Reform options for peer review

SIR — Ernst, Saradeth and Resch (*Nature* **363**, 296; 1993) present valuable but worrying data on the reliability of the peer-review system of a medical journal. If their findings were found to be the rule rather than the exception, a number of reforms merit consideration: (1) there is a strong case, as suggested by the authors, for all reputable scientific journals to conduct annual blind tests of the quality of their reviewers; (2) the number of reviewers could be increased with the aim of increasing reliability; (3) reviewers could be identified to authors in an attempt to increase accountability and encourage open debate; (4) there seems to be no reason for submitted manuscripts to indicate the name or institution of the authors.

In addition we need to openly address the question of why these inconsistencies occur. Finally, in agreement with the authors, we emphasize that a balanced review system is essential for scientists, whose careers are strongly influenced by acceptance or rejection of publications.

R. H. Bradshaw
N. E. Bubier
Animal Behaviour Research Group,
Department of Zoology,
South Parks Road, Oxford OX1 3PS, UK

SIR — Ernst *et al.* suggest drawbacks with the peer review of academic articles, but overlook authors while drawing attention to a matter of concern for editors. What they describe as reproducibility is the persistent problem of reliability between markers of any examination. They describe a criterion referenced marking system and deduce, quite reasonably, that perfect reliability is seldom if ever achieved.

Any teacher or student in higher education will know that there is always going to be a variance between the judgements of individuals where there is a high degree of inference. The criteria Ernst *et al.* suggest, vary considerably in the degree of inference required; however, the criteria they suggest are all high inference variables. In practice, high reliability is associated with low inference. Low inference is often used with matters of little importance.

Take for example "linguistic merit" — one would not reasonably expect there to be a high measure of reliability between expert raters unless there were ample exemplars of the anchor points to the scale. Examples that instanced "high linguistic ability" would need to be contrasted with instances of "low linguistic merit" — and the other points on the scale would require similar examplars. With criterion referenced assessments, reliability can be improved with training and models.

Your correspondents miss a significant point with regard to editorial responsibility. Editors of learned journals exercise judgement in order to maintain standards in science. They do so imperfectly, but they accept, along with most readers, that the final judgement must lie with the reader. Errors are likely to occur in any activity mediated by humans. Examinations are no less reliable or valid than refereeing scientific publications. The lesson to learn is that reliability can be improved if models of performance are made explicit and those making judgements are properly trained.

Ray McAleese
Institute for Computer Based Learning,
Heriot-Watt University,
Edinburgh EH14 4AS, UK

SIR — Ernst *et al.* are a little subjective in their analysis of the results from their anonymous peer-reviewing exercise. They sent the same paper to 45 experts who were asked to assess the manuscript according to nine quality criteria. Ernst *et al.* tell us that the reviewers' responses were so poorly reproducible as to render peer review unacceptable.

So let us review their study. Out of 45 potential referees, only 31 (69 per cent) returned adequately completed reviews. Is this response rate unacceptable, acceptable, fair or good? It can scarcely be excellent. Further, although the data are rather poorly presented and described (quality of table = unacceptable), it seems that in only two of the nine criteria did all 31 respondents offer an opinion (scientific merit and statistical methods — which says something about current vanities). In six of the criteria, at least one of the respondents failed to offer an opinion. Two (6.5 per cent) of the respondents were even unable to make a comment about their overall judgement of the paper. These inadequacies are not discussed by Ernst and colleagues. I would thus judge the methodology and statistical methods of their study to lie somewhere between acceptable and unacceptable.

And what of the discussion? Ernst *et al.* write: "the results . . . demonstrate disappointingly poor reproducibility with extreme judgements ranging from unacceptable to excellent for most criteria". But I see it differently. On the basis of the results given, only 4.4 per cent of the opinions suggested the paper was unacceptable; 21.0 per cent of the opinions suggested the manuscript was excellent and 74.6 per cent found it acceptable, fair or good. In other words, three-quarters of the referee sample thought the paper was worth publishing, with modification. This is not an extreme range of opinions but pretty good agreement.

Some readers might think it a mite unfair to submit a short letter to *Nature* to this type of criticism. But it was submitted as a serious study of the peer-review system in order to invite comment, raise awareness, stimulate further research and so on. This then raises the question of whether *Nature* should have published such a questionable study. And the answer, of course, is "yes". The study may have flaws but it is interesting. Indeed, although the sharply critical among us might be tempted to criticize the study into the ground and deem it unacceptable, more generously minded individuals might say that the conclusions and methodology are acceptable or even good and some (perhaps those who have a grudge against the peer-review system) might hail the study as excellent. For, like it or not, the peer-review system is little more than a market research exercise (using a very small sample) to judge the response of the entire population (scientists likely to read the paper) to the matter in hand. Regarded as a crude opinion-polling exercise, peer-reviewing can be easily improved. Editors need to base their publishing judgements on the opinions of quite a few reviewers (probably five or six — but don't ask for statistical justification). They also need to make sure that they constantly vary their reviewers so that the sample remains representative. (Why should one or two individuals control all the papers on a particular subject published in a journal?). And editors need to ask their reviewers the right questions ("Do you agree with the authors' conclusions?" rather than "Comment on the scientific merit"). By contrast, regular, blind tests of the quality of reviewing, as Ernst *et al.* demand, may be well-intentioned but are futile.

Simon P. Wolff
University College London Medical School,
Rayne Institute,
5 University Street,
London WC1E 6JJ, UK

SIR — Ernst *et al.* describe the poor reproducibility of the peer-review system for publication or rejection of submitted articles. For my part, I would take encouragement from these statistics, which suggest that the selection of referees has little impact on the outcome.

Of greater concern is the finding that the main reason for rejection of the paper was criticism of the statistical methods employed; four referees assessed the statistical methods to be unacceptable, but seven judged them to be excellent. Application of a statistical method to analyse a particular dataset is either right or wrong and the 20 referees who thought that the statistical methods were acceptable, fair or good were probably not sure.

There are cases in peer-reviewed journals not only of the application of the wrong statistical methods but also of misinterpretation of the statistical findings. I

© 1993 Nature Publishing Group