

Peer review process: Assessments by applicant-nominated referees are biased, inflated, unreliable and invalid

HERBERT W. MARSH^{1,2}, NIGEL W. BOND³, & UPALI W. JAYASINGHE¹

¹*Self-concept and Learning Facilitation (SELF) Research Centre, ²Department of Educational Studies, University of Oxford, United Kingdom, and ³University of Western Sydney, Bankstown Campus, New South Wales, Australia*

Abstract

How trustworthy are peer reviews by applicant-nominated assessors (ANAs)? For Australian Research Council (ARC) proposals ($N = 2,330$) with at least one ANA and one assessor nominated by the funding panel (PNAs), ANAs gave substantially higher ratings in all nine discipline panels (covering sciences, social sciences, and humanities). Compared to reviews by PNAs, ANA ratings were less related to ratings by other assessors, 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 PNA for different proposals, ratings in the role of ANA were biased whereas those by the same person in the role of PNA were not. ANA ratings of ARC grant proposals are biased, inflated, unreliable, and invalid, leading the ARC to abandon use of ANAs. Particularly if replicated in other situations, the results have important implications for other evaluations based on ANAs.

Research track records of academic psychologists in Australia, like those of academics in all disciplines throughout the world, are fundamentally influenced by the peer review process. Particularly in psychology departments and other academic settings, it is frequently used to evaluate grant proposals, journal submissions, job applications, promotions, monographs, textbooks, PhD theses, and a variety of other academic products (Bornmann & Daniel, 2005; Chubin, 1994; Cicchetti, 1991; Jayasinghe, Marsh, & Bond, 2001, 2003; Marsh & Ball, 1981, 1989, 1991). More broadly, this process serves a gatekeeper role, acting as the final arbiter of what is valued and acceptable in many areas of public life. Nevertheless, most published studies of the peer review process focus on results from one specific discipline; few take a cross-disciplinary perspective that is necessary to evaluate the generalisability of conclusions. Even though many published notes and comments offer suggestions about how to improve the process, remarkably few provide empirical support. Given the central importance of the peer review process to science, there is surprisingly little evidence of the use of scientific methods to evaluate the peer review process. However, the quantitative social

science research tools used by psychologists (with their focus on reliability, validity, and bias) are uniquely appropriate to evaluate the peer review process.

A problem common to all peer review processes is that the ratings given to the same submission by different assessors typically differ, sometimes substantially. This resulting lack of reliability in peer reviews is, perhaps, the most important weakness of the peer review process. In order to provide a common benchmark, Marsh and Ball (1981, 1989, 1991) defined single-rater reliability as the correlation between two independent assessors of the same submissions across a large number of different submissions; it can also be derived from analysis of variance and multilevel modelling (Jayasinghe, 2004). This single-rater reliability can then be used to estimate the reliability of the mean rating based on varying numbers of raters, using the widely known Spearman–Brown equation (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, Cicchetti (1991)

reported single-rater reliabilities of between .17 and .37 ($Mdn = .33$) based on nine analyses of reviews of submissions to the (American) National Science Foundation, suggesting that the reliability of assessments of grant proposals may be comparable to those reported for journal submissions. Commenting on results such as these, Jefferson (2001) claimed, “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).

As highlighted by Grimm (2005), some journals and funding bodies allow applicants to nominate applicant-nominated assessors (ANAs). However, based on three papers presented at the Fifth International Congress of Peer Review and Biomedical Publication and interviews with journal editors, Grimm found no clear agreement on their value. Existing research does not resolve questions such as, should decision-making bodies encourage use of ANAs, and are authors taking advantage of this option advantaged or disadvantaged? Hence, the purpose of the present investigation was to systematically evaluate the quality of assessments given by ANAs to grant proposals submitted to the Australian Research Council (ARC).

Focusing on a “within proposal” perspective, we compare ratings of the same proposal by ANAs and assessors nominated by the ARC funding panel (PNAs). This allows us to control the many sources of variation associated with a particular proposal (e.g., quality of the proposal). Because the final ARC panel rating of each proposal is based on a critical integration of ratings and written comments by all external assessors, a subsequent rejoinder of these external evaluations by authors of the proposal, and an independent evaluation of the proposal by ARC panel members with appropriate expertise, the final ARC panel rating provides one appropriate criterion against which to validate reviews by ANAs and PNAs. Finally, because 555 of the external assessors served as ANAs on at least one proposal and as a PNA on at least one other proposal, we were able to compare ratings of the same assessor in the role of PNA and ANA: a within-assessor perspective.

Methods

Data used in this study were based on all externally reviewed proposals in the 1996 round (see ARC, 1996). This encompassed 10,023 reviews of 2,331 proposals by 6,233 assessors for nine ARC discipline panels covering all areas of science, social science and the humanities; an average of 4.3 ratings per proposal. Assessor’s country was classified into four regions: Australia (56.6%), North America (19.6%),

Europe (18.7%), and “Other” (Asia, Africa, South America and New Zealand, 5.1%), but ANAs were more likely than PNAs to come from different countries. The dependent variables used in this article are project ratings (overall quality of proposal) and researcher ratings (overall quality of the research team), or a weighted average of these two ratings (.4 for researcher ratings, .6 for project ratings, based on ARC guidelines). Our primary focus is assessor type (ANAs vs. PNAs). Most proposals (75.7%) had one ANA. However, some proposals had no ANAs (18.9%), typically because the authors of the proposal did not nominate an assessor or the assessor did not respond when asked by the ARC, and a few proposals had more than one ANA (5.4%). We compared ratings by ANAs and PNAs with a synergistic combination of single-level analyses (e.g., t tests and repeated-measures ANOVAs) and multilevel analyses with a cross-classified structure to account for the fact that some assessors evaluate more than one proposal (Goldstein, 2003; Jayasinghe, 2004; Jayasinghe et al., 2003).

Results

Do ANAs give systematically higher ratings to the same proposal than PNAs?

For each proposal having at least one PNA and one ANA, the mean project and researcher ratings were calculated separately for PNAs and for ANAs. Paired t tests, based on the mean of ANAs and the mean of PNAs for the same proposal, were used to evaluate the effect of assessor type. This within-proposal perspective is important because it controls for the effect of the proposal by holding it constant. The mean of project ratings of ANAs (87.2) was significantly higher than that of PNAs (80.2), $t(1,888) = 25.9$, $p < .001$. Similarly, the mean researcher rating of ANAs (88.8) was significantly higher than that of PNAs (83.4), $t(1,884) = 25.5$, $p < .001$. Thus, ANAs gave systematically higher ratings than the corresponding PNAs who reviewed the same proposal. Next, we examined whether ratings of ANAs were higher than those of PNAs for each of the nine disciplinary panels in the ARC, using repeated measures of multivariate analysis of variance (Figure 1). For both researcher and project ratings; ANAs gave statistically higher ratings than PNAs ($ps < .001$). Whereas this main effect of assessor type (i.e., PNA vs ANA) interacted significantly with panel type, tests of simple main showed that ANA ratings were significantly higher than PNA ratings in each of the nine panels (all $ps < .001$). Multilevel modelling allows a more flexible approach to the within-proposal perspective in which assessor ratings (Level 1) are nested under proposal (Level 2). Using this alternative

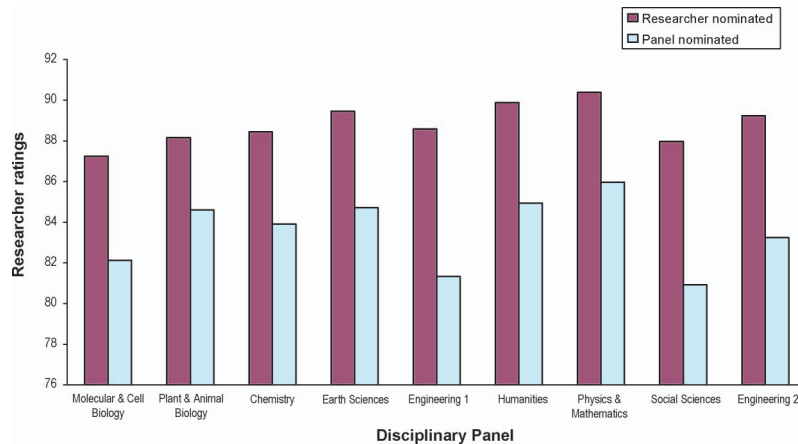


Figure 1. Researcher ratings for applicant (researcher)-nominated assessors and funding panel-nominated researchers in each of the nine discipline areas (results for project ratings show a similar pattern; see Table 1).

approach, we again found that ANA ratings were significantly higher than PNA ratings for both project and researcher ratings, and that the direction (and statistical significance) of this effect generalised over the nine panels representing all science, social science, and humanity disciplines (Figure 1).

Effect of ANAs on reliabilities of project and researcher ratings

The single-assessor reliability (intra-proposal correlation) was calculated using multilevel cross-classified models based on an overall assessment for each assessor (project and researcher ratings were weighted .6 and .4, respectively, as specified in the ARC guidelines). In this approach, the single-assessor reliability increases when variance between assessors within proposals decreases (i.e., there is better agreement among assessors of the same proposal) or variance between proposals increases (i.e., there are larger differences between proposals). The resulting single-assessor reliability was .20 (or .52 based on 4.3 assessors per proposal, the average number of assessors for each ARC grant).

Next we evaluated whether reliability increases when the overall rating is adjusted for the ANA bias. Total variance was reduced (from 1.00 to .933, 6.8%) after adjusting for researcher-nominated assessors. As expected, assessor level variance decreased (from .802 to .723, 9.9%) and proposal level variance increased (from .146 to .162, 11.0%). Therefore, the single-assessor reliability increased (from .20 to .23, 13%) when the overall assessor rating for a proposal was corrected for the bias associated with ANAs. These results demonstrate that the systematic bias associated with ANAs contributes to the unreliability of the peer reviews and that controlling for this effect results in more reliable ratings.

Final panel rating as a criterion

In the ARC peer review process, the best single index of the quality of a proposal is the final panel rating, which is based on a critical review of assessments by all assessors, a critical reading of the proposal by panel members, and a subsequent rejoinder to the external assessments of the proposal by authors of the proposal. Using final panel rating as a validity criterion, we compared the validity of ANAs and PNAs.

In order to compute correlations between final panel ratings and overall assessor ratings (weighted average of project and researcher ratings) by ANAs and PNAs, one ANA (when there were more than one researcher-nominated assessors in the proposal) and up to three panel-nominated assessors of the same proposal were randomly selected for each proposal. Consistent with a priori predictions, final panel ratings were systematically (and statistically significantly, $ps < .01$) less correlated with ANAs ($r = .52$) than with the three PNAs (rs of .62, .61, and .59). These results suggest that ANAs are less valid than PNAs, in relation to final panel ratings.

Analyses of ratings by assessors who reviewed different proposals as an ANA and a PNA

Now we extend the logic of our within-proposal perspective (ANA and PNA ratings of the same proposal) to incorporate a within-assessor perspective. More specifically, 555 assessors reviewed different proposals as both an ANA and a PNA; a total of 1,707 reviews: 1,069 as PNAs and 638 as ANAs. Using these data we ask whether the same assessor gives systematically higher ratings as an ANA than when the same person serves as a PNA.

Initially, we focused on PNA ratings by reviewers who served as both ANAs and PNAs. For each

proposal, mean project and researcher ratings were computed for the remaining PNA assessors who did not review proposals as both an ANA and a PNA (and excluding ratings by any other ANAs). Despite the large N and powerful test, the results showed that there was no significant difference between the ratings of these assessors (when they rated proposals as PNAs) and the remaining PNAs for either project ($p = 0.14$) or researcher ($p = 0.79$) ratings. Thus, when in the role of PNA, these assessors (who were also ANAs on different proposals) did not differ from other PNAs. Next, we focused on ANA ratings by these same assessors who served as both ANA and PNA. For each proposal, mean project and researcher ratings were computed for the remaining PNA assessors who did not review proposals as both PNAs and ANAs. Both project ($t(637) = 10.8$, $p < 0.001$) and researcher ($t(637) = 10.7$, $p < 0.001$) ratings were substantially higher than those of the remaining assessors when these assessors reviewed proposals in the role of ANAs. In summary, these assessors rated proposals as substantially higher than other assessors when in the role of ANA, but ratings by these same assessors did not differ from other reviewers when they served as PNAs.

Does the bias in ratings by ANAs vary as a function of nationality?

In a multilevel analysis, three dummy variables were constructed for North American (United States and Canada), European, and Other (Asian, African, New Zealand and South American) assessors.

Australian assessors, who gave the lowest ratings, were considered the base (left out) category. When only country was considered (Models 1 and 3 in Table 1), project and researcher ratings by Europeans, “others” and particularly North Americans were all significantly higher than Australians’ ratings. Next we added assessor type (1 = ANA, 0 = PNA; Models 2 and 5) to the analyses, again showing that ANAs gave significantly higher ratings than PNAs. However, there was a substantial decrease in the size of country effects after controlling for assessor type. Whereas ratings by North Americans were still substantially higher than those by Australians, there were no significant differences between Australians and Europeans. Assessors from other countries gave higher project ratings than Australians, but did not differ for researcher ratings. Hence, controlling for assessor type substantially reduced the apparent sizes of country differences. The reason is that ANAs were disproportionately from outside of Australia so that part of the apparently higher ratings by assessors from other countries was due to the ANA bias. Hence, after controlling for this bias the only substantial effects of country were the systematically higher ratings by North American (ANA and PNA) assessors. For example, much of the difference between Australian and non-Australian assessors (Figure 2) was due to ANAs. North American assessors (PNAs and ANAs) gave substantially higher ratings than assessors from other countries and the positive bias in ANAs was particularly large for the North American assessors. Whereas controlling for the country of assessors

Table 1. Multilevel models for assessor type as a function of country

Source	Project ratings			Researcher ratings		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Fixed						
Main effects						
Type		.232 (.009)	.221 (.009)		.212 (.009)	.200 (.009)
North American	.135 (.010)	.086 (.010)	.078 (.010)	.149 (.010)	.102 (.010)	.096 (.010)
European	.037 (.010)	.005 (.010)	.007 (.010)	.013 (.010)	-.016 (.009)	-.016 (.009)
Other	.043 (.010)	.028 (.009)	.028 (.009)	.024 (.009)	.010 (.009)	.010 (.009)
Interactions						
Type.N Am			.039 (.009)			.038 (.008)
Type.Eur			.015 (.009)			.025 (.009)
Type.Other			.015 (.009)			.017 (.008)
Random						
Level 2	.146 (.010)	.161 (.010)	.161 (.010)	.206 (.012)	.222 (.012)	.221 (.012)
Level 1	.805 (.013)	.741 (.012)	.739 (.012)	.728 (.012)	.674 (.011)	.672 (.011)
-2*loglikelihood	26,843.5	26,219.5	26,198.1	26,367.4	25,799.1	25,774.7

Note: N America, European and Other = dummy variables for assessors from North America, Europe and “Other” regions respectively. Type = dummy variable for assessor type: applicant-nominated assessor (ANA) = 1, panel-nominated assessor (PNA) = 0. Type.NAm, Type.Eur, and Type.Other = interactions between assessor type and country. Parameter estimates more than twice the size of the corresponding standard errors (in parentheses) are statistically significant ($p < .05$).

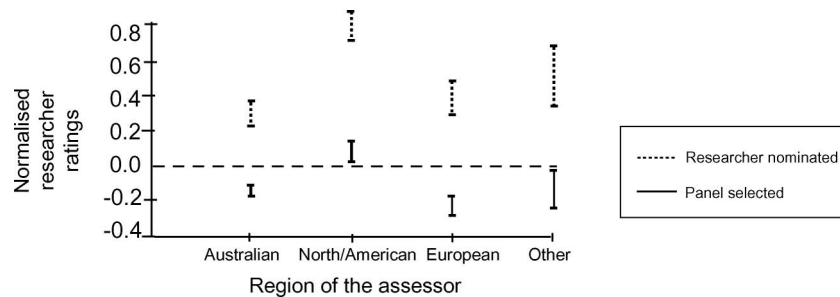


Figure 2. Confidence intervals for researcher ratings by applicant (researcher)-nominated assessors and panel-nominated assessors from different countries (results based on project ratings show a similar pattern of results).

reduced the size of the ANA bias slightly, the effect was still highly significant and generalised well over assessors from different countries.

Our results are consistent with observations by Wood (1997), who noted that ARC discipline panel members reported problems of variation in scoring practices between assessors from different countries (e.g., assessors from the United States being more lenient than those from Germany or the United Kingdom). However, our findings also demonstrate that part of the problem is associated with assessor type.

Discussion

Overall, 20% of ARC assessors were ANAs and 81% of proposals had at least one ANA. Both project and researcher ratings by ANAs were significantly higher than those of PNAs. This positive bias in ANA ratings generalised over different panels covering the sciences, social sciences, and humanities, and generalised over responses by assessors from different countries. Assessor level variance decreased by 7% after adjusting for the higher ratings given by ANAs, so that there was better agreement between assessors after controlling the ANA bias. Furthermore, the same assessors gave higher ratings when they rated proposals in the role of ANA but not when the same person made ratings in the role of PNA. Controlling for the ANA bias improved reliability by 13%.

Our results indicate that ANA assessments of ARC grant proposals are biased, less reliable and less valid than PNAs. 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. Because of the diverse ways in which we evaluated ANAs and the generalisability of our results over large numbers of assessors from different academic disciplines and from different countries, the generalisability of our results is strong. An important qualification of our results is that there are apparently no other statistically rigorous studies of the effects of ANAs on peer review ratings, and so it is difficult to ascertain

the generalisability of our results. Although clearly beyond the scope of the present investigation, it would also be informative to know how editors and decision makers utilised ANAs in other contexts in an attempt to counter problems with them such as those clearly identified in the present investigation. Nevertheless, particularly if replicated in other situations, our results also have policy implications not only for the ARC, but also for many other situations in which applicants are able to choose their external assessors.

It is also important to place our findings within the larger context of the quality of the peer review process. Reliability estimates of ARC peer review ratings (e.g., .52 based on responses by an average of 4.3 assessors) fall far short of acceptable standards of reliability. However, these results are not unlike those reported in other evaluations of the reliability of peer reviews of journals and grant proposals reviewed earlier, in which reliability estimates are consistently below acceptable standards broadly endorsed in psychological research. Furthermore, it is important to note that these reliability estimates are likely to underestimate that of the reliability of the final decision taken by panels who independently review the proposals, the research track records of proposal authors, assessors' written comments, and a one page rejoinder by the authors of proposals in response to the external reviews – as well as the numerical ratings by external assessors considered here. Furthermore, initially culled proposals that did not satisfy technical proposal requirements, or were deemed to be un-fundable in a preliminary evaluation, were not sent out for external review (and, thus, not included in this study). If these initially culled proposals had been included, reliability estimates would likely have been higher. Although the reliability of ARC proposal evaluations was not the focus of the present investigation per se (Jayasinghe, 2004), one way to increase the reliability of peer review ratings of grant proposals is to exclude ANAs (for further discussion of other possible strategies, see Jayasinghe, 2004). Because there is so little rigorous research on ANAs it is difficult to evaluate the generalisability of our results to

other applications in which decisions are based on judgments made by peers nominated by the person being evaluated. We suspect, however, that in other situations that do not place so many constraints on the objectivity of the reviewers (e.g., letters of reference for job applications), ratings by ANAs are likely to be even more biased, less reliable, and less valid.

Acknowledgements

This research was funded in part by a grant from the Australian Research Council. We would like to express our thanks to the Australian Research Council and to Professor Max Brennan, former Chair of the Australian Research Council, for assistance in providing the data used in this study. A more detailed description of materials, statistical analyses, and overall research project are available from the Jayasinghe (2004) PhD thesis that is available electronically (<http://self.uws.edu.au/Theses/Jayasinghe/list.htm>); also see Jayasinghe, Marsh & Bond, 2001; 2003).

References

- Australian Research Council (ARC). (1996). *ARC members handbook*. Canberra: Author.
- Bornmann, L., & Daniel, H. D. (2005). Selection of research fellowship recipients by committee peer review. Reliability, fairness and predictive validity of Board of Trustee decisions. *Scientometrics*, *63*, 297–320.
- 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. *Behavioural and Brain Sciences*, *14*, 119–135.
- Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London: Hodder Arnold.
- Grimm, D. (2005). Suggesting of excluding reviewers can help get your paper published. *Science*, *309*, 1974.
- Jayasinghe, U. W. (2004). Peer review in the assessment and funding of research by Australian Research Council. Doctoral dissertation submitted to the University of Western Sydney, Australia. <http://self.uws.edu.au/Theses/Jayasinghe/list.htm>
- 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 multi-level 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.
- Jefferson, T. (2001). Corrections and clarifications. *Science*, *294*, 1463.
- 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. *Behavioural and Brain Sciences*, *14*, 157–158.
- Wood, F. Q. (1997). *The peer review process* (Report No. 54). Canberra: National Board of Employment, Education and Training.