J. R. Statist. Soc. A (2003)
166, *Part* 3, *pp.* 279–300

A multilevel cross-classified modelling approach to peer review of grant proposals: the effects of assessor and researcher attributes on assessor ratings

Upali W. Jayasinghe, Herbert W. Marsh and Nigel Bond

University of Western Sydney, Australia

[Received January 2001. Final revision July 2002]

Summary. The peer review of grant proposals is very important to academics from all disciplines. Although there is limited research on the reliability of assessments for grant proposals, previously reported single-rater reliabilities have been disappointingly low (between 0.17 and 0.37). We found that the single-rater reliability of the overall assessor rating for Australian Research Council grants was 0.21 for social science and humanities (2870 ratings, 1928 assessors and 687 proposals) and 0.19 for science (7153 ratings, 4295 assessors and 1644 proposals). We used a multilevel, cross-classification approach (level 1, assessor and proposal cross-classification; level 2, field of study), taking into account that 34% of the assessors evaluated more than one proposal. Researcher-nominated assessors (those chosen by the authors of the research proposal) gave higher ratings than panel-nominated assessors chosen by the Australian Research Council, and proposals from more prestigious universities received higher ratings. In the social sciences and humanities, the status of Australian universities had significantly more effect on Australian assessors than on overseas assessors. In science, ratings were higher when assessors rated fewer proposals and apparently had a more limited frame of reference for making such ratings and when researchers were professors rather than non-professors. Particularly, the methodology of this large scale study is applicable to other forms of peer review (publications, job interviews, awarding of prizes and election to prestigious societies) where peer review is employed as a selection process.

Keywords: Australian Research Council; Cross-classified models; Grant proposal funding; Interrater reliability; Multilevel modelling; Peer review process

1. Introduction

The peer review process is highly valued and frequently used for evaluating the academic merits of grant proposals, journal submissions, academic promotions, job applications, monographs, text-books, doctoral theses and a variety of other academic products (Cicchetti, 1991). In the peer review process, assessors evaluate the attributes of academic documents according to specific criteria and then decision makers (e.g. editors or granting bodies) use these assessments and, perhaps, apply additional criteria to decide whether or not to accept a manuscript, to fund a proposal, to grant a promotion, etc. Clearly, the peer review process is extremely important to the academic community.

The precise details of the peer review process, however, vary widely. Nevertheless, whichever, peer review process is used the ratings that are given to the same submission by different assessors

Address for correspondence: Herbert W. Marsh, SELF Research Centre, University of Western Sydney, Bankstown Campus, Locked Bag 1797 Penrith South DC, NSW 1797, Australia.
E-mail: h.marsh@uws.edu.au

0964–1998/03/166279

normally differ from each other—sometimes substantially. To provide a common bench-mark, Marsh and Ball (1981, 1989, 1991) defined the single-rater reliability as the correlation between two independent assessors of the same submissions across a large number of different submissions. This single-rater reliability can then be used to estimate the reliability of the mean rating based on varying numbers of raters. On the basis of their review of previous research, Marsh and Ball (1989) reported that the mean single-rater reliability for journal submissions was only 0.27. Although there is little research on the reliability of assessments of grant proposals, Cicchetti (1991) reported single-rater reliabilities of between 0.17 and 0.37 (median 0.33) based on nine analyses of reviews of (American) National Science Foundation grant submissions. This suggests that the reliability of assessments of grant proposals may be comparable with those reported for journal submissions, but there is too little research on grant proposals to draw any clear conclusions.

By way of countering the criticisms that referees selected by the editors are biased, some journals allow their authors to nominate at least one of the referees (Daniel, 1993). A similar policy was adapted by the Deutsche Forschungsgemeinschaft in nominating reviewers (Neidhardt (1988), as cited in Daniel (1993)). Yet no reports are available regarding the effect of author-nominated reviewers on the reliability, fairness or validity of manuscript review (Daniel, 1993).

Although there are many opinions about the peer review of research funding, little quantitative analysis is available on its operation or outcomes (Osmond, 1983; Garfield, 1986). Naylor (1994) pointed out that manuscripts are more commonly accepted with revision but research proposals are rarely given this opportunity. He further suggested that applicants for research funds should be given an opportunity to resubmit within weeks of a decision or re-entered in the next competition. The peer review process for grant proposals also differs from that used for manuscripts in at least one important way which alters the review process. Assessors for academic journals may not know the name of the authors of manuscripts that they are asked to review (i.e. some variation of a blind review process is used) and are only asked to evaluate the quality of the manuscript. In contrast, the typical peer review process for assessing grant proposals requires assessors to evaluate the research track record of the research team, on the basis of personal knowledge and track record summaries provided, as part of the grant application. Hence, particularly in the case of grant proposals, there has also been considerable interest in what characteristics of the researchers (e.g. age, gender, academic rank and institutional affiliation) are associated with peer review assessments and whether these constitute a potential bias in the process.

For example, Emery *et al.* (1992) noted that the age of the author had no effect on success in winning a US National Institute of Health first independent research support and transition award. Further, the difference in the rate of funding for male and female applicants was not statistically significant. Cole *et al.* (1977) found that professional age had no significant influence on the ratings and final outcome of the US National Science Foundation peer review system. Hammel (1980) established that scientific productivity increased with age, with some evidence of flattening, but not necessarily declining, with age. Bazeley (1998) reported that professors were more likely to be funded than grant applicants who were below the level of professor. Eminent authors were likely to be more successful in achieving publication and in receiving quick reviews (Lindsey, 1976, 1977; Mahoney, 1985; Zuckerman and Merton, 1971). Therefore, owing to 'accumulative advantage' successful scientists tend to be more productive, whereas less productive scientists are likely to decline further in their productivity (Allison and Stewart, 1974). Cole *et al.* (1977), however, emphasized that it is to be expected that leading scientists whose research work had been valued highly by other scientists are more likely to submit highly rated proposals.

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 and behavioural sciences. For example, Lindsey (1978) noted that there was a lack of paradigm development and consensus in the social and behavioural sciences. He further emphasized that the social and behavioural sciences deal with social reality that is based on political and cultural assumptions. Therefore, he concluded, they are less able to use value-free objective criteria than in the physical sciences. Furthermore, the industrial and technological success of the physical and biological sciences directly translated their benefits into serviceable programmes or products whereas the process has been less spectacular in social and behavioural sciences (Lindsey, 1978). According to Cicchetti (1991), there did not appear to be any formal studies of the reliability of peer review undertaken for manuscripts or abstract submissions to journals in the physical sciences. He emphasized, however, that there was a belief that levels of agreement between referees were substantially higher for journals in the physical sciences than in other areas. This conclusion appears to have been based on a brief analysis made by Zuckerman and Merton (1971) about *Physical Review*, which is one of the most prestigious journals in the physical sciences. To pursue this distinction, in the present investigation we analysed separately the results from the social science and humanity panels and from the science panels.

1.1. Summary of analytic approaches used in previous research

Typically, analyses in previous research on peer review studies were carried out by using single-level models using analytic techniques such as correlation, multiple regression, analysis of variance, discriminant function analysis and Z-tests for proportions. Cole *et al.* (1977) tested matching hypotheses of reviewer and applicant departments and the relative eminence of the reviewer and applicant by using analyses of variance. Similarly, gender bias in peer review was examined by using Z-tests for proportions (Gilbert *et al.*, 1994). However, characteristics associated with assessors (in our research, assessor ratings are at level 1, cross classified by assessors and proposals) that are clustered into proposals, pose special problems for analysis. These include the appropriate levels of analysis, aggregation bias, heterogeneity of regression and associated problems of model misspecification due to a lack of independence between measurements at different levels. Thus, it is inappropriate to pool ratings without regard for the correlation between ratings from the same assessor for different proposals, or between ratings from different assessors for the same proposal. Ratings for proposals or from assessors in the same field of study may also be correlated. Multilevel analyses allow researchers to consider multiple units of analysis within the same analysis simultaneously.

Another complication with peer review data is that it is typical that some or most assessors review more than one academic product (e.g. manuscripts and funding proposals), but no one assessor typically reviews all the academic products. Hence, even if multilevel analyses were used, the analyses would be compromised by this problem of cross-classification. There have been a few attempts to deal with the problem in single-level analyses. For example, Marsh and Ball (1981) included a ‘fixed effect’ for each of the reviewers who evaluated a relatively large number of journal submissions to make a correction for the rater response bias of each reviewer and analysed the data by using analysis of variance. Importantly, in the present investigation, we used a multilevel cross-classified model in which ratings were nested within assessors, and ratings within proposals. However, assessors and proposals were non-nested since an assessor may rate more than one proposal and a proposal may be rated by more than one assessor. The use of multilevel cross-classified models is potentially important in our study because one in three assessors rated more than one proposal. In summary, the multilevel approach has many advantages over generally inappropriate alternative approaches that have typically been used in peer review studies. Because peer review data have a multilevel structure, the study provides an important methodological advantage over most research done in peer review using single-level models.

1.2. The present investigation: aims and objectives

The aims of this paper are as follows.

(a) The first aim is to compute the reliabilities for external reviewer ratings of the Australian Research Council (ARC) large grant programme. There is little research on the reliability of ratings of grant proposals, but previously reported single-rater reliabilities were disappointingly low.

(b) Of particular interest is whether researcher-nominated external reviewers give systematically higher or lesser reliable ratings. A critical concern in the ARC peer review process is whether there are systematic differences in the ratings that are offered by those external reviewers who were nominated by the applicant and those nominated by the ARC panels.

(c) A large number of external reviewers rated only one proposal per year; approximately a third of the assessors rated more than one and a few evaluated many proposals (up to 26). It is important to examine whether the number of proposals reviewed by each external reviewer has any effect on the quality of ratings.

(d) A large number of external reviewers from countries other than Australia were asked to review ratings. An important issue is whether ratings by overseas external reviewers differ from each other and from ratings given by Australian external reviewers.

(e) The question of whether the gender of either the external reviewer or the researcher has an effect on ratings is of general interest. Of particular interest is whether there is a gender bias in ARC ratings (i.e. a researcher gender by external reviewer gender interaction).

(f) An important policy question is whether the age of researchers has any effect on external reviewer ratings.

(g) It is appropriate to investigate whether there is any effect of well-established senior researchers and/or prestigious older universities on ratings of the proposal.

(h) We wish to test Lindsey's (1978) hypothesis that ratings are more reliable in the sciences than in the social sciences and humanities and to evaluate whether the effects of researcher and assessor characteristics differ in this broad discipline classification.

The structure of the paper is as follows. Section 2 summarizes briefly the procedure that is adopted by the ARC in the allocation of large research grants. The discussion of data and statistical modelling is given in Section 3. Section 4 summarizes the results by using tables and figures. Section 5 provides a discussion of important results and the final section outlines implications on policy of the study and future research directions.

2. Australian Research Council large grant allocations

The ARC has established a peer review process for the evaluation of grant proposals submitted to its large grants scheme (Australian Research Council, 1996). In Australia, the ARC is the main research funding agency for non-medical research. Applicants submitted proposals according to a prescribed format to one of nine discipline panels (Molecular and Cell Biology, Plant and Animal Biology, Chemistry, Earth Sciences, Engineering 1 (electrical and computing), Engineering 2 (civil, mechanical and chemical), Humanities, Physics and Mathematics, and Social Sciences; see Australian Research Council (1996)) that cover the wide range of disciplines which are eligible for funding through this scheme. In an interesting variation to the typical peer review process, applicants were invited to nominate the assessors whom they would like to evaluate their proposals (researcher-nominated assessors) and a justification about why the researcher-nominated assessors were appropriate. Applicants were also given the option of nominating assessors who they felt should not evaluate the proposal.

In a preliminary examination, two members from each panel scrutinized the proposal and determined whether the proposal was sufficiently competitive to be sent out for further assessment or culled. In 1996, these initial culls accounted for around 22% of the total applications. The remaining proposals were sent to external assessors for assessments. The two members who had read the proposal chose five external assessors and two reserve assessors. The initial group of five assessors typically included one external reviewer nominated by the researcher (researcher-nominated assessor). Other external assessors (panel-nominated assessors) were selected either from professionals known to panel members through their professional associations or from an extensive database of potential reviewers and their areas of expertise, compiled by the ARC. Reserve assessors were contacted when the original assessors had not completed their assessments by a cut-off date. (For further discussion of this process, see Bazeley (1998), Bond and Hesketh (1998) and Australian Research Council (1996)).

Assessors were asked to provide two global ratings on a scale of 1–100—one of the quality of the proposal itself (the project rating) and one of the research track record of the researcher team (the researcher rating). These were supplemented with more specific ratings (on a scale of 1–7) about the proposal and a separate evaluation of the research track record of each researcher on the research team. Assessors were also asked to provide written comments that were used as part of the feed-back that was given to applicants. (For more detail, see Australian Research Council (1996).) The total (overall) score representing each proposal for the 1996 round, as computed by the ARC, was a weighted average of the global project rating (weighted 0.6) and the global researcher rating (weighted 0.4) of the proposal, averaged across all external assessors.

After the external assessments had been completed, applicants were given an opportunity to make a one-page response to the assessors' reports received for their projects. Panels would then meet again to consider the assessors' reports and the applicants' responses to all the material that was available for each proposal. Although the committee could raise or lower the weighted average response by external assessors in response to comments by the proposal authors and their assessment of the materials, typically they did not change the ratings at all or only by a small amount. The panels then assigned a final (single) rating (the committee score) for the proposal. The applications were then ranked according to the committee scores. (The weighted average of external assessors before adjustment by the ARC panel correlated 0.95 with the final committee score after adjustment.) Within each panel, the funds for proposals were allocated starting from the best-ranked proposal and working down until the allocated funding was exhausted (Bond and Hesketh, 1998).

Marsh and Bazeley (1999) and Bazeley (1998) considered data from a commissioned study of opportunities for early career researchers (Bazeley *et al.*, 1996). Apparently, this small study was the first external evaluation of ARC assessments based on data provided by the ARC. Consistent with the overall design of the study of Bazeley *et al.* (1996) data were considered from a sample of academic disciplines (physics, engineering, psychology, history, nursing and social work). Of the 488 proposals that were considered for funding in 1995, 175 were not externally reviewed, 202 were sent to external reviewers but not funded and 111 were actually funded. For the proposals that were sent to external reviewers, the estimated reliabilities of the mean of four independent assessors' ratings were 0.49 for the quality of the proposal and 0.63 for the quality of the research team (Marsh and Bazeley, 1999). For these data (Bazeley, 1998), female researchers were under-represented (11.4% of first researchers were women) and success rates were slightly lower for women (18.3%) than for men (21.2%). Bazeley suggested that the gender imbalance was due to the lack of seniority of women in the academic system. Professors were more likely to be funded than researchers who had a lower academic rank. More established researchers, however, were not affected by including younger researchers in the team. Age had no significant effect on obtaining research grants and younger researchers were not disadvantaged. As noted by some ARC panel members a specific problem in using assessors from other countries is the variation in scoring practices between assessors from different countries (Wood, 1997). For example, Wood reported that assessors from America are generally considered to be more lenient than those from the UK or Germany.

3. Methods

3.1. Data and variables

Data for the present investigation are based on all externally reviewed proposals from the 1996 round of the ARC large grants scheme. Thus, the data consisted 2331 proposals that were rated by 6233 external assessors (653 or 22% of the proposals, which were culled in a preliminary internal evaluation in which they were judged to be ineligible or uncompetitive and, thus, not sent out for review by external assessors). The external assessors provided a total of 10023 reviews.

The response variables that were used in the different analyses presented in this paper are two global overall ratings (the overall quality of the proposal and of the research team). Characteristics of the assessor and the first-named researcher of the research team to be considered included nomination (assessor nominated by the researcher or nominated by the ARC panel), the number of proposals reviewed by each assessor, region (country) of the assessor, gender of the assessor and researcher, age of the researcher, academic title of the researcher and the university status of the researcher (Table 1).

As stated previously, a decision was made to analyse data from social science and humanities panels separately from data from science-based panels (Molecular and Cell Biology, Plant and Animal Biology, Physics and Mathematics, Chemistry, Earth Sciences, Engineering 1 (electrical and computing) and Engineering 2 (civil, mechanical and chemical)).

3.2. Statistical procedure

Peer review ratings may have a hierarchical structure if each reviewer reviews only one submission or a cross-classified structure if some reviewers review more than one submission. The response variables can be either categorical (e.g. accept, reject or resubmit for journal articles or to fund or not to fund for grant proposals) or continuous (e.g. assessor ratings for the quality of grant proposals). Fitting single-level regression models is typically inappropriate if the data have a hierarchical or cross-classified structure. Fitting single-level models would only be appropriate if the estimated intraproposal correlation was not statistically significant (i.e. the ratings were completely unreliable), but this is unlikely to be the case. Furthermore, the estimation of correct standard errors requires a consideration of an appropriate multilevel structure of the data (Goldstein, 1995; Goldstein *et al.*, 1998).

In the multilevel analyses, it is not necessary to have the same number of lower level units within each higher level unit for the data to be balanced (Goldstein *et al.*, 1998). For example, the number of reviewers assigned to evaluate an outcome (proposal, manuscript, job applicant, etc.) typically varies from case to case. Because peer review studies typically have unbalanced

Table 1. Summary description of the data†

Variable	*Assessors*		*Ratings*	
	Number	*%*	*Number*	*%*
Assessor attributes				
Gender of the assessor				
Male	4112	90.5	7111	90.2
Female	430	9.5	772	9.8
Researcher-nominated assessors	2016	32.4	2016	20.1
Region of the assessor				
Australia	3429	56.6	6296	64.2
North America	1186	19.6	1511	15.4
Europe	1132	18.7	1555	15.8
Other regions	311	5.1	449	4.6
Number of proposals reviewed by each				
1 proposal	4100	65.8	4100	40.9
2 proposals	1241	19.9	2482	24.8
More than 2 proposals	892	14.3	3441	34.3

	First researchers	
	Number	*%*
Researcher attributes		
Gender of the first researcher		
Male	1964	84.3
Female	365	15.7
Title of researcher		
Professor	540	23.2
Non-professor	1789	76.8
University status of first researcher		
Older, more prestigious (pre-1987) universities	1999	87.1
Newer (post-1987) universities	296	12.9
Age of first researcher		
Young (less than 40 years)	597	25.6
Older (40 years or more)	1734	74.4

†All percentages were calculated excluding missing values. The total number of cases may differ from one category to the next depending on the missing values of the category.

data, the ability to handle unbalanced data is an important advantage of multilevel modelling. This is in contrast with traditional repeated measures multivariate analysis of variance, which requires balanced data.

In summary, the multilevel approach offers many advantages over generally inappropriate alternative approaches that are typically used in peer review studies. Because this is apparently the first study that has used multilevel modelling to analyse peer review data, the multilevel modelling approach illustrated here provides an important advance over most research that has been done on peer review.

3.3. Multilevel models

In the ARC data for 1996, some assessors reviewed more than one proposal. Thus, there was an assessor and proposal cross-classification at level 1 and fields of study at level 2. In the ARC

1996 round there were up to eight assessors for each of 2331 proposals and 34% of assessors rated more than one proposal. There were a total of 144 fields of study employed by the ARC (although a few of them did not have any proposals in 1996) that were then assigned to one of the nine discipline panels. The Physics and Mathematics Panel had the highest number of fields of study with 31 fields of study and the Molecular and Cell Biology Panel had the lowest number of fields of study, with three in the 1996 round. The influence of any remaining panel effect (within social science or science) on ratings was controlled by the presence of field of study as the top level in the models. Normal plots of residuals for both project and researcher ratings showed no evidence of outliers.

In extensive preliminary analyses, the coefficient of each explanatory variable was allowed to vary randomly, one at a time at the assessor and field of study levels, for each of the response variables. Little or no evidence of variation in any of these coefficients was found. As the fixed and random coefficients at levels 1 and 2 were nearly unchanged by the inclusion of random coefficients, random-intercept models were fitted. The variance components (or random-intercept) model assumes that the only variations between proposals and between fields of study are in the intercept of the model.

The dependent variable is denoted by $Y_{(ij)k}$, referring to the rating given by assessor i to the jth proposal from kth field of study. Thus, for example, a two-level multilevel cross-classified model with assessor and proposal cross-classification at level 1 and field of study at level 2 can be written as

$$Y_{(ij)k} = X_{(ij)k}\beta + u_k + e_{ik} + e_{jk}$$

where $X_{(ij)k}$ is a vector of explanatory variables defined at the assessor or researcher level and the assessor and proposal effects are modelled by the level 1 random variables e_{ik} and e_{jk} and level 2 random variable u_k. The level 1 and level 2 variances of the model are given by $\mathrm{var}(e_{ik}) = \sigma^2_{ei}$, $\mathrm{var}(e_{jk}) = \sigma^2_{ej}$ and $\mathrm{var}(u_k) = \sigma^2_u$. The random effects were assumed to be normally distributed.

We fitted cross-classified models by creating a dummy variable for each proposal and allowing the coefficients of these to vary randomly at the proposal level with an equal variance. Further details of multilevel cross-classified models are available elsewhere (Goldstein, 1995; Rasbash *et al.*, 1999; Rasbash and Goldstein, 1994; Snijders and Bosker, 1999). All cross-classified models fitted in this paper are based on the iterative generalized least squares procedure in MLwiN (version 1.1) (Goldstein *et al.*, 1998).

3.4. Reliabilities of Australian Research Council ratings

The single-rater reliabilities or intraproposal correlations were derived from analysis of variance (Winer (1971) and Shout and Fleiss (1979); see Marsh and Ball (1981, 1989)) and multilevel modelling (Goldstein, 1995). When different sets of reviewers evaluate each proposal, the intraproposal correlation coefficient (ρ) derives from an analysis-of-variance model across reviewers:

$$\rho = \frac{\mathrm{MSS} - \mathrm{MSE}}{\mathrm{MSS} + (N-1)\,\mathrm{MSE}}$$

where MSS is the mean sum of squares between proposals, MSE is the mean-square error and N is the number of assessors per proposal.

In multilevel modelling, the intraproposal correlation is defined as the correlation between two independent ratings of the same proposal. In our cross-classified model the correlation between two ratings for the same proposal is given by

$$\rho = \mathrm{corr}(y_{(i'j)k}, y_{(ij)k}) = \frac{\sigma_u^2 + \sigma_{ej}^2}{\sigma_u^2 + \sigma_{ej}^2 + \sigma_{ei}^2}$$

where σ_u^2 is the between field of study variance, σ_{ej}^2 is the between-proposal variance and σ_{ei}^2 is the between-assessor variance.

The intraproposal correlation is sometimes called a single-rater reliability in the peer review literature (e.g. Marsh and Ball (1981)). The reliabilities of ratings of the 1996 ARC round were computed by substituting the single-rater reliabilities ρ computed using the variance estimates from the base-line variance component cross-classified models and the average number of assessors (N) in the Spearman–Brown formula (see Marsh and Ball (1981)):

$$r_m = \frac{N\rho}{1 + (N-1)\rho}.$$

3.5. Significance of parameters

If the ratio of the parameter estimate to its standard error is greater than 1.96 (the critical value under a normal distribution) then the parameter estimate is significantly different from 0 (Goldstein *et al.*, 1998). When two models are nested (i.e. the parameters estimated in one model are a proper subset of those in a second model), the difference of deviances of the two models (the χ^2-difference with degrees of freedom equal to the difference in the number of parameters estimated) can, under appropriate conditions, be used to test whether the difference between the two models is statistically significant.

3.6. Modelling procedures

Some preliminary cross-classified analyses were conducted. The academic title of the assessor was not statistically significant and, thus, not considered further. Although the cubic age effect of assessor age had a marginally significant effect, we excluded the age of the assessor from further analyses as the dates of birth of 74% of assessors were not available. The effect of the number of researchers in the research team was also not significant. The variables that were non-significant in the preliminary analyses were not considered further.

The characteristics that are considered in this paper are the nomination of assessor (1 for researcher nominated; 0 for panel nominated), the region of the assessor, the number of proposals reviewed by each assessor, the gender of the assessor (1 for females; 0 for males) and the first researcher (1 for females; 0 for males), the academic title of the first researcher of the research team (1 for professor; 0 for others) and the university status of the first researcher of the team (1 for pre-1987 universities; 0 for post-1987 universities). The region of the assessor was represented by American assessors (1 for North American assessors; 0 for others), European assessors (1 for European assessors; 0 for others) and assessors from other regions (1 for Asian, African, New Zealand and South American assessors; 0 for others) where Australia was the base-line or ‘left-out’ category. Preliminary analyses showed a significant cubic (but no linear or quadratic) effect of the first researcher’s age on ratings. Thus, only the cubic effect of age was included in the analyses.

In the present investigation, we examined the simultaneous influence of assessor and researcher characteristics on the project (the overall quality of proposal) and researcher (the overall quality of the research team) ratings. Multilevel cross-classified models were used with project and researcher ratings as the response variables and assessor and researcher attributes as the explanatory variables.

The response variables project and researcher ratings have skewed distributions with more than 40% of ratings being 87 or above out of 100. Hence, project and researcher ratings were transformed to normal scores (Goldstein, 1995; Goldstein *et al.*, 1998). Continuous explanatory variables were standardized. In general, we assumed that the probability of missing data was independent of any of the random variables used in the multilevel models (see Rasbash *et al.* (1999) for details).

4. Results

4.1. Reliabilities of Australian Research Council ratings

The single-assessor reliabilities of global ratings calculated from the multilevel cross-classified models were marginally higher than those produced by the techniques of analysis of variance. Of greater relevance, however, was the finding that reliability estimates based on cross-classification models were also marginally larger than those based on models without cross-classification. This reflected that in cross-classified models individual assessor effects (leniency or stringency) were taken into account. Furthermore, in supplementary analyses, we found that the reliabilities were systematically higher for assessors who evaluated more proposals.

The single-rater reliabilities in social sciences and humanities were 0.179 ($p < 0.001$) for the quality of the proposal (the project rating) and 0.255 ($p < 0.001$) for the quality of the research team (the researcher rating). The corresponding reliability estimates for science were 0.167 and 0.223 for the project and researcher ratings respectively. The reliability estimates for the project and researcher ratings, based on the average number of assessors per proposal (4.1), were 0.47 and 0.58 for social sciences and humanities and, based on the average of 4.2 assessors, 0.46 and 0.55 for science respectively. These results indicated that ratings in science were certainly no more reliable than ratings in social sciences and humanities—thus refuting the Lindsey (1978) hypothesis. Results in Fig. 1 demonstrate that confidence intervals around the estimated proposal effects are wide, reflecting the substantial amount of unreliability and probable error in the ratings of each proposal.

4.2. Gender effects in the peer review of grant proposals

It is important to examine whether researchers are disadvantaged as a function of their gender and whether this effect varies as a function of the gender of the assessor. Of particular interest is a ‘gender matching hypothesis’, suggesting that female assessors give systematically higher ratings to female researchers and that male assessors give systematically higher ratings to male researchers. Because of our interest in this question and as there was a substantial number of missing values (23%) for the gender of the assessor, we decided to look at this issue separately, before the more general analysis of other assessor and researcher characteristics.

As described earlier, multilevel cross-classified models were fitted separately for social sciences and humanities (2401 ratings by 1580 assessors for 673 proposals from 28 fields of study) and for science (3235 ratings by 1621 assessors for 1436 proposals from 106 fields of study). (The number of cases was substantially lower than those in the main analyses because of the substantial amount of missing data associated with assessor gender.) Separate models were fitted with researcher and project ratings, using normalized ratings as the response variables. Female first researcher (1, female; 0, male), female assessor and their interaction were the explanatory variables. Importantly, the significance of this cross-level interaction between the gender of the researcher and the gender of the assessor could not be tested appropriately by using traditional single-level analyses.

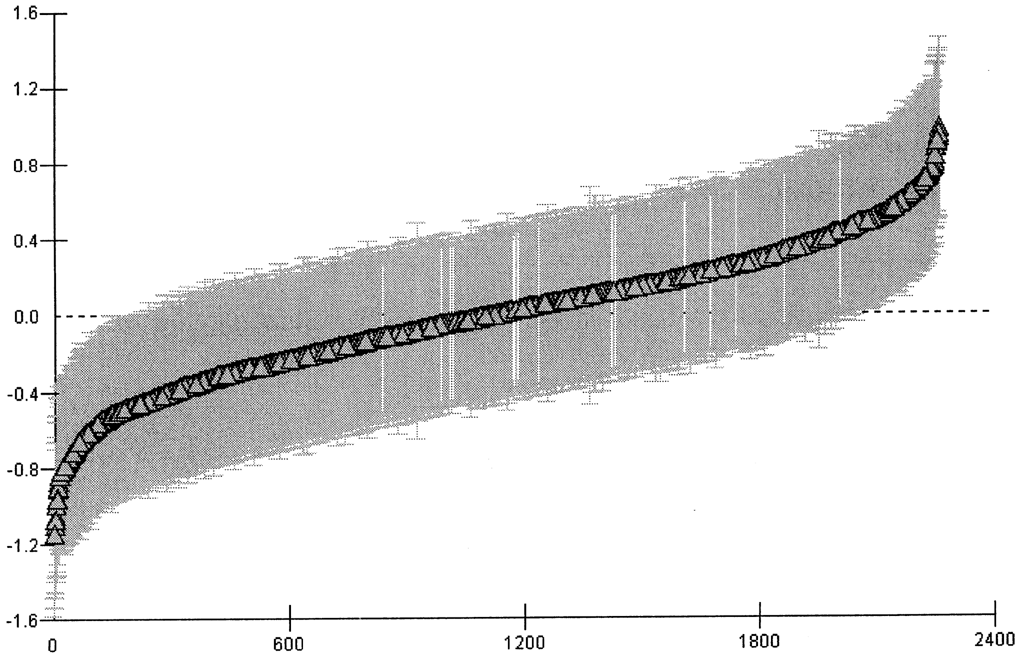

Fig. 1. Confidence intervals (plus or minus 1.96 standard errors) of proposal residuals for researcher ratings, illustrating the substantial error variation that is associated with each proposal (for further detail see Goldstein and Healy (1995), Goldstein (1995) and Snijders and Bosker (1999)): note that a very large number of proposals means that these confidence intervals merge together to form a continuous interval rather than discrete intervals

Initially, we fitted a base-line variance component model for project and researcher ratings. The final model contained researcher gender, assessor gender and their interaction in addition to the intercept. The addition of three gender variables to the base-line model decreased the deviance by 10.0 and 5.9 for project and by 12.6 and 8.1 for researcher ratings of social science and science respectively. The effects of assessor gender were small, but statistically significant and negative for both ratings in social sciences and humanities (Table 2). Female assessors gave significantly lower ratings than did male assessors in social sciences and humanities, whereas female researchers received lower ratings than male researchers in science. However, the researcher gender by assessor gender interaction (Table 2) was not statistically significant for all ratings in both disciplines. In supplemental analyses, the coefficient of each gender variable was allowed to vary randomly at each level, but this addition was not statistically significant, indicating that the relative lack of gender effects was consistent across different fields of study and different assessors.

In summary, there was little or no significant effect of gender and, of particular importance, there was no evidence to suggest that female assessors favoured female researchers or male assessors favoured male researchers. We interpreted these results to indicate that there is no evidence based on our results to suggest a gender bias in ARC ratings.

4.3. Main analyses of assessor and researcher characteristics

In this section we pursue a more complete analysis of assessor and researcher characteristics. These effects on project and researcher ratings were analysed in several stages. Firstly, a base-line

Table 2. Cross-classified analyses for genders of the researcher and assessor†

Parameter	*Estimate for the following disciplines:*			
	Social science and humanities		*Science*	
	Project ratings	*Researcher ratings*	*Project ratings*	*Researcher ratings*
Fixed effects: main effects				
Intercept	0.114	0.073	0.010	0.023
Female 1st researcher	0.099 (0.059)	0.062 (0.062)	−0.090 (0.053)	**−0.132 (0.056)**
Female assessor	**−0.142 (0.071)**	**−0.143 (0.069)**	−0.093 (0.058)	−0.060 (0.057)
Interaction				
Female 1st researcher × female assessor	−0.034 (0.103)	−0.090 (0.100)	0.033 (0.153)	−0.041 (0.151)
Random effects				
σ_u^2 (between-field variance)	0.025 (0.013)	0.022 (0.013)	**0.059 (0.013)**	**0.050 (0.012)**
σ_{ej}^2 (between-proposal variance)	**0.171 (0.026)**	**0.243 (0.028)**	**0.107 (0.013)**	**0.167 (0.015)**
σ_{ei}^2 (between-assessor variance)	**0.918 (0.031)**	**0.831 (0.028)**	**0.749 (0.017)**	**0.689 (0.016)**

†Standard errors are given in parentheses. Values in bold are statistically significant.

model of variance components was evaluated. Secondly, assessor attributes and their interactions (but not researcher attributes) were added to the base-line model. Thirdly, the final model contained all assessor and researcher attributes with their interactions. All possible two-way interactions were introduced and, because no three-way interactions were significant, they were not included in the models. For these analyses, a small number of proposals from organizations other than universities were excluded. After listwise deletion of missing values, main analyses for social sciences and humanities included 2738 ratings (95% of the total available) given by 1858 assessors for 667 proposals from 28 fields of study. Similarly, the main analyses for science included 6660 ratings (93% of the total available), 4055 assessors, 1580 proposals and 108 fields of study.

4.3.1. Variance components for the main analyses

This section examines the variances due to assessors, proposals and fields of study before and after adjusting for assessor and researcher characteristics.

4.3.1.1. Base-line variance component model. We began by fitting a base-line (variance component) model of project and researcher ratings. This model explained how the total variance was partitioned into variance components associated with assessor, proposal and field of study. The results indicated that in social sciences and humanities there was significant variation between fields of study (project, $\sigma_u^2 = 0.035$, standard error SE = 0.014; researcher, $\sigma_u^2 = 0.030$, SE = 0.015), between proposals (project, $\sigma_{ej}^2 = 0.162$, SE = 0.023; researcher, $\sigma_{ej}^2 = 0.245$, SE = 0.025) and between assessors (project, $\sigma_{ei}^2 = 0.906$, SE = 0.028; researcher, $\sigma_{ei}^2 = 0.803$, SE = 0.025). Similarly, the values for field of study, proposal and assessor variances in science were 0.056 (SE = 0.012), 0.099 (SE = 0.011) and 0.773 (SE = 0.015) for project ratings and 0.042 (SE = 0.011), 0.163 (SE = 0.013) and 0.713 (SE = 0.014) respectively for researcher ratings. Thus, employing multilevel cross-classified models is well justified. In social sciences and humanities, for project rating, 82.1% of the variance is due to assessors, 14.7% is due to the proposal and 3.2% is due to the field of study. For researcher ratings the corresponding percentages were 74.5%, 22.7% and 2.8% respectively. Similarly, in science percentages of assessor, proposal and field of study variances were 83.3%, 10.7% and 6.0% for project ratings and 77.7%, 17.8% and 4.6% for researcher ratings respectively. Whereas systematic variation due to the proposal was disappointingly low in both disciplines, it was systematically higher for researcher ratings than for project ratings, indicating why researcher ratings were more reliable. Although most of the variance in both ratings was due to differences between assessors of the same proposal, there was significant variation due to proposals and fields of study.

4.3.1.2. Variance components for model with assessor characteristics. The introduction of assessor variables into the base-line model led to small decreases in the variance components for field of study for both project ratings (from 0.035 to 0.029 in social sciences and humanities and from 0.056 to 0.042 in science) and researcher ratings (from 0.030 to 0.027 in social sciences and humanities and from 0.042 to 0.031 in science). This indicated that the differences between fields of study were reduced somewhat after controlling for the effects of assessor attributes. This reduction was marginally higher in science than in social sciences and humanities. Controlling for assessor attributes *increased* somewhat the variance components for proposals for both project ratings (from 0.162 to 0.172 in social sciences and humanities and from 0.099 to 0.112 in science) and researcher ratings (from 0.245 to 0.255 in social science and humanities and from 0.163 to 0.175 in science). These results are important and interesting in that controlling for the effects of explanatory variables typically leads to reductions in the variance components, not increases. In the present investigation, however, the reliability of differentiation between proposals was improved somewhat by controlling for assessor characteristics. This suggests, perhaps, that the assessors' characteristics evaluated here were not sources of valid differences between the proposals.

4.3.1.3. Variance components for model with assessor and researcher characteristics. The introduction of research variables into the model with assessor attributes led to small decreases in the variances for the field of study for project ratings (from 0.029 to 0.025 for social sciences and humanities and from 0.042 to 0.036 for science) and researcher ratings (from 0.027 to 0.021 for social sciences and humanities and from 0.031 to 0.025 for science). This indicates that the differences between fields of study were reduced somewhat after controlling for the effects of researcher attributes. Controlling for researcher attributes also decreased somewhat the variances for proposals for both project (from 0.172 to 0.149 for social sciences and humanities and from 0.112 to 0.099 for science) and researcher (from 0.255 to 0.209 for social sciences and humanities and from 0.175 to 0.144 for science) ratings. Hence, the reliability in differentiation between proposals was reduced somewhat by controlling for researcher characteristics, suggesting, perhaps, that the researcher characteristics that were evaluated here were valid sources of differentiation between the proposals.

4.3.2. Simultaneous effects of assessor and researcher characteristics

In this section, our main focus is on the results of the final model that simultaneously evaluated assessor and researcher characteristics (Table 3). However, results of the preliminary model with assessor characteristics only are also discussed if these results differed from those of the final model. The introduction of assessor variables (the nationality of the assessor, the researcher-nominated status, the number of proposals reviewed by each assessor and inter-

Table 3. Cross-classified analyses for characteristics of the researcher and assessor†

Parameter	*Estimate for the following disciplines:*			
	Social science and humanities		*Science*	
	Project ratings	*Researcher ratings*	*Project ratings*	*Researcher ratings*
Assessor main effects				
Intercept	−0.399	−0.494	−0.366	−0.357
Researcher-nominated assessors	**0.557 (0.131)**	**0.472 (0.123)**	**0.518 (0.083)**	**0.476 (0.080)**
American assessors	**0.880 (0.228)**	**0.571 (0.217)**	−0.014 (0.100)	−0.017 (0.098)
European assessors	**0.459 (0.217)**	0.157 (0.207)	−0.022 (0.090)	−0.154 (0.088)
Assessors from other regions	**0.621 (0.286)**	**0.580 (0.273)**	−0.094 (0.155)	0.022 (0.151)
Number of proposals reviewed	0.038 (0.119)	−0.039 (0.114)	**−0.051 (0.024)**	**−0.074 (0.023)**
Assessor interaction effects				
Nominated and American assessor	**0.299 (0.142)**	0.253 (0.135)	**0.218 (0.074)**	**0.206 (0.071)**
Nominated and European assessor	0.105 (0.137)	0.218 (0.131)	0.076 (0.077)	0.105 (0.075)
Nominated and assessor from other regions	0.111 (0.279)	0.209 (0.266)	0.184 (0.120)	0.149 (0.117)
Nominated and number of proposals reviewed	−0.011 (0.124)	−0.118 (0.118)	−0.063 (0.053)	−0.075 (0.051)
American and number of proposals reviewed	−0.095 (0.320)	−0.404 (0.305)	**−0.252 (0.096)**	**−0.264 (0.094)**
European and number of proposals reviewed	−0.216 (0.245)	−0.182 (0.234)	−0.050 (0.065)	−0.031 (0.063)
Other regions and number of proposals reviewed	−0.021 (0.364)	0.063 (0.347)	0.102 (0.104)	0.110 (0.101)
Researcher main effects				
Professor 1st researcher	0.153 (0.144)	0.178 (0.151)	**0.270 (0.089)**	**0.312 (0.093)**
University status 1st researcher	**0.266 (0.085)**	**0.292 (0.088)**	**0.262 (0.057)**	**0.246 (0.059)**
Cubic age 1st researcher	−0.061 (0.037)	**−0.081 (0.038)**	0.010 (0.014)	0.022 (0.014)
Researcher interaction effects				
Professor and university 1st researcher	0.208 (0.154)	**0.320 (0.162)**	−0.117 (0.092)	0.034 (0.097)
Professor and cubic age 1st researcher	0.012 (0.021)	−0.020 (0.022)	−0.012 (0.008)	0.000 (0.009)
University and cubic age 1st researcher	0.058 (0.038)	**0.102 (0.040)**	0.008 (0.014)	−0.011 (0.015)
Researcher × assessor interactions				
Nominated assessor and professor 1st researcher	0.096 (0.113)	0.120 (0.107)	−0.052 (0.066)	−0.058 (0.064)
Nominated assessor and university 1st researcher	−0.069 (0.132)	−0.080 (0.124)	−0.062 (0.082)	−0.080 (0.077)
Nominated assessor and cubic age 1st researcher	0.004 (0.019)	−0.031 (0.018)	0.003 (0.008)	−0.003 (0.008)
American assessor and professor 1st researcher	−0.131 (0.172)	−0.059 (0.165)	0.129 (0.075)	0.014 (0.073)
American assessor and university 1st researcher	**−0.713 (0.205)**	**−0.544 (0.197)**	−0.013 (0.099)	0.058 (0.096)
American assessor and cubic age 1st researcher	−0.002 (0.039)	−0.003 (0.037)	0.016 (0.009)	0.015 (0.009)
European assessor and professor 1st researcher	−0.045 (0.157)	−0.002 (0.151)	0.085 (0.073)	0.032 (0.071)
European assessor and university 1st researcher	**−0.511 (0.216)**	−0.329 (0.206)	−0.098 (0.091)	−0.029 (0.089)
European assessor and cubic age 1st researcher	−0.023 (0.025)	0.005 (0.024)	−0.002 (0.009)	−0.002 (0.008)
Other region assessor and professor 1st researcher	**−0.547 (0.265)**	**−0.795 (0.254)**	0.149 (0.121)	0.128 (0.117)
Other region assessor and university 1st researcher	**−0.638 (0.300)**	**−0.606 (0.286)**	0.166 (0.156)	−0.083 (0.152)
Other region assessor and cubic age 1st researcher	0.120 (0.062)	**0.144 (0.059)**	−0.022 (0.013)	−0.016 (0.013)
Proposals reviewed and professor 1st researcher	0.078 (0.108)	0.099 (0.104)	0.011 (0.026)	0.008 (0.025)
Proposals reviewed and university 1st researcher	−0.097 (0.124)	−0.066 (0.119)	0.004 (0.025)	−0.002 (0.024)
Proposals reviewed and cubic age 1st researcher	−0.027 (0.017)	−0.022 (0.016)	−0.000 (0.005)	−0.005 (0.005)

(*continued*)

Table 3 (*continued*)

Parameter	*Estimate for the following disciplines:*			
	Social science and humanities		*Science*	
	Project ratings	*Researcher ratings*	*Project ratings*	*Researcher ratings*
Random effect				
σ_u^2 (between-field variance)	**0.025 (0.012)**	0.021 (0.011)	**0.036 (0.009)**	**0.025 (0.007)**
σ_{ej}^2 (between-proposal variance)	**0.149 (0.020)**	**0.209 (0.022)**	**0.099 (0.010)**	**0.144 (0.011)**
σ_{ei}^2 (between-assessor variance)	**0.794 (0.025)**	**0.701 (0.022)**	**0.692 (0.014)**	**0.635 (0.013)**

†Standard errors are given in parentheses. Values in bold are statistically significant.

actions between these variables) led to highly significant decreases in the deviances for both project ratings (268.7 for social sciences and humanities and 587.0 for science, 12 degrees of freedom and $p < 0.001$) and researcher ratings (270.9 for social sciences and humanities and 626.4 for science, $p < 0.001$). Similarly, the addition of researcher variables (academic title, university status and cubic age of the first researcher) and their interactions with assessor and researcher variables was significant for both project ($\chi^2(21) = 82$ for social sciences and humanities and $\chi^2(21) = 107$ for science) and researcher ($\chi^2(21) = 112$ for social science and $\chi^2(21) = 172.7$ for science) ratings (p-values less than 0.001). Finally, in supplemental analyses, we ascertained that the exclusion of non-significant effects had almost no effect on the parameter estimates.

The largest effect was for researcher-nominated assessors ($p < 0.001$) even after controlling for researcher characteristics. Overall, researcher-nominated assessors gave higher ratings than panel-nominated assessors. However, this difference interacted significantly with nationality (North American *versus* Australian) for all except researcher ratings in the social sciences and humanities (even this interaction was significant in the assessor model). An inspection of these interactions (Fig. 2) indicated that the difference between researcher-nominated and panel-nominated assessors was somewhat larger for ratings by North American assessors than Australian assessors.

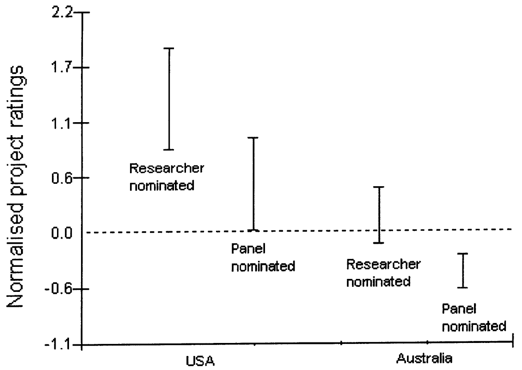

Fig. 2. 95% confidence interval for the interaction between North American nationality of assessors and researcher-nominated status of assessors for project ratings in social sciences and humanities

In science, the number of proposals reviewed by an assessor was significantly and negatively related to their ratings, but this effect also interacted with nationality (North American *versus* Australian—see Table 3). Whereas assessors who reviewed more proposals gave lower ratings, this negative effect of the number of proposals assessed was larger for North American assessors than for Australian assessors (Fig. 3). In social sciences and humanities, there was no significant effect of the number of proposals reviewed by an assessor.

There were significant effects of the nationality of the assessors that are of particular interest, but the interpretations were complicated by interactions particularly with the university status (see Table 3). In the social sciences and humanities, Australian assessors tended to give lower ratings than did overseas assessors. However, Fig. 4 illustrates that the effect of assessors' nationality was larger for researchers from post-1987 (newer, less prestigious) Australian universities. Hence, the relative status of the Australian university from which the proposal came had significantly more effect on Australian assessors than on overseas assessors. Interestingly, in science we did not find any significant effect of region of assessor on ratings (although European assessors gave significantly lower ratings before adjusting for researcher characteristics).

University status had a highly significant effect for all ratings in both disciplines (Table 3) such that proposals from pre-1987 (older, more prestigious) universities received higher ratings than those from post-1987 universities. However, as discussed earlier, this effect interacted with the nationality of the assessor in the social sciences and humanities such that the systematically higher ratings for proposals from pre-1987 universities were due primarily to ratings by Australian assessors (Fig. 4).

The positive effect of being a professor was evident in science for both project and researcher ratings. In the social sciences and humanities, the positive effect of being a professor was not significant but did interact significantly with university status for researcher ratings (Table 3). An inspection of Fig. 5 indicates that the advantage of being a professor was more clearly evident in older, well-established (pre-1987) universities than in newer universities. The effect of professor status in the social sciences and humanities also interacted with the effect of assessors from other regions such that these assessors rated proposals by professors marginally lower than proposals by non-professors, whereas Australian assessors rated proposals by professors as marginally higher (Fig. 6).

The effects of age, after controlling for other variables, was not significant for ratings in the sciences but was significant for researcher ratings in the social sciences and humanities. The

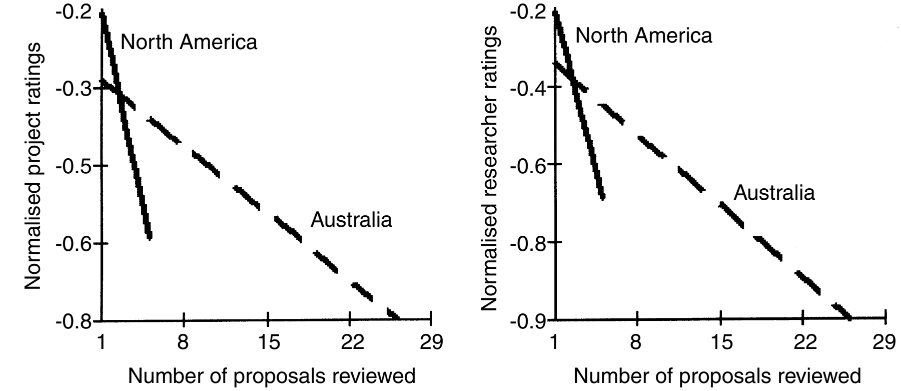

Fig. 3. Interaction between North American nationality of assessors and the number of proposals reviewed for project and researcher ratings in science

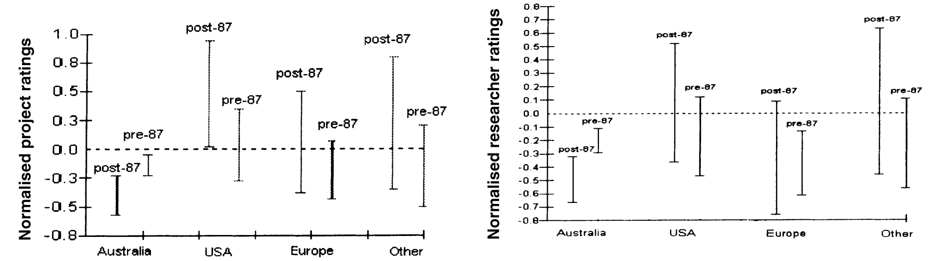

Fig. 4. 95% confidence interval for interaction between assessors' nationality and the status of the university for project and researcher ratings in social sciences and humanities

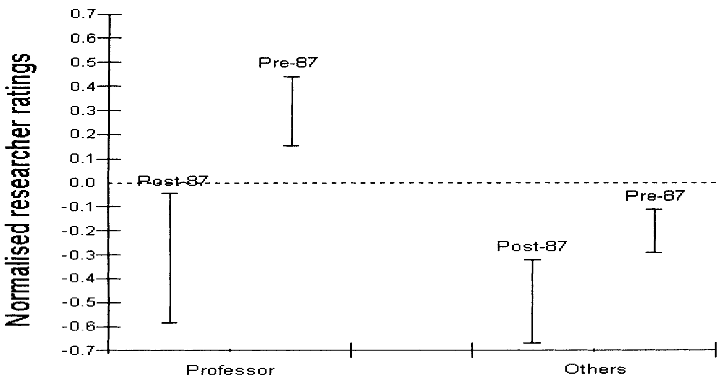

Fig. 5. 95% confidence interval for the interaction between professor rank and university status in the social sciences: researcher ratings received by professors in pre-1987 and post-1987 universities

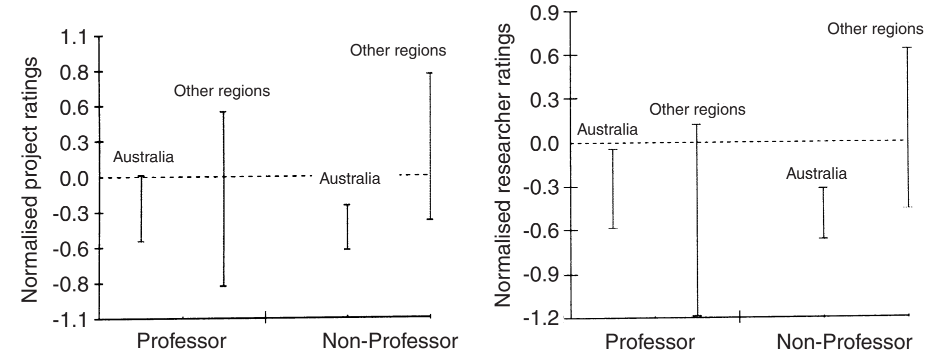

Fig. 6. 95% confidence interval for researcher and project ratings given to researchers with different academic titles in social science and humanities

cubic age effect of first researcher was significantly negative for researcher ratings in the social sciences, but this effect interacted with the effect of university status and with the effect of assessors from other regions. Fig. 7 illustrates that researcher ratings for younger researchers were similar for pre-1987 and post-1987 universities, but that older researchers from pre-1987 universities received better ratings than older researchers from post-1987 universities. Researchers from pre-1987 universities received somewhat higher ratings with increasing age, whereas

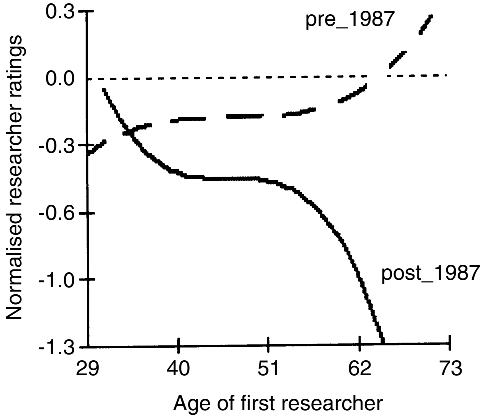

Fig. 7. Interaction between age of researcher and university status for researcher ratings in social sciences and humanities

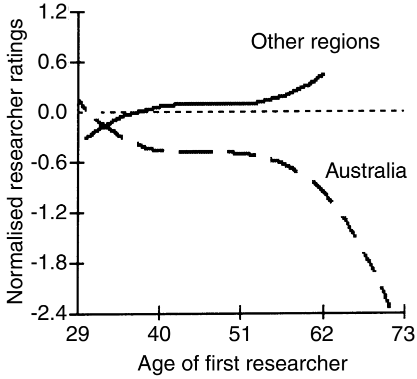

Fig. 8. Interaction between age of researcher and nationality (Australia *versus* other regions) for researcher ratings in social sciences and humanities

researchers from post-1987 universities received lower ratings with increasing age (see Fig. 7). This difference apparently reflects the fact that many older researchers at post-1987 universities were initially appointed at a time when these institutions were colleges of advanced education rather than research universities. In the social sciences the age of the researcher also interacted with the effect of assessors from other regions for researcher ratings. Whereas assessors from Australia and other regions tended to give similar ratings to the youngest researchers, Australian assessors gave systematically lower ratings as researchers became older (Fig. 8).

5. Discussion

5.1. Reliabilities

Estimates of reliability for the evaluations of grant proposals in the ARC scheme were disappointingly low. The single-rater reliabilities were low for researcher ratings (0.255 in social sciences and humanities; 0.223 in science) and even lower for project ratings (0.179 in social sciences and humanities; 0.167 in science). On the basis of the average number of assessors per proposal (4.1 for social sciences and humanities; 4.2 for science), the estimates of reliability for the average proposal assessment were 0.47 and 0.46 for project ratings and 0.58 and 0.55 for researcher ratings respectively. Whereas these estimates are unacceptably low, the estimates of reliability were equally low in the sciences and in the social sciences and humanities. Hence, there was no support for the contention that evaluations of grant proposals are more reliable in the sciences than in social sciences and humanities.

Although disappointingly low, these estimates of reliability underestimated the reliability of the actual panel decisions. In the ARC programme the final decision is taken by panels, who also review the proposal, assessors' comments, authors' rejoinders, etc. Furthermore, the 22% of initially culled proposals that were not reviewed by external assessors were not included in the analyses of reliability (because there were no external assessments of these proposals). If they had been included, the between-proposal variance would probably have increased substantially and, consequently, the reliability would have improved. Because of the above-mentioned factors, the final decision about ARC proposals was likely to have been more reliable than indicated by the estimates of reliability based on the assessor ratings (for a further discussion, see Marsh and Bazeley (1999)). However, the reliability of the dichotomous decision (e.g. to fund or not to fund) was likely to be lower than the reliability of the reasonably continuous scores on which this decision was based.

5.2. Researcher and assessor attributes

The differences in success rates for male and female researchers were small or non-significant. In this sense, there was no gender bias in the grant evaluations. Furthermore, this lack of researcher gender difference did not interact with assessor gender. In particular, female assessors did not favour female researchers; nor did male assessors favour male researchers. Hence, there was no support for a 'matching hypothesis'. Surprisingly, however, only 15.3% of the total number of researchers (including initial culls) were females. Thus, females were substantially under-represented among those researchers who applied for ARC grants even though those who did apply fared no better or worse than the predominantly male applicants.

When the first-named researcher on the research team was a professor, both the project ratings and particularly the researcher ratings were higher in science. In supplementary analyses, we found that the inclusion of a professor in the research team as a co-author (not the first-named researcher) also had an additional positive effect on ratings. Common sense suggests that academic rank may be a valid source of variance and this interpretation is supported by two related findings. First, the effect of professor status was significantly stronger for researcher ratings than for project ratings (see Table 3). Individuals become professors at least in part because of their well-established research track records, so ratings of research track records should be higher for professors than for non-professors. Whereas it is possible that professors should also be able to write better grant proposals than non-professors, it is also possible for non-professors to write outstanding proposals as well. Hence, the finding that the advantage associated with being a professor was greater for researcher ratings than project ratings offered support for the interpretation that this is a valid source of variation in the ratings. Second, controlling for the effect of being a professor resulted in a less reliable differentiation between the proposals. Again, this suggested that professor rank might be a valid source of variation in the proposal evaluations. Indeed, it is surprising, perhaps, that the effect of being a professor was not even greater. The apparent explanation for this finding is an implicit reverse discrimination introduced by the ARC in that applicants were only asked to provide a list of publications during the previous 5 years and assessors were explicitly instructed to rate the research track record in relation to opportunity. This may also explain why Australian assessors rate particularly older researchers more negatively in that these assessors are more likely to be aware of these policies, which are designed to provide a greater opportunity to younger, less senior researchers.

The status of the university had a significant and substantial effect in both social sciences and humanities, and in sciences. In social sciences and humanities, however, this effect was only evident in ratings by Australian assessors (Fig. 4).

6. Policy implications and future directions

6.1. Researcher-nominated assessors

Researcher-nominated assessors gave systematically higher ratings than did ARC panel-nominated assessors. Supplemental analyses showed that the reliabilities of ratings increased after adjusting for the effects of nominated assessors and that the reliabilities of ratings by researcher-nominated assessors were systematically lower than the reliabilities of panel-nominated assessors. These results suggest that ratings by researcher-nominated assessors were positively biased.

The rationale for having researcher-nominated assessors was that researchers would be more likely to have at least one assessor who had a good understanding of the proposal and would be likely to make the peer review process more credible to the academic community. Although the ARC did not actually present a systematic evaluation of researcher-nominated ratings, there was at least the implication that they were not systematically higher than ratings by ARC-nominated assessors and that they were sometimes substantially lower. Hence, our results present a very different picture—showing that researcher-nominated assessors gave systematically higher ratings than ARC panel-nominated assessors. Clearly, a proposal that did not have any researcher-nominated assessors was likely to be disadvantaged relative to proposals that did, and the few proposals that had multiple researcher-nominated assessors were likely to be particularly advantaged. Also, because the ARC could not control communication between researchers and their researcher-nominated assessors, it was possible that researchers put inappropriate pressure on their researcher-nominated assessors. Importantly, our results also suggest that ratings for researcher-nominated assessors were more positively biased and less reliable than those by ARC panel-nominated assessors. Hence, despite the potentially adverse reactions by researchers who appreciated being able to nominate their own assessors, the ARC decided to drop the use of researcher-nominated assessors. In this respect, our results have already had important implications on policy for the ARC peer review process (Bond and Hesketh, 1998). These results also have important implications on policy for many other situations in which proposers or applicants can choose some or all of the external assessors who provide assessments of their suitability as part of an assessment process (e.g. most job applications).

6.2. Number of proposals reviewed by each assessor

It was observed that the number of proposals reviewed by each assessor had a significant negative effect on global assessor ratings in science. This effect was little affected by controlling for other assessor characteristics. Apparently, assessors who reviewed more proposals had a broader frame of reference against which to evaluate each proposal. The more proposals an assessor reviewed, the better the quality of the ratings they provided. As the number of proposals increased, the ability to discriminate between the proposals also increased and, thus, the reliabilities of proposals. This interpretation was supported by supplemental analyses of reliabilities. In particular, the single-rater reliabilities were computed separately for global ratings provided by assessors who rated fewer than three proposals and three or more proposals (overall 34% or 3441 of ratings were provided by assessors who rated three or more proposals). The single-rater reliabilities of assessors who rated three or more proposals were systematically higher for both project ($\rho = 0.159$) and researcher ($\rho = 0.267$) ratings than those of assessors who rated fewer than three proposals ($\rho = 0.147$ for project ratings and $\rho = 0.192$ for researcher ratings). Interestingly, this effect of the number of proposals reviewed on the assessor ratings was even larger for North American assessors than for Australian assessors in science.

Apparently, assessors who reviewed a single proposal tended to give higher ratings in part because they did not know the quality of the other proposals in the same subdisciplinary area and had no adequate frame of reference for determining what constituted a fundable proposal. Thus, for example, Perlman (1982) suggested that one method of improving interrater reliabilities would be for reviewers to review sets of papers simultaneously rather than reviewing individual papers periodically over a long period of time. Unfortunately, in our study 66% of assessors reviewed only one proposal and 20% two proposals, which partially explains why the reliability of the ratings was so low.

The results of the present investigation suggest that there may be relatively little systematic bias in peer ratings for the large ARC scheme—at least for the characteristics that were considered here. Like nearly all studies of peer reviews, however, our results indicated that there was substantial variation between ratings by different assessors. Hence, the reliability of the peer reviews is not adequate by most standards—particularly given the importance of the decisions based on these peer reviews and the high regard given to the peer review process in evaluating a wide range of academic outcomes. A critical direction for future research is how to improve the reliability of peer reviews. One approach to this problem would be to assess all reviewers in terms of their leniency of grading and to control for this effect when evaluating their reviews (i.e. to discount the low ratings by extremely harsh reviewers and the high ratings by extremely lenient reviewers). This approach, however, would only be practical for reviewers who rated a large number of proposals (e.g. Marsh and Ball (1981)) and is not feasible in a scheme where most reviewers assess only one or a small number of proposals. Alternatively, it may be feasible for a relatively small number of assessors to evaluate a relatively large number of proposals within their subdiscipline area of expertise. This would substantially reduce between-assessor variation in that the same assessors would evaluate all proposals within the same subdiscipline or field of study. Although a radical departure from the existing ARC scheme, this strategy is consistent with the manner in which the ARC does preliminary culls (based on evaluations by two panel members with relevant expertise) and is consistent with peer review processes used by some academic journals where many or most of the assessments are conducted by members of a relatively small editorial board. In support of this strategy, we found that the quality of ratings apparently improved when the same assessor was asked to evaluate more proposals.

Acknowledgements

This research was funded in part by a grant from the ARC. We would like to thank the ARC and Professor Max Brennan, former Chair of the ARC, for their assistance in providing the data that are used in this study, and Ken Rowe for comments on the analyses and earlier drafts of this paper. We are also grateful to the Joint Editor and two referees for helpful comments and suggestions.

References

Allison, P. D. and Stewart, J. A. (1974) Productivity differences among scientists: evidence for accumulative advantage. *Am. Sociol. Rev.*, **39**, 596–606.

Australian Research Council (1996) Large research grants guidelines for 1996. In *ARC Members' Handbook 1996*. Canberra: Australian Research Council.

Bazeley, P. (1998) Peer review and Panel decisions in the assessment of Australian Research Council Project grant applicants: what counts in a highly competitive context? *Higher Educ.*, **35**, 435–452.

Bazeley, P., Kemp, L., Stevens, K., Asmar, C., Grbich, C., Marsh, H. W. and Bhatal, R. (1996) *Waiting in the Wings: a Study of Early Career Academic Researchers in Australia*. Canberra: National Board of Employment, Education and Training.

Bond, N. and Hesketh, B. (1998) A report on the "Reader Trial" conducted by Panel A7a-Psychology and Education. University of Western Sydney, Sydney.

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

Cole, S., Rubin, L. and Cole, J. R. (1977) Peer review and the support of science. *Scient. Am.*, **237**, 34–41.

Daniel, H. D. (1993) *Guardians of Science: Fairness and Reliability of Peer Review* (transl. by E. R. William). Weinnheim: VCH Weinnheim.

Emery, J. A., Meyers, H. W. and Hunter, D. E. (1992) NIH FIRST awards: testing background factors for funding against peer review. *J. Soc. Res. Admin.*, **24**, part 2, 7–15.

Garfield, E. (1986) Refereeing and peer review: Part 1, opinion and conjecture on the effectiveness of refereeing. *Curr. Contents*, **31**, 3–11.

Gilbert, J. R., Williams, E. S. and Lundberg, G. D. (1994) Is there gender bias in JAMA's peer review process? *J. Am. Med. Ass.*, **272**, 139–142.

Goldstein, H. (1995) *Multilevel Statistical Model*, 2nd edn. London: Arnold.

Goldstein, H. and Healy, M. J. R. (1995) The graphical presentation of a collection of means. *J. R. Statist. Soc.* A, **158**, 175–177.

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse, G. and Healy, M. (1998) *A User's Guide to MLwiN*. London: Institute of Education.

Hammel, E. (1980) Report of the Task Force on Faculty Renewal. Population Research Center, University of California at Berkeley, Berkeley.

Lindsey, D. (1976) Distinction achievement and editorial board membership. *Am. Psychol.*, **31**, 799–804.

Lindsey, D. (1977) Participation and influence in publication review proceedings: a reply. *Am. Psychol.*, **32**, 579–585.

Lindsey, D. (1978) *The Scientific Publication System in Social Science*. San Francisco: Jossey-Bass.

Mahoney, M. J. (1985) Open exchange and epistemic progress. *Am. Psychol.*, **40**, 29–39.

Marsh, H. W. and Ball, S. (1981) The interjudgmental reliability of reviews for the *Journal of Educational Psychology*. *J. Educ. Psychol.*, **73**, 872–880.

Marsh, H. W. and Ball, S. (1989) The peer review process used to evaluate manuscripts submitted to academic journals: interjudgemental reliability. *J. Exptl Educ.*, **57**, 151–169.

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

Marsh, H. W. and Bazeley, P. (1999) Multiple evaluations of grant proposals by independent assessors: confirmatory factor analysis, evaluations of reliability, validity and structure. *Multiv. Behav. Res.*, **34**, 1–30.

Naylor, C. D. (1994) Reviewing applied research grant proposals: can we learn from medical journals? *Can. Med. Ass. J.*, **150**, 1207–1209.

Neidhardt, F. (1988) *Selbststeuerung in der Forschungsförderung: das Gutachterwesen der DFG Opladen.* Westdeutscher Verlag.

Osmond, D. H. (1983) Malice's Wonderland: research funding and peer review. *J. Neurbiol.*, **14**, 95–112.

Perlman, D. (1982) Reviewer "bias": do Peters and Ceci protest too much? *Behav. Brain Sci.*, **5**, 231–232.

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D., Langford, I. and Lewis, T. (1999) *A User's Guide to MLwiN—Version 2.0*. London: Institute of Education.

Rasbash, J. and Goldstein, H. (1994) Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model. *J. Educ. Behav. Statist.*, **19**, 337–350.

Shout, P. E. and Fleiss, J. L. (1979) Intraclass correlations: uses in assessing rater reliability. *Psychol. Bull.*, **86**, 420–428.

Snijders, T. A. B. and Bosker, R. J. (1999) *Multilevel Analysis: an Introduction to Basic and Advanced Multilevel Modeling*. London: Sage.

Winer, B. J. (1971) *Statistical Principles in Experimental Design*. New York: McGraw-Hill.

Wood, F. Q. (1997) *The Peer Review Process*. Canberra: National Board of Employment, Education and Training.

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