extending informetrics: An opinion paper. *Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics*, *2*, 442–449.

Whitley, R., & Gläser, J. (Eds.). (2007). *The changing governance of the sciences: The advent of research evaluation systems*. Dordrecht, the Netherlands: Springer.

Whitman, N., & Eyre, S. (1985). The pattern of publishing previously rejected articles in selected journals. *Family Medicine*, *17*(1), 26–28.

Wiener, S., Urivetsky, M., Bregman, D., Cohen, J., Eich, R., Gootman, N., et al. (1977). Peer review: Inter-reviewer agreement during evaluation of research grant evaluations. *Clinical Research*, *25*, 306–311.

Wiley, S. (2008). Peer review isn't perfect ... But it's not a conspiracy designed to maintain the status quo. *The Scientist*, *22*(11), 31.

Wilson, J. D. (1978). Peer review and publication. *Journal of Clinical Investigation*, *61*(4), 1697–1701.

Wood, F. Q., & Wessely, S. (2003). Peer review of grant applications: A systematic review. In F. Godlee & T. Jefferson (Eds.), *Peer review in health sciences* (2nd ed., pp. 14–44). London: BMJ Books.

Yalow, R. S. (1982). Is subterfuge consistent with good science? *Bulletin of Science Technology & Society*, *2*(5), 401–404.

Young, S. N. (2003). Peer review of manuscripts: Theory and practice. *Journal of Psychiatry & Neuroscience*, *28*(5), 327–330.

Ziman, J. (2000). *Real science. What it is, and what it means*. Cambridge, UK: Cambridge University Press.

Zuckerman, H., & Merton, R. K. (1971a). Patterns of evaluation in science: Institutionalisation, structure and functions of the referee system. *Minerva*, *9*(1), 66–100.

Zuckerman, H., & Merton, R. K. (1971b). Sociology of refereeing. *Physics Today*, *24*(7), 28–33.