Improving the Peer-Review Process for Grant Applications

Reliability, Validity, Bias, and Generalizability

Herbert W. Marsh — *University of Oxford*
Upali W. Jayasinghe — *University of New South Wales*
Nigel W. Bond — *University of Western Sydney*

Peer review is a gatekeeper, the final arbiter of what is valued in academia, but it has been criticized in relation to traditional psychological research criteria of reliability, validity, generalizability, and potential biases. Despite a considerable literature, there is surprisingly little sound peer-review research examining these criteria or strategies for improving the process. This article summarizes the authors' research program with the Australian Research Council, which receives thousands of grant proposals from the social science, humanities, and science disciplines and reviews by assessors from all over the world. Using multilevel cross-classified models, the authors critically evaluated peer reviews of grant applications and potential biases associated with applicants, assessors, and their interaction (e.g., age, gender, university, academic rank, research team composition, nationality, experience). Peer reviews lacked reliability, but the only major systematic bias found involved the inflated, unreliable, and invalid ratings of assessors nominated by the applicants themselves. The authors propose a new approach, the reader system, which they evaluated with psychology and education grant proposals and found to be substantially more reliable and strategically advantageous than traditional peer reviews of grant applications.

Keywords: peer review, grant proposals, bias, validity, reliability

The purpose of this article is to critically evaluate and propose strategies to improve the peer-review process for grant applications. The peer-review process is highly valued but widely criticized as the primary basis for evaluating what is good in academic settings. In psychology departments and other academic settings, peer review is used to evaluate grant proposals, journal submissions, job applications, promotions, tenure, monographs, textbooks, doctoral theses, doctoral and postdoctoral applications, and other academic products (Bornmann & Daniel, 2005; Chubin, 1994; Cicchetti, 1991; Jayasinghe, Marsh, & Bond, 2001, 2003; Marsh & Ball, 1981, 1989, 1991). In addition, peer review provides constructive feedback to authors that is useful in revising or implementing their work (Nickerson, 2005). More broadly, this process serves a gatekeeper role, acting as the final arbiter of what is valued and acceptable, a filtering system to establish what research findings are trustworthy—a seal of approval. Here we briefly consider previous research on peer review as a generic process, and then we present results from our research program, which focuses specifically on the peer review process for reviewing grant proposals.

Despite the importance of peer review and its long history of controversy, there is surprisingly little empirically rigorous research in this area. For example, on the basis of experience with the National Science Foundation and the National Institutes of Health in the United States, Chubin (1994) claimed that the peer-review process is so highly valued that it is often considered to be sacrosanct, above reproach, and not subject to serious scrutiny. Indeed, Goldbeck-Wood (1999) suggested that it "is traditionally surrounded by an almost religious mystique" (p. 44). However, Peters and Ceci (1982) noted the "possibility of response bias in the peer-review process (e.g., institutional affiliation, paradigm confirmation or theory support, editor–author friendship, 'old boy networks')" (p. 188). In a systematic review of the peer-review processes in biomedical re-

Herbert W. Marsh, Department of Education, University of Oxford, Oxford, United Kingdom; Upali W. Jayasinghe, Centre for Primary Health Care and Equity, University of New South Wales, Sydney, Australia; Nigel W. Bond, School of Psychology, University of Western Sydney, New South Wales, Australia.

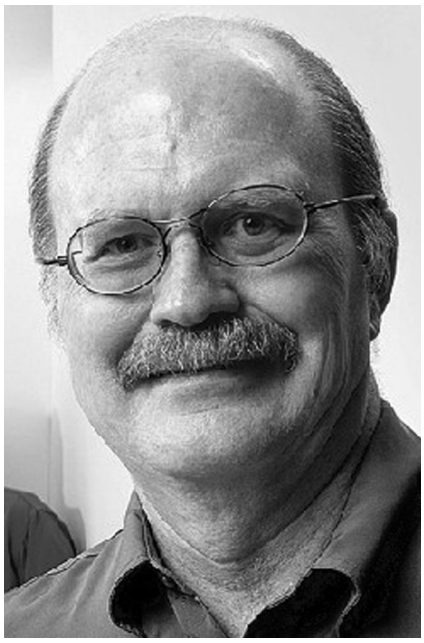
This research was funded in part by a grant from the Australian Research Council to Herbert W. Marsh that helped support doctoral research undertaken by Upali W. Jayasinghe when all three authors were at the University of Western Sydney.

We would like to express our thanks to the Australian Research Council and particularly to Max Brennan, former chair of the Australian Research Council, for their assistance in providing the data used in this study. We would also like to thank Alison O'Mara for helpful comments on drafts of this article.

A more detailed description of materials, statistical analyses, and the overall research project is available from the Jayasinghe (2003) doctoral dissertation, which is available at http://self.uws.edu.au/Theses/Jayasinghe/list.htm; also see Jayasinghe, Marsh, and Bond, 2001, 2003, 2006; Marsh and Ball, 1981, 1989; and Marsh, Bond, and Jayasinghe, 2007.

Correspondence concerning this article should be addressed to Herbert W. Marsh, Department of Education, University of Oxford, 15 Norham Gardens, Oxford OX2 6PY, United Kingdom. E-mail: herb.marsh@education.ox.ac.uk

Vol. 63, No. 3, 160–168 DOI: 10.1037/0003-066X.63.3.160



Herbert W. Marsh

search, Jefferson, Rudin, Brodney, and Davidoff (2006) found that good research on the peer-review process was so rare that almost no conclusions were warranted, particularly about constructive alternatives and interventions designed to improve peer reviews. They also noted the poor generalizability of results based on small idiosyncratic samples, which led Jefferson (2001) to claim, "If I manufactured a drug called peer review and applied to the Food and Drug Administration for its registration on the basis of currently available evidence, they would collapse laughing" (p. 1463).

An editorial in the *British Medical Journal* (Tite & Schroter, 2006) argued that it is ironic that there is so little evidence of the scientific method being employed to evaluate the peer-review process used to select scientific articles. Seeking ways to improve peer reviews, Schroter et al. (2004) found that a short-term peer-review training package had a small effect on the quality of reviews but that the effect was no longer significant in a six-month follow-up. After comparing peer reviews by reviewers nominated by the authors with those by reviewers selected by the editors of 10 medical journals, Schroter, Tite, Hutchings, and Black (2006) concluded that editors should be cautious about relying on the recommendations of author-nominated reviewers.

Research Grants Funding by the Australian Research Council

In this article we summarize major findings from our research program of peer reviews of grant applications (Jayasinghe, Marsh, & Bond, 2001, 2003, 2006; Marsh, Bond, & Jayasinghe, 2007), which is based on our collaboration with the Australian Research Council (ARC; 1996). The ARC is the main source of funding for basic research in Australia; it covers all science, social science, and humanities disciplines and obtains reviews from external reviewers from all over the world: Australia (56.6%), North America (United States and Canada, 19.6%), Europe (18.7%), and other areas (Asia, Africa, South America, and New Zealand, 5.1%). Much of our research is based on a database of 2,331 proposals rated by 6,233 external assessors, who provided a total of 10,023 reviews—an average of 4.3 assessors per proposal. Using this large database, we were able to test the generalizability of the results across disciplines and across assessors from different nationalities. Methodologically, we pioneered the use of multilevel cross-classified models (Level 1, assessor and proposal cross-classification; Level 2, field of study), taking into account both the fact that 34% of the assessors evaluated more than one proposal and the lack of independence of data at different levels, issues that are largely ignored in the single-level model approach that is widely used in peer-review research (for further discussion, see Jayasinghe et al., 2003).

In the ARC peer-review process for reviewing grant applications, applicants submitted proposals to one of nine discipline panels and nominated assessors to evaluate their proposals (applicant-nominated assessors, or ANAs). In a preliminary evaluation, the panel culled proposals (22%) that were ineligible or deemed uncompetitive (after having been read by at least two panel members). Each proposal was then sent to four external assessors nominated by the panel (panel-nominated assessors, or PNAs) and one ANA. In order to minimize conflict of interest, assessors were not selected if they were from the same institution as or had co-authored any publications with any of the applicants during the past five years. Assessors rated the proposals on the quality of the proposal (*project ratings*) and the research track record of the applicants (*researcher ratings*), which were weighted .6 and .4, respectively, by the ARC. These ratings were supplemented by ratings of the originality, methodology, and scientific/theoretical merit of the proposal, the research track record of each member of the research team, and written comments (for a more detailed discussion of results based on these supplemental ratings, see Jayasinghe, 2003; also see Marsh & Ball, 1989). External reviews were sent to the applicants, who provided a brief, one-page rejoinder. ARC panel members then evaluated all materials, including the rejoinder, to determine a final panel rating that was the basis of funding. Within each panel, panel members allocated appropriate funds to the proposals (typically not the full amounts requested), starting from the best-ranked proposal and working down the list until the available funding was exhausted. The probability of success (including proposals that were initially culled) was 21%.

Single-Rater Reliability: Agreement Among Different Assessors of the Same Grant Proposal

Among the many criticisms directed at the peer-review process, the most basic, broadly supported, and damning is its failure to achieve acceptable levels of agreement among independent assessors, which results in unreliable peer

Upali W. Jayasinghe

reviews (Callaham, Baxt, Waeckerle, & Wears, 1998; Cicchetti, 1991; Jayasinghe, 2003; Marsh & Ball, 1981). In order to provide a common benchmark, Marsh and Ball (1981, 1989, 1991) defined single-rater reliability as the correlation between two independent assessors' ratings of the same submissions across a large number of different submissions. It can also be derived from analysis of variance (Shrout & Fleiss, 1979) and multilevel modeling (Goldstein, 2003; Jayasinghe et al., 2003; Snijders & Bosker, 1999). This single-rater reliability can then be used to estimate, with the Spearman–Brown equation, the reliability of a mean rating based on varying numbers of raters (Marsh & Ball, 1989). For overall assessments based on 16 peer-review studies of journal articles (Cicchetti, 1991), single-rater reliabilities varied from .19 to .54 (*Mdn* = .30). Although there is less research on the reliability of assessments of grant proposals, which are the focus of the present investigation, Cicchetti (1991; also see Klahr, 1985) reported single-rater reliabilities between .17 and .37 (*Mdn* = .33) for nine analyses of reviews of submissions to the (American) National Science Foundation. These figures suggest that the reliability of assessments of grant proposals may be comparable to the reliability of assessments of journal submissions.

We (Jayasinghe et al., 2001, 2003, 2006; Marsh et al., 2007) found that the single-rater reliabilities for reviews of grant applications obtained for the ARC large grant scheme were .15 for the quality of the proposal and .21 for the quality of the research team. We then applied the Spearman–Brown equation to estimate the reliabilities based on an average of 4.3 external assessors per proposal (the average number of assessors per proposal for the ARC peer-review process), which were .44 (quality of the project) and .53 (quality of the research team). Thus, assessors are better able to differentiate reliably between the track records of researchers than between the quality of the proposed projects. We used the Spearman–Brown equation again to determine that it would require at least 6 assessors per proposal to achieve more acceptable reliability estimates of .71 (project) and .82 (researcher). We emphasize that these results underestimate the true reliability of the process (since 22% of the initial proposals were not included because they were culled as uncompetitive and since the final decision was based on more information than just external assessor ratings, including narrative summaries by the assessor and a response to the reviews by the author; for further discussion of this issue, see Jayasinghe, 2003, and Marsh & Ball, 1981, 1989). Nevertheless, the interrater reliability estimates are not adequate, falling well below acceptable levels of .8 (or even .9). Indeed, Helmstadter (1964) demonstrated that in order to successfully differentiate between two cases that differ by one quarter of a standard deviation with an 80% probability, a reliability of .94 is needed.

Jayasinghe et al. (2001) demonstrated that the standard error of the ratings (based on measurement error) was 4.6. When we constructed 95% confidence intervals for each proposal, few proposals were significantly different from the cutoff value for funding. Hence, for most successful and unsuccessful grant proposals, the decision of whether or not to fund was based substantially on chance, whether the random error happened to be positive or negative.

Does Peer-Review Reliability of Grant Applications Differ According to Discipline?

It has been argued that peer-review outcomes are less prone to error and bias in the physical and biological sciences than in the social sciences and humanities (e.g., Lindsey, 1978; also see Cicchetti, 1991; Zuckerman & Merton, 1971). Because our database is ideally suited to test this hypothesis, we did separate analyses for grant proposals from the sciences and grant proposals from the social sciences and humanities (Jayasinghe et al., 2003). We found that the single-rater reliabilities across the social sciences and humanities were marginally higher than those for science panels (.18 vs. .17 for project; .26 vs. .23 for researcher). Hence, ratings in science were certainly no more reliable than ratings in the social sciences and humanities. Clearly, the problem of low reliabilities of the peer reviews of grant applications generalizes over a range of disciplines.

How Trustworthy Are Peer Reviews of Grant Applications by Applicant-Nominated Assessors?

Some funding bodies allow applicants to nominate their own assessors (ANAs), but the merits of this strategy have not been tested (Grimm, 2005). In our database, most grant proposals (81%) had at least one ANA and one assessor nominated by the funding panel (PNA). Using a "within proposal" perspective, we (Marsh et al., 2007) compared ratings of the same proposal by ANAs and PNAs. This

Nigel W. Bond

allowed us to control the many sources of variation associated with a specific proposal—particularly, the quality of the proposal, which was necessarily held constant when two assessors reviewed the same proposal. We then extended the logic of our "within proposal" perspective (ANA and PNA ratings of the same proposal) to incorporate a "within assessor" perspective in which 555 assessors reviewed different proposals as both an ANA and a PNA.

In each of the nine discipline panels, ANA ratings of grant proposals were half a standard deviation higher than PNA ratings, were less related to ratings by other assessors, were less related to the ARC final assessment, and contributed to the unreliability of peer reviews. Furthermore, when the same assessor was both an ANA and a PNA for different proposals, the assessor's ratings in the role of ANA were biased, whereas those by the same person in the role of PNA were not. It is not surprising, perhaps, that the ANA assessors who were specifically invited by an applicant to serve as a reviewer and agreed to do so would feel a dual responsibility as a critical reviewer and as an advocate of the applicant. This is consistent with the finding that the behavior of the same reviewer differed systematically when in the role of an anonymous PNA and when in the role of an ANA. Nevertheless, this unique feature of our data provides a particularly strong basis for the evaluation of the ANA bias. These results led the ARC to discontinue their use of ANAs despite the potentially adverse reactions of researchers, who appreciated being able to nominate their own assessors. These results provide clear answers to Grimm's (2005) questions about the use of ANAs for reviewing grant proposals: (a) No, funding bodies should not encourage the use of ANAs—at least not for deciding whether to fund grants, and (b) if funding bodies allow ANAs, then applicants should take advantage of this option because they are likely to be disadvantaged if they do not do so.

How Does the Nationality of Assessors Affect Ratings of Grant Applications and the ANA Bias?

Ratings of grant applications made by Australian assessors were significantly lower (relative to an *SD* of 11.0 on the 0–100 scale; all *p*s $< .05$) than those made by assessors from North American (by 3.8 points, effect size [ES] = .34), European (1.1 points, ES = .09), and other countries (1.8 points, ES = .16). Because geographic region and researcher-nominated status were confounded (ANAs were more likely to come from outside of Australia than were PNAs chosen by the ARC), after adjusting for the ANA bias, we found the differences were reduced to 2.6 points for North American ($p < .05$), 0.2 points for European (*ns*), and 1.6 points for other countries ($p < .05$). Thus, Australian assessors tended to be harsher than assessors from other countries, but some of this difference can be explained in terms of the ANA biases.

North American assessors—PNAs and ANAs—gave higher ratings than assessors from other countries, and their reviews of grant applications were somewhat less reliable (in relation to agreement with other reviewers of the same application) and somewhat less valid (in relation to the final panel rating). These results are consistent with suggestions that Americans are part of a culture that is comfortable being generous in their evaluations—particularly after having accepted an invitation from an applicant to serve as an ANA. However, further research is needed to determine how much of this effect is a response bias that is due to nationality per se or to other confounding factors and whether the results generalize to peer reviews collected in the United States and other countries. In particular, it may be that Australian assessors were more critical because they were sometimes competing for the same scarce funds. Nevertheless, American assessors, particularly ANAs, were generous even in relation to reviewers from other countries.

Effects of External Assessor and Applicant Attributes on the Success of Grant Proposals

Does the number of grant proposals assessed by assessors make a difference? An important problem in most peer-review processes is that external assessors typically view only one or a very few submissions. In our database, a total of 4,100 (65.8%) external assessors assessed only one proposal. We (Jayasinghe, 2003; Jayasinghe et al., 2001, 2003) have suggested that assessors who rated only a single grant proposal did not have a sufficient frame of reference for translating subjective impressions about the quality of a proposal and the quality of a research track record onto the numerical scale that constituted the basis of the peer-review ratings. Hence, even though two different assessors might give the same proposal very different marks on a rating scale (owing to their tendencies to be relatively lenient or harsh),

they might agree substantially on the rank order of a proposal relative to other proposals if given access to all the other proposals. We found that ratings by assessors who rated three or more proposals were significantly harsher, more reliable (in relation to other reviews of the same proposal), and more valid (in relation to the final panel rating). However, even among the 15 assessors who evaluated 10 or more proposals, we found that some reviewers were consistently more lenient in their ratings than others. Correcting for these assessor response biases resulted in higher reliabilities and more reliable differentiation among the grant proposals (Jayasinghe, 2003).

Academic rank: Are professors more likely to be funded? For our grant application data, applicants with the rank of professor were disproportionately successful, whereas those with the titles of Dr., Mr., or Ms. were significantly underrepresented in the group that obtained grants. The results were similar for applicants named first in a proposal and all applicants in a proposal and were consistent across different disciplines. However, the finding that professors are evaluated more favorably is not unexpected and should not be interpreted to mean that the assessment process is biased (see Jayasinghe et al., 2001; Marsh & Bazeley, 1999). Indeed, particularly in the Australian context, where the title of professor is harder to achieve and is less frequent than, for example, in North America, professors typically have to have successful research track records, including research grants, in order to become professors. Hence, it is reasonable that professors should have a substantial advantage over nonprofessors, particularly in terms of their research track records. Consistent with these speculations, professor status was significantly ($p < .001$) more related to researcher ratings ($r = .14$) than to proposal ratings ($r = .08$), providing preliminary support for the construct validity of grant proposal ratings in relation to the title of the researcher, rather than a bias interpretation.

Does university affiliation influence a grant applicant's success? Australian universities are classified into four groups that are roughly associated with university status and prestige, here called Groups A through D (see Jayasinghe, 2003), with Group A having the highest status and prestige. The percentages of the total grant applications and the success rates, respectively, for the four groups were in line with expectations: Group A, 51.0% and 59.4%; Group B, 33.1% and 31.8%; Group C, 9.9% and 6.1%, Group D, 4.7% and 2.2%; and nonuniversity institutions, 1.4% and 0.5%. Group A's success rate was significantly greater than its representation among the total number of grant application proposals, whereas Groups B, C, and D and other nonuniversity institutions had success rates significantly less than their rates of representation in the grant application process. Although so small as to be of little practical significance, it is unclear whether this institutional affiliation effect on grant proposal ratings represents a source of validity (researchers from more prestigious universities are stronger researchers) or a source of bias.

Do older grant applicants get better ratings? Grant applicant age explained only about half of 1% of the variance in ratings of grant applications. Success rates did not differ significantly for different age groups. However, a polynomial regression analysis resulted in a small but highly significant cubic effect of age: an increase in ratings up to the age of 40 (25.6% of the applicants), almost no change in ratings between the ages of 40 and 60 (68.7% of the applicants), and another increase in ratings for the researchers who were over 60 (5.7% of applicants; see Jayasinghe, 2003). However, supplemental analyses suggested interesting interpretations of effects at both ends of the age continuum.

Young researchers (under the age of 40) were less likely to have their proposals culled in the initial round. When culled proposals were included, there were no significant differences between young researchers and other researchers in the probability of being funded. This suggests that, consistent with the ARC policy to encourage early career researchers, young researchers were given the benefit of the doubt at the time of initial culling. This strategy would probably enhance their confidence and allow them to obtain potentially valuable feedback from external assessors even if they were eventually unsuccessful. Retaining more early career proposals that would otherwise have been culled apparently led to lower evaluations for young applicants whose proposals were not culled.

There is also a counterexplanation for the increase in the ratings of researchers older than 60. Until recently (and at the time of our study), the typical age of retirement in Australia was 60; beyond this age, a generous pension scheme meant that there was little financial benefit in continuing to work full time. An honorary title (e.g., emeritus professor) was a viable compromise for committed researchers that allowed them more flexibility to pursue their research at little loss of after-tax salary. Hence, we speculate that the age effects observed beyond the age of 60 were self-selection effects; highly successful researchers continued to apply for grants after the age of 60 and continued to be highly successful; less successful researchers were more likely to put their efforts into other activities after the retirement age of 60.

Gender of the grant applicants and external assessors: Are women applicants disadvantaged? There are two quite different perspectives to this issue (Jayasinghe, 2003). From one perspective, because only 15.3% of the grant applicants were women, women are substantially underrepresented among those researchers who apply for ARC grants, and this represents a worrisome bias. However, from another perspective, the percentage of successful applications by female researchers (15.2%) was almost exactly proportional to their representation. When the gender of only the first-named investigator in the proposal was considered, the success rate was 21% for both men and women. More detailed analyses on second- and third-named researchers also indicated that the success rate did not differ significantly for men and women. Furthermore, the (non)effect of gender did not interact significantly with panel, demonstrating that the

lack of a gender effect generalized well over the nine social science, humanities, and science disciplines. Hence, for this large multidisciplinary archive of peer reviews, there was no evidence of a gender bias in the reviews of applications by men and women researchers.

We also tested a "matching hypothesis" that external assessors would give higher ratings to researchers of the same sex (for further discussion, see Jayasinghe, 2003). Overall, the percentage of female assessors was only 9%, but the percentages of female assessors were substantially higher in the social sciences (34%) and the humanities (23%). To evaluate the matching hypothesis, we considered proposals with at least one male and one female assessor. For this analysis, effects that were due to researcher gender, assessor gender, and their interaction were all statistically nonsignificant and were consistent across the discipline panels. When these interaction effects were evaluated with the more powerful multilevel cross-classified models that allowed the use of all of the available data, the interaction effects remained statistically nonsignificant. (Jayasinghe, 2003, also found no support for a matching hypothesis in terms of applicant/assessor age, academic title, and prestige of university affiliation; also see Bornmann & Daniel, 2007). In summary, there was no support for either a gender bias based on applicant gender or a gender-matching hypothesis.

Our research into peer reviews of grant applications is notable in contradicting widely cited claims of gender bias in peer reviews (e.g., Wenneras & Wold, 1997). Our research is clearly more generalizable than most such research in that it is based on a large number of applicants as well as external assessors from all over the world and includes proposals from the sciences, the social sciences, and the humanities. However, our findings do not argue that there are necessarily no gender biases associated with other, more idiosyncratic peer-review processes—only that it is possible to have a peer-review process without systematic gender bias in the evaluation of proposals by female grant applicants. Indeed, in a recently published meta-analysis based on 66 different peer-review studies of grant applications, Bornmann, Mutz, and Daniel (2007) found that estimated gender effects varied from 22.1% in favor of men to 22.9% in favor of women applying for grants. Although there was a small gender effect in favor of men overall that was marginally significant ($.01 < p < .05$) because of the large sample sizes, a majority of the individual studies showed no significant gender effect. From a different perspective, however, our study is highly consistent with most other research showing that women are substantially underrepresented in the numbers who apply for grants—even if there is no gender bias in the review of their grant proposals.

Combined Effects of Grant Applicant and Assessor Attributes

We subsequently examined the combined effects of grant applicant and assessor attributes—including the type of assessor (PNA vs. ANA), the nationality of the assessor, the number of proposals reviewed by each assessor, the academic rank of the first-named researcher, and the university type of the first-named researcher. Controlling for assessor characteristics led to small increases in reliable differentiation among the proposals, suggesting that those assessor characteristics were a potential source of bias. In contrast, controlling for grant applicant characteristics led to somewhat less reliable differentiation among proposals, suggesting that these were valid sources of variance rather than sources of bias. The juxtaposition of these two results is important; it suggests that more bias is associated with assessor characteristics but that applicant characteristics are more likely to contribute to validity.

For the combined analysis, ANA/PNA differences had the largest effect (ANAs giving higher ratings than PNAs), and the size of this effect actually increased after we controlled for other characteristics in the combined model. Similarly, North American assessors continued to give higher ratings than assessors from other countries, and this difference was particularly large for ANAs. Also, assessors who evaluated more proposals gave lower ratings that were more reliable and valid. Applicants with the academic title of professor and those from older, more prestigious universities also received higher ratings, although these two variables did not interact. Hence, this model of combined characteristics resulted in a pattern of findings that was similar to the pattern that resulted from analyses of each of these characteristics considered separately. In summary, we have critically evaluated a wide variety of effects associated with characteristics of grant applicants, characteristics of assessors, and their interaction, but the only major source of systematic bias that we found was the inflated, unreliable, and invalid ratings given by assessors who were nominated by the applicants themselves (ANAs).

The Reader System: How to Improve Peer Reviews of Grant Applications in Psychology and Education

Although we are appropriately critical of the peer-review process for grant applications, we have also highlighted possible strategies to improve peer-review reliability for grant proposals (e.g., excluding ANAs). We also found that reliability is better for assessors who evaluate more proposals but that even frequently used assessors had systematic response biases (leniency or harshness) that detracted from the reliability of their assessments. On the basis of these results, we proposed and tested a simple, straightforward process to enhance the peer-review reliability of grant applications: the reader trial system (Jayasinghe et al., 2006).

In the reader system, small numbers of expert readers (typically three or four) were used for each of a few selected subdisciplines within psychology (cognition, developmental, educational, learning, perception, and physiology) and education (secondary, tertiary, and policy) that are part of the social science panel in the ARC. Readers were chosen on the basis of their research expertise and broad knowledge in their subdiscipline. The same readers reviewed all the proposals (between 16 and 25) in their subdiscipline, rated the quality of both the proposal and the

researcher, provided written comments, and were paid a small emolument. Because all readers read all of the grant proposals in their subdiscipline, each had a similar frame of reference from which to evaluate any given proposal. Also, by using a ranking procedure, we eliminated differences in leniency/harshness as a source of disagreement between the ratings of different readers (for a more detailed description of the reader trial implementation, see Jayasinghe, 2003; Jayasinghe et al., 2006).

Single-rater reliabilities were much higher for the reader system than for the traditional ARC approach for both project (.30 vs. .17) and researcher (.63 vs. .24) ratings. Based on an average of 4.3 readers per proposal (the average number used in the traditional ARC peer-review process), the reliability of the researcher ratings was an acceptable .88 for the reader system. For both approaches, the researcher ratings were much more reliable than the project ratings. Also, in the traditional approach, these two ratings were so highly correlated (.79) as to have little discriminant validity, whereas this correlation was substantially smaller (.43) for the reader system. From a practical perspective, this result makes the two ratings more useful. From a theoretical perspective, this result supports our interpretation that the traditional peer-review ratings of grant proposals are substantially biased by a halo effect that has been substantially controlled through application of the reader system.

There are many potential advantages to the reader system beyond the increased reliability of the peer-review ratings of grant proposals. The reader system is streamlined and could provide substantial savings in time for staff (materials are sent to fewer people, so there is no need to maintain large databases) and the academic community (only a few hundred readers would be employed rather than the 6,500 assessors). Also, readers who are carefully selected, committed, and paid to do the job are likely to provide more useful peer evaluations. Their performance could be more closely monitored for quality control and to provide feedback to improve their ratings. Because the readers could meet in a central location and are relatively small in number, it would be possible to devise peer-review training programs that included a detailed, ongoing monitoring of results that would overcome potential problems in the sustainability of benefits associated with such interventions (e.g., Schroter et al., 2004). Furthermore, at the least, the rationale underlying our reader system should have broad applicability to the many forms of the peer-review process in different academic settings as well as those used in the broader public community.

However, there are also potential disadvantages of the reader system for evaluating grant proposals that require further scrutiny. As the number of readers is substantially less than the number of assessors in the traditional approach, there are added concerns about potential areas of bias or conflicts of interest (although these should also be easier to monitor and detect in the reader system). Also, there may be the need for additional outside expertise if, for example, a particular proposal falls outside of the areas of expertise of the readers in a particular subdiscipline. Within our study, the effects of the ranking procedure, the large number of proposals evaluated by each assessor, and the control for response biases were confounded. Hence, further research is needed to determine which of these characteristics are important. Thus, although our results based on the reader system are clearly promising for peer reviews of grant applications, there needs to be further research testing the reader system's applicability and generalizability to other forms of peer review. However, whereas the reader system may be new to the ARC, related approaches have been proposed and given a trial elsewhere, such as the peer-review process of the National Science Foundation (Klahr, 1985) and that of the Heart and Stroke Foundation of Ontario (Hodgson, 1995). Indeed, our reader system also has some features in common with the peer-review process used by some journals that have large editorial boards that conduct most of the reviews with only occasional assistance from external (ad hoc) reviewers.

Directions for Further Research

Clearly, there is a need for more systematic intervention research designed to improve the peer-review process for grant applications, such as our reader system research and the research program by Bornmann, Daniel, and colleagues (e.g., Bornmann & Daniel, 2005; also see related research for the peer review of journal articles by the *British Medical Journal,* e.g., Tite & Schroter, 2006). Because most peer-review research, including our research, is correlational, it provides a weak basis for any causal inferences—particularly in evaluating potential biases. Although 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. Whereas it might be possible to fund grants that were rejected by the peer-review process and to compare outcomes based on these grants with outcomes based on successful grants, this type of study would be ethically dubious and probably unacceptable to funding agencies. There is surprisingly little longitudinal peer-review research evaluating the consistency of results based on the same applicants or the same assessors over time as well as systematically evaluating differences associated with variations in how the peer-review process is operationalized. Finally, we note the relevance of meta-analysis to synthesizing the growing body of peer-review research, as illustrated by the Bornmann et al. (2007) meta-analysis of gender differences in peer-review studies. More generally, authors of peer-review studies, experimental and nonexperimental, need to evaluate critically the validity of their interpretations within a construct validity perspective.

The focus of much of our research program has been on the peer-review process for evaluating grant proposals, whereas other peer-review studies focus on evaluation of journal articles, fellowship applications, and other academic products. A major difference is that journal submissions are likely to be anonymous (although it is often possible to surmise who an author is), whereas the evaluation of the research

track record of authors is typically a critical component in evaluations of grant proposals, particularly fellowship applications and job applications. Hence, potential biases associated with applicant characteristics such as those considered here (gender, age, academic rank, institutional affiliation) are likely to be a more critical feature. Also, grant proposals and fellowship applications are typically evaluated in one stage (although sometimes preliminary proposals are short-listed for consideration in a second stage based on more detailed proposals), whereas journal submissions typically go through at least two cycles of review and revision. Nevertheless, we suspect that many of the issues, concerns, and, perhaps, even the results addressed in our research will generalize to other peer-review applications. More generally, there is a need for peer-review research to more clearly articulate what constitutes the peer-review process, to systematically evaluate how the different components that comprise various peer-review strategies contribute to the effectiveness of peer reviews, and to determine the extent to which results generalize across different applications of the peer-review process.

We have focused on improving the reliability of peer reviews for grant applications, but this research should be seen as a means to improving the review process rather than as an end in itself. Whereas reliability sets an upper limit on validity, increasing reliability does not necessarily enhance validity. Indeed, a common strategy is to include assessors with different perspectives—a strategy that might improve validity but would probably result in apparently less reliable responses. An important limitation in our research program with grant applications, and in peer-review research more generally, is that we had no fully appropriate external criteria against which to validate the outcomes of the peer-review process. We argued that the final panel decision (based on the integration of all available information) was the best criterion available for validating and testing potential biases in the ratings by individual assessors. However, a more external criterion is needed to validate the results of the panel decision itself. Whereas subsequent publications, citation counts, and journal impact values might provide a reasonable basis for validating peer reviews of journal articles (Daniel, 2005) or applications for doctoral or postdoctoral positions (Bornmann & Daniel, 2005), these are less viable options for grant proposals in that there is not a clear one-to-one relation between one particular grant and subsequent publications that might not result until many years after the funding decision. However, at least in terms of ratings of researcher quality, the previous track record of researchers does provide a viable validity criterion (see Marsh & Bazeley, 1999). Nevertheless, the need to find more suitable validity criteria remains a critical issue for research into the review of grant applications and for peer-review research more generally.

REFERENCES

Australian Research Council. (1996). *ARC members' handbook.* Canberra, Australian Capital Territory, Australia: Author.

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

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

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

Callaham, M. L., Baxt, W. G., Waeckerle, J. F., & Wears, R. L. (1998). Reliability of editors' subjective quality ratings of peer reviews of manuscripts. *Journal of the American Medical Association, 280,* 229–231.

Chubin, D. E. (1994). Grants peer review in theory and practice. *Evaluation Review, 18,* 20–30.

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

Daniel, H. D. (2005). Publications as a measure of scientific achievement and scientists' productivity. *Learned Publishing, 18,* 143–148.

Goldbeck-Wood, S. (1999). Evidence on peer review—Scientific quality control or smokescreen? *British Medical Journal, 318,* 44–45.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London: Hodder Arnold.

Grimm, D. (2005, September 23). Suggesting or excluding reviewers can help get your paper published. *Science, 309,* 1974–1975.

Helmstadter, G. (1964). *Principles of psychological measurement.* New York: Appleton-Century-Crofts.

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

Jayasinghe, U. W. (2003). *Peer review in the assessment and funding of research by the Australian Research Council.* Unpublished doctoral dissertation, University of Western Sydney, New South Wales, Australia.

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,* 343–364.

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

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

Jefferson, T. (2001, November 16). Corrections and clarifications. *Science, 294,* 1463.

Jefferson, T., Rudin, M., Brodney, F. S., & Davidoff, F. (2006). Editorial peer review for improving the quality of reports of biomedical studies. *Cochrane Database of Methodology Reviews,* Issue 1, Art. No. MR000016.

Klahr, D. (1985). Insiders, outsiders, and efficiency in a National Science Foundation panel. *American Psychologist, 40,* 148–154.

Lindsey, D. (1978). *The scientific publication system in social science.* San Francisco: Jossey-Bass.

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

Marsh, H. W., & Ball, S. (1989). The peer review process used to evaluate manuscripts submitted to academic journals: Interjudgmental reliability. *Journal of Experimental Education, 57,* 151–169.

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

Marsh, H. W., & Bazeley, P. (1999). Multiple evaluations of grant proposals by independent assessors: Confirmatory factor analysis evaluations of reliability, validity, and structure. *Multivariate Behavioral Research, 34*(1), 1–30.

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

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

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

Schroter, S., Black, N., Evans, S., Carpenter, J., Godlee, F., & Smith, R.

(2004). Effects of training on quality of peer review: Randomised controlled trial. *British Medical Journal, 328,* 673–675.
Schroter, S., Tite, L., Hutchings, A., & Black, N. (2006). Differences in review quality and recommendations for publication between peer reviewers suggested by authors or by editors. *Journal of the American Medical Association, 295,* 314–317.
Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428.
Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London: Sage.
Tite, L., & Schroter, S. (2006). Evidence based publishing. *British Medical Journal, 333,* 366.
Wenneras, C., & Wold, A. (1997). Nepotism and sexism in peer-review. *Nature, 387,* 341–343.
Zuckerman, H., & Merton, R. K. (1971). Patterns of evaluation in science: Institutionalisation, structure and functions of the referee system. *Minerva, 9,* 66–100.