Nepotism and sexism in peer-review

In the first-ever analysis of peer-review scores for postdoctoral fellowship applications, the system is revealed as being riddled with prejudice. The policy of secrecy in evaluation must be abandoned.

Christine Wennerås and Agnes Wold

Throughout the world, women leave their academic careers to a far greater extent than their male colleagues[1]. In Sweden, for example, women are awarded 44 per cent of biomedical PhDs but hold a mere 25 per cent of the postdoctoral positions and only 7 per cent of professorial positions. It used to be thought that once there were enough entry-level female scientists, the male domination of the upper echelons of academic research would automatically diminish. But this has not happened in the biomedical field, where disproportionate numbers of men still hold higher academic positions, despite the significant numbers of women who have entered this research field since the 1970s.

Reasons for lack of success

Why do women face these difficulties? One view is that women tend to be less motivated and career-oriented than men, and therefore are not as assiduous in applying for positions and grants. Another is that women are less productive than men, and consequently their work has less scientific merit. Yet another is that women suffer discrimination due to gender. We decided to investigate whether the peer-review system of the Swedish Medical Research Council (MRC), one of the main funding agencies for biomedical research in Sweden, evaluates women and men on an equal basis. Our investigation was prompted by the fact that the success rate of female scientists applying for postdoctoral fellowships at the MRC during the 1990s has been less than half that of male applicants.

Our study strongly suggests that peer reviewers cannot judge scientific merit independent of gender. The peer reviewers overestimated male achievements and/or underestimated female performance, as shown by multiple-regression analyses of the relation between defined parameters of scientific productivity and competence scores.

...the credibility of the academic system will be undermined in the eyes of the public if it does not allow a scientific evaluation of its own scientific evaluation system.

In the peer-review system of the Swedish MRC, each applicant submits a curriculum vitae, a bibliography and a research proposal. The application is reviewed by one of 11 evaluation committees, each covering a specified research field. The individual applicant is rated by the five reviewers of the committee to which he or she has been assigned. Each reviewer gives the applicant a score between 0 and 4 for the following three parameters: scientific competence; relevance of the research proposal; and the quality of the proposed methodology. The three scores given by each reviewer are then multiplied with one another to yield a product score that can vary between 0 and 64. Finally, the average of the five product scores an applicant has received is computed, yielding a final score that is the basis on which the applicants to each committee are ranked.

The MRC board, which includes the chairmen of the 11 committees, ultimately decides to whom the fellowships will be awarded. Usually each committee chooses between one and three of the top-ranked applicants. Of the 114 applicants for the 20 postdoctoral fellowships offered in 1995, there were 62 men and 52 women, with a mean age of 36 years, all of whom had received a PhD degree within the past five years. Most of the female applicants had basic degrees in science (62 per cent), and the rest had medical (27 per cent) or nursing (12 per cent) degrees; the corresponding figures for the male applicants were 38, 59 and 3 per cent.

Traditionally, peer-review scores are not made public, and indeed the MRC officials initially refused us access to the documents dealing with evaluation of the applicants. In Sweden, however, the Freedom of the Press Act grants individuals access to all documents held by state or municipal authorities. Only documents defined as secret by the Secrecy Act are exempt, for example those that may endanger the security of the state, foreign relations or citizens' personal integrity. Accordingly, we appealed against the refusal of the MRC to release the scores.

In 1995, the Administrative Court of Appeal judged the evaluation scores of the MRC to be official documents. Hence, to our knowledge, this is the first time that genuine peer-reviewer evaluation sheets concerning a large cohort of applicants has become available for scientific study.

We found that the MRC reviewers gave female applicants lower average scores than male applicants on all three evaluation parameters: 0.25 fewer points for scientific competence (2.21 versus 2.46 points); 0.17 fewer points for quality of the proposed methodology (2.37 versus 2.54); and 0.13 fewer points for relevance of the research proposal (2.49 versus 2.62). Because these scores are multiplied with each other, female applicants received substantially lower final scores compared with male applicants (13.8 versus 17.0 points on average). That year, four women and 16 men were awarded postdoctoral fellowships.

As shown by these figures, the peer reviewers deemed women applicants to be particularly deficient in scientific competence. As it is generally regarded that this parameter is related to the number and quality of scientific publications[2–5], it seemed reasonable to assume that women earned lower scores on this parameter than men because they were less productive. We explored this hypothesis by determining the scientific productivity of all 114 applicants and then comparing the peer-reviewer ratings of groups of male and female applicants with similar scientific productivity.

Productivity variables

We measured the scientific productivity of each applicant in six different ways. First, we determined the applicant's total number of original scientific publications, and second, the number of publications on which the applicant was first author. Both figures were taken from the applicant's bibliography, which we double-checked in the Medline database. (We call these measures 'total number of publications' and 'total number of first-author publications'.)

To take into account the fact that the prestige of biomedical journals varies widely, we constructed measures based on journals' impact factors. The impact factor of a scientific journal is listed in the independent Institute of Scientific Information's *Journal Citation Reports*, and describes the number of times an average paper published in a particular journal is cited during one year. Our third measure was to add together the impact factors of each of the journals in which the applicant's papers were published, generating the 'total impact measure' of the applicant's total number of publications.

Fourth, we generated the 'first-author impact measure' by adding together the impact factors of the journals in which the applicant's first-author papers appeared. The unit of measure for both total impact and first-author impact is 'impact points',

with one impact point equalling one paper published in a journal with an impact factor of 1.

Fifth, using the science citation database, we identified the number of times the applicant's scientific papers were cited during 1994, which yielded the measure 'total citations'. And sixth, we repeated this procedure for papers on which the applicant was first author, giving the measure 'first-author citations'.

Did men and women with equal scientific productivity receive the same competence rating by the MRC reviewers? No! As shown in Fig. 1 for the productivity variable 'total impact', the peer reviewers gave female applicants lower scores than male applicants who displayed the same level of scientific productivity. In fact, the most productive group of female applicants, containing those with 100 total impact points or more, was the only group of women judged to be as competent as men, although only as competent as the least productive group of male applicants (the one whose members had fewer than 20 total impact points).

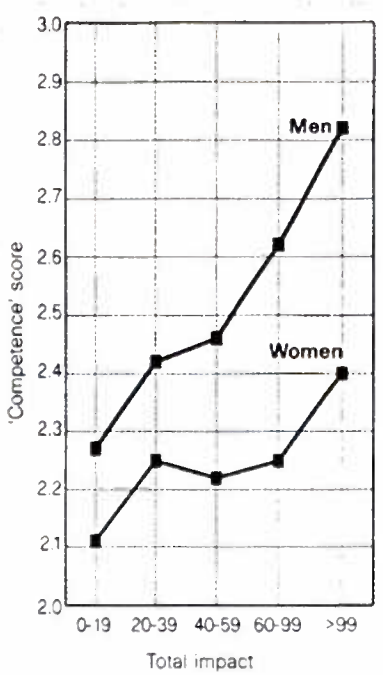

Figure 1 **The mean competence score given to male (red squares) and female (blue squares) applicants by the MRC reviewers as a function of their scientific productivity, measured as total impact. One impact point equals one paper published in a journal with an impact factor of 1. (See text for further explanation.)**

Why women score low

Although the difference in scoring of male and female applicants of equal scientific productivity suggested that there was indeed discrimination against women researchers, factors other than the applicant's gender could, in principle, have been responsible for the low scores awarded to women. If, for example, women were mainly to conduct research in areas given low priority by the MRC, come from less-renowned universities, or have less collaboration with academic decision-makers, their lower scores could depend on such factors, rather than on their gender *per se*.

To determine the cause of women's lower scores, we performed a multiple-regression analysis, which reveals the factors that exert a primary influence on a certain outcome (for example competence scores) and the size of such an influence. Multiple regression permits the elimination of factors whose influence on a certain outcome merely reflects their dependence on other factors.

In the multiple-regression analysis, we assumed that the competence scores given to applicants are linearly related to their scientific productivity. We constructed six different multiple-regression models, one for each of the productivity variables outlined above. In each of these models, we determined the influence of the following factors on the competence scores: the applicant's gender; nationality (Swedish/non-Swedish); basic education (medical, science or nursing school); scientific field; university affiliation; the evaluation committee to which the applicant was assigned; whether the applicant had had postdoctoral experience abroad; whether a letter of recommendation accompanied the application; and whether the applicant was affiliated with any of the members of the evaluation committee. The last piece of information is noted on the MRC evaluation protocols, in which case the reviewer in question is not allowed to participate in the scoring of that applicant. It was as frequent for female (12 per cent) as for male (13 per cent) applicants to be associated with a committee member.

The outcome of the regression analysis is shown in Table 1. Three out of the six productivity variables generated statistically significant models capable of predicting the competence scores the applicants were awarded: total impact, first-author impact and first-author citations. The model that provided the highest explanatory power was the one based on total impact ($r^2 = 0.47$). In all three models, we found two factors as well as scientific productivity that had a significant influence on competence scores: the gender of the applicant and the affiliation of the applicant with a committee member.

According to the multiple-regression model based on total impact, female applicants started from a basic competence level of 2.09 competence points (the intercept of the multiple regression curve) and were given an extra 0.0033 competence points by the reviewers for every impact point they had accumulated. Independent of scientific productivity, however, male applicants received an extra 0.21 points for competence. So, for a female scientist to be awarded the same competence score as a male colleague, she needed to exceed his scientific productivity by 64 impact points (95 per cent confidence interval: 35–93 impact points).

This represents approximately three extra papers in *Nature* or *Science* (impact factors 25 and 22, respectively), or 20 extra papers in a journal with an impact factor of around 3, which would be an excellent specialist journal such as *Atherosclerosis, Gut, Infection and Immunity, Neuroscience* or *Radiology*. Considering that the mean total impact of this cohort of applicants was 40 points, a female applicant had to be 2.5 times more productive than the average male applicant to receive the same competence score as he ((40 + 64)/40 = 2.6).

Friendship bonus

According to the same multiple-regression model, applicants who were affiliated with a committee member received competence scores 0.22 points higher than applicants of the same gender and scientific productivity who lacked such ties (Table 1). This 'affiliation bonus' was worth 67 impact points (confidence interval: 29 to 105 impact points). Hence, an applicant lacking personal ties with the reviewers needed to have 67 more impact points than an applicant of the same sex who was associated with one of the reviewers, to be perceived as equally competent. So, although MRC policy does not allow 'biased' reviewers to participate in the scoring of applicants they are associated with, this rule was insufficient, as the 'neutral' committee members compensated by raising their scores when judging applicants affiliated with one of their peers.

Because the affiliation bonus was of the same magnitude as the 'male gender' bonus, a woman applicant could make up for her gender (−0.21 competence points) by being affiliated with one of the reviewers (+0.22 competence points). On the other hand, a female (−0.21 competence points) lacking personal connections in the committee (−0.22 competence points) had to present an additional 131 impact points to the MRC reviewers to receive the same competence score as a male applicant affiliated with one of the reviewers.

Such a level of productivity was attained by only three of the 114 applicants, one male and two female. Hence, being of the female gender and lacking personal connections was a double handicap of such severity that it could hardly be compensated for by scientific productivity alone.

The two other regression models, based on first-author impact and first-author citations, yielded almost identical results to the first with regard to the effect of gender and affiliation (Table 1). This congruity was not a statistical artefact due to a high degree of interrelation between the three productivity variables, as the total impact and first-author

Table 1 **Factors that significantly influenced peer reviewers' rating of scientific competence, according to three multiple regression models.**

	Scientific productivity			*Additional points given by the reviewers for the following factors*			*Size of the influence of the non-scientific factors in productivity equivalents*		
Multiple regression model based on:	r^2	Intercept	Competence points per productivity unit	Male gender	Reviewer affiliation	Recommendation letter	Male gender	Reviewer affiliation	Unit of measure
Total impact	0.47	2.09	0.0033 *<0.00005**	0.21 *<0.00005*	0.22 *0.0008*	0.10 *0.04*	64 (35–93)†	67 (29–105)	Impact points
First-author impact	0.44	2.13	0.0094 *<0.0001*	0.24 *<0.00005*	0.20 *0.005*	NS	25 (14–36)	21 (6–36)	Impact points
First-author citations	0.41	2.17	0.0054 *0.001*	0.23 *<0.00005*	0.23 *0.001*	NS	42 (23–61)	42 (17–67)	Citations during 1994

* Italicized numbers indicate *P*-values for the variable in question.
† Numbers in parentheses indicate 95% confidence interval.
NS, not statistically significant, *P*-value > 0.05.

impact of the applicants were only moderately correlated ($r = 0.63$), as were total impact and first-author citations ($r = 0.62$). We therefore believe that male gender and reviewer affiliation were real determinants of scientific competence in the eyes of the MRC reviewers.

The applicant's nationality, education, field of research or postdoctoral experience did not influence competence scores in any of the models. A letter of recommendation had a positive effect on the competence score in the model based on total impact, but not in the two others (Table 1). By contrast, the evaluation committee that rated individual applicants did influence competence scores, as some committees were 'sterner' in their evaluation of competence than the rest (data not shown). However, an applicant who was assigned to a 'tough' committee had the same chance of being awarded a fellowship as other applicants, as fellowships were distributed based on the rank the applicant acquired within his or her committee and not on absolute score values.

Changing the system

The peer-review system, characterized as "the centrepiece of the modern scientific review process"[6], has been criticized on many grounds, including poor inter-reviewer reliability[7] and because reviewers may favour projects confirming their own views[8]. Our study is the first analysis based on actual peer-reviewer scores and provides direct evidence that the peer-review system is subject to nepotism, as has already been suggested anecdotally[8–10].

One might argue that young researchers affiliated with peer reviewers are part of a scientific élite that has received superior training and are therefore more competent than average applicants. Indeed, applicants with such ties had higher total impact levels on average than applicants without such connections (data not shown). Hence, applicants with personal alliances justly benefited from higher competence scores because of their higher scientific productivity. However, on top of that, they were given extra competence points not warranted by scientific productivity. We see no reason why an applicant who manages to produce research of high quality despite not being affiliated with a prestigious research group should not be similarly rewarded.

Several studies have shown that both women and men rate the quality of men's work higher than that of women when they are aware of the sex of the person to be evaluated, but not when the same person's gender is unknown[11–13]. It is somewhat surprising that the results of these studies have not discouraged the scientific community from relying on evaluation systems that are vulnerable to reviewer prejudice.

An interesting question that we could not address here is whether the harsher evaluation of female researchers was due to the paucity of women among the peer reviewers. The small number of women reviewers (5 out of 55) and their uneven distribution among the MRC's committees made a statistical analysis of their scoring behaviour impossible. However, a few studies have indicated that female evaluators may be more objective in assessing the achievement of women than their male counterparts[14]. Nevertheless, we are not confident that a simple increase in the percentage of women reviewers would solve the problem of gender-based discrimination.

If gender discrimination of the magnitude we have observed is operative in the peer-review systems of other research councils and grant-awarding organizations, and in countries other than Sweden, this could entirely account for the lower success rate of female as compared with male researchers in attaining high academic rank. The United Nations has recently named Sweden as the leading country in the world with respect to equal opportunities for men and women, so it is not too far-fetched to assume that gender-based discrimination may occur elsewhere. It is therefore essential that more studies such as ours are conducted in different countries and in different areas of scientific research.

An in-depth analysis of other peer-review systems can be achieved only if the policy of secrecy is abandoned. We could perform our study only because of the Swedish Freedom of the Press Act. It is often claimed that secrecy in scoring will protect reviewers from improper influences. But our results cast doubt on these claims. It has also been suggested that the recruitment of peer reviewers of high quality would be impeded if reviewers were not granted anonymity. Such fears seem to be exaggerated because, although reviewer evaluation scores have been accessible to everyone in Sweden since the court ruling of 1995, there have been no large-scale defections of peer reviewers from the evaluation committees.

Most important, the credibility of the academic system will be undermined in the eyes of the public if it does not allow a scientific evaluation of its own scientific evaluation system. It is our firm belief that scientists are the most suited to evaluate research performance. One must recognize, however, that scientists are no less immune than other human beings to the effects of prejudice and comradeship. The development of peer-review systems with some built-in resistance to the weaknesses of human nature is therefore of high priority. If this is not done, a large pool of promising talent will be wasted.

Christine Wennerås is in the Department of Medical Microbiology and Immunology and Agnes Wold is in the Department of Clinical Immunology at Göteborg University, Guldhedsgatan 10, S-413 46 Göteborg, Sweden
(e-mail: agnes.wold@immuno.gu.se).

Acknowledgements. We thank documentalist Ann-Marie Holst for assistance, and Maria Wold-Troell, Svante Wold and Christer Andersson for statistical advice. The study was supported by a grant from the Swedish Ministry of Education.

1. Widnall, S. E. *Science* **241**, 1740–1745 (1988).
2. Cole, S., Cole, J. R. & Simon, G. A. *Science* **214**, 881–886 (1981).
3. Long, J. S. *Social Forces* **71**, 159–178 (1992).
4. Sonnert, G. *Social Stud. Sci.* **25**, 35–55 (1995).
5. Sonnert, G. & Holton, G. *Am. Sci.* **84**, 63–71 (1996).
6. Glantz, S. A. & Bero, L. A. *J. Am. Med. Assoc.* **272**, 114–116 (1994).
7. Ernst, E., Resch, K. L. & Uher, E. M. *Ann. Intern. Med.* **116**, 958 (1992).
8. Forsdyke, D. R. *FASEB J.* **7**, 619–621 (1993).
9. Caliza, L. & Gerbilsa, S. *Nature* **374**, 492 (1995).
10. Perez-Enciso, M. *Nature* **378**, 760 (1995).
11. Goldberg, P. *Trans-Action* **5**, 28–30 (1968).
12. Nieva, V. F. & Gutek, B. A. *Acad. Manag. Rev.* **5**, 267–276 (1980).
13. O'Leary, V. E. & Wallston, B. S. *Rev. Pers. Soc. Psychol.* **2**, 9–43 (1982).
14. Frieze, I. H. in *Women and Achievement: Social and Motivational Analyses* (eds Mednick, M. T., Tangri, S. S. & Hoffman, L. W.) 158–171 (Hemisphere, Washington DC, 1975).

Nature adds: This article was peer-reviewed by three males.