Supporting and enhancing peer review in the *BJGP*

The *BJGP* has for many years operated on an open peer review system, in which a minimum of two peer reviewers report on each original research article considered for publication and where the identities of the authors and reviewers are known to each other. Although peer review remains the 'gatekeeper' to research publication, its efficacy and reliability are still a topic of controversy. There is concern about the variation in the quality of peer review, both within and between journals.[1,2] Editorial decisions such as the choice of reviewers, the interpretation of their comments, and the need to navigate between reviews offering divergent advice add to the difficulties. Formal training for reviewers is rare. Recently the ability of the system to identify fraud and plagiarism has been questioned. A 2007 Cochrane review has highlighted the urgent need for high-quality research into the outcomes of peer review.[3]

One place to focus efforts at improvement is at the level of the individual reviewer. Until now *BJGP* reviewers have not routinely received feedback on their performance, although they do receive a copy of the other review(s) and the editor's comments sent to the manuscript authors. While the quality of reviews carried out for the *BJGP* is almost uniformly good, we are now committed to implementing a more formal feedback system to help new reviewers, support existing reviewers, and further improve the quality of future reviews and publications.

EXISTING TOOLS

We examined the literature to identify existing tools used to assess the quality of peer reviews. These tools have often been devised to provide a quantitative measure of quality for comparison purposes in research studies. Most comprise a numerical scoring system to rate reviews ranging from 4-point to 100-point scales, some providing a single global score and others with multiple scores for subcategories.[4,5]

We found four published reports of the validation of existing numerical scales.

A subjective scoring system was tested in a prospective observational study carried out at the *Annals of Emergency Medicine*.[6] The scale, defined simply as 'review quality', ranged from 1 (poor) to 5 (excellent). The inter-rater reliability (IRR) was found to be moderate (intra-class correlation [ICC] = 0.44, $P<0.001$), but more positively it was noted that the tool was simple to use and easy to implement.

A more complex 5-point scale, developed by the editors of *Obstetrics and Gynaecology*, was validated and tested in a prospective observational study of the quality of 247 reviews submitted to the *Dutch Journal of Medicine*.[7] On this scale, ranging from 1 (unacceptable) to 5 (exceptional), each point was defined clearly in prose. For example, *'5 (exceptional): The rare outstanding critique that is comprehensive, objective, and insightful. Evaluates purpose of the study, study design, scientific validity, and conclusions by numbering questions and constructive suggestions to be addressed by the author. Includes comments to the editor about whether this is something new and important and useful to our readers.'* The IRR was reasonable (ICC = 0.62, 95% confidence interval = 0.55 to 0.68) and the intra-observer variability ranged from 0.66 to 0.88, demonstrating adequate test-retest reliability. The authors claimed that this score is faster and simpler for daily use than more complex systems. A weakness, however, was that scores did not follow a normal distribution, resulting in floor and ceiling effects.

The third was a more complex instrument, designed for the editors of the *Journal of Vascular and Interventional Radiology (JVIR)*. It consists of seven differently weighted subcategories reflecting both review content and format that contribute to the overall score, with a maximum of 14 points.[8] Its validity was tested in a prospective observational study of 53 *JVIR* reviews. IRR was good (ICC = 0.84, $P<0.001$) but content validation was only assessed subjectively. There are several problems regarding the transferability of this system. Firstly, the *JVIR*'s reviewers use grade sheets to help evaluate the manuscript, which were incorporated into the instrument. Secondly, the instrument was tested predominantly by reviewers themselves, not journal editors. Lastly, the instrument could be criticised for the weighting of some of its attributes. For example, timeliness could contribute up to 3 points (21% of the total score). A mediocre review completed in a timely manner may therefore score relatively highly. Indeed, when this instrument was utilised by another group in a prospective observational study, the scoring was changed to reduce the relative importance of timeliness in the overall score.[9]

Finally, the Review Quality Instrument (RQI) was developed to be a simple, reliable, and valid scale for future studies of peer review by an editorial group from the *BMJ*.[10] The scale rates seven aspects of the review, each on a 5-point Likert scale. The RQI underwent several revisions before being accepted, with testing of the tool occurring at each stage. It was subsequently used by the authors in a large randomised trial studying the effect of blinding and unmasking of reviewers on the quality of 934 reviews.[11] From this it was shown that scores had a normal distribution, with no evidence of floor or ceiling effects. IRR of the total mean score was good (weighted κ [κw] = 0.83), but was variable across subcategories (κw 0.49–0.73). The worst-performing sections included scores for importance, contructiveness, substantiation, and interpretation. The authors suggested that better training and guidelines might improve the variability in raters' scores. A further limitation is that this scoring system also puts a greater burden on editors' time (2–10 minutes) than simpler scales. Despite these drawbacks, the RQI has been applied widely in randomised trials and observational studies.

IMPACT OF FEEDBACK ON REVIEW QUALITY

Only one article was identified that specifically addressed how the provision of written feedback alters subsequent review quality.[12] This randomised trial carried out at the *Annals of Emergency Medicine* focused on

"We are now committed to implementing a more formal feedback system to help new reviewers, support existing reviewers, and further improve the quality of future reviews and publications."

poor to moderate rated reviewers based on editors' subjective quality rating (1–5 scale). Two feedback interventions were tested. The low-quality reviewers (median quality score of ≤3) received the standard journal measures, a brief summary of the specific content goals for a quality review, and the editor's numerical rating of their review. The moderately rated reviewers (median quality score of ≤4) experienced a more in-depth intervention, receiving in addition the editor's ratings of the other reviews of the same manuscript and a copy of an exemplary review of a further manuscript. Neither intervention improved reviewer performance on subsequent manuscripts. If anything, there was a more negative trend for the poorly rated reviewers, though this was not statistically significant. The authors concluded that the written feedback was ineffective. However, it is important to note that neither feedback intervention provided specific details of the problems with the review, and the subjective quality score is unlikely to have been meaningful for the reviewers. Additionally, the reviewers involved in the study did not self-select; it is likely that feedback is a more effective educational tool for those who actively seek it.

SURVEY OF *BJGP* REVIEWERS

Given the large body of literature in educational research to suggest that feedback does improve performance and the current culture of self-improvement within medicine, we still felt this was a timely opportunity to use feedback in the improvement of peer review at the individual level. We surveyed *BJGP* reviewers to determine where the focus of feedback should lie, the form it should take, and who would most benefit from it.

We invited 120 *BJGP* reviewers with a range of experience to complete a short online questionnaire regarding feedback for peer review. There was a 58% response (70 reviewers). Although reviewers with different levels of experience responded, including those who had not yet reviewed for the journal, there was little difference in the responses between those who were more and those who were less experienced.

Most reviewers (93%) said that they would value feedback for their reviews and would find it useful for improving future reviews. A further 65% felt that feedback on every review would be the most appropriate, although there was a significant subgroup (17%) who said that only initial reviews warranted feedback. Individualised written feedback was the most popular format (47%), followed by a number of scores in subcategories (31%). A single score was the least popular (21%).

The questionnaire also allowed free text responses, with several common themes emerging. Reviewers felt the current system of receiving decision letters and a copy of other review(s) was already very helpful. While routine personalised feedback would be greatly appreciated and useful for professional development, there was recognition that this may overload the journal. Feedback for novice reviewers and annual reports were among suggestions of valuable forms of feedback. There was also support for peer review workshops as a training tool.

Box 1. The *BJGP* review scoring criteria

Grade	Description
A	An excellent and timely review, providing a set of comments that are comprehensive, insightful, and clear, and are informed by a close familiarity with the topic and/or the methodology of the study. There is a clear recommendation on acceptance for publication, consistent with these comments, which are structured, immediately comprehensible to the authors, and which can act as a constructive guide to redrafting and resubmission. There are useful comments to the editor about matters such as the novelty, importance, and likely interest to readers of the *BJGP*. These top-class reviews often suggest additional literature and references for consideration by the authors.
B	A very good review, with useful and timely guidance for the editor, clear comments to the authors, and sufficient detail for resubmission and redrafting, although perhaps with less subject or methodological expertise, less incisiveness, and perhaps also missing some key details. Like Grade A reviews, these reviews are likely to run to at least 40 or 50 lines of comment, providing sufficient material to not only help authors improve their manuscript but also to reflect on their methods, findings, and interpretation.
C	An adequate review that is still useful, but may not provide a comprehensive opinion or absolutely clear advice to the editor. This may be problematic when a more detailed review has come to a different conclusion about quality or a different recommendation on acceptance, so that a further review may be needed to supplement the shortcomings in the Grade C report.
D	An evaluation that is too brief and superficial to be useful. It not only fails to identify significant shortcomings in the study, but also is too thin to be used as a basis for rejection. A very short review of this kind recommending acceptance can be equally unhelpful, particularly when it has to be weighed against a more guarded opinion in a more detailed report.
E	A review that is short, dismissive, or mildly offensive, with evidence of bias or personal animosity, with no attempt to provide objective or constructive comments and with very weak academic/intellectual content.

ADDRESS FOR CORRESPONDENCE

Abigail Moore
University of Oxford, Department of Primary Health Care Sciences, New Radcliffe House, Radcliffe Observatory Quarter, Woodstock Road, Oxford, OX2 6NW

E-mail: abigail.moore@medsci.ox.ac.uk

A NEW FEEDBACK SYSTEM FOR THE *BJGP*

Taking into account the literature review, the questionnaire results, the *BJGP*'s AllenTrack submission software, and the time and work burden of providing feedback by the editorial staff, the *BJGP* will now be introducing a new feedback system which will offer:

- routine provision of feedback on a 5-point scoring system based on the system devised by the editors of *Obstetrics and Gynaecology* (Box 1),[7] adapted for the A, B, C, D, E system already built into AllenTrack, which allows reviewers to see their mean review rating; and
- narrative feedback on request on up to two occasions per year, providing comments along the lines of the review criteria, expanding and advising where necessary. This may be of particular interest to new reviewers.

CONCLUSION

Peer review is an imperfect system, but is probably the best method we have to safeguard original research publication. We hope that the new feedback system implemented at the *BJGP* will go some way to improving the quality and consistency of the journal's own peer review process.

Abigail Moore,
Foundation Doctor, Oxford University Hospitals, Oxford, UK.

Roger Jones,
BJGP Editor, RCGP, London.

Provenance
Freely submitted; not externally peer reviewed.

REFERENCES

1. Gasparyan AY, Kitas GD. Best peer reviewers and the quality of peer review in biomedical journals. *Croat Med J* 2012; **53(4):** 386–389.
2. Marusic A, Lukic IK, Marusic M, *et al.* Peer review in a small and a big medical journal: case study of the *Croatian Medical Journal* and the *Lancet*. *Croat Med J* 2002; **43(3):** 286–289.
3. Jefferson T, Rudin M, Brodney Folse S, Davidoff F. Editorial peer review for improving the quality of reports of biomedical studies. *Cochrane Database Syst Rev* 2007; **18:** Mr000016.
4. McNutt RA, Evans AT, Fletcher RH, Fletcher SW. The effects of blinding on the quality of peer review. A randomized trial. *JAMA* 1990; **263(10):** 1371–1376.
5. Schroter S, Tite L, Hutchings A, Black N. Differences in review quality and recommendations for publication between peer reviewers suggested by authors or by editors. *JAMA* 2006; **295(3):** 314–317.
6. Callaham ML, Baxt WG, Waeckerle JF, Wears RL. Reliability of editors' subjective quality ratings of peer reviews of manuscripts. *JAMA* 1998; **280(3):** 229–231.
7. Landkroon AP, Euser AM, Veeken H, *et al.* Quality assessment of reviewers' reports using a simple instrument. *Obstet Gynecol* 2006; **108(7):** 979–985.
8. Feurer ID, Becker GJ, Picus D, *et al.* Evaluating peer reviews. Pilot testing of a grading instrument. *JAMA* 1994; **272(2):** 98–100.
9. Bingham CM, Higgins G, Coleman R, Van Der Weyden MB. The *Medical Journal of Australia* internet peer-review study. *Lancet* 1998; **352(9126):** 441–445.
10. van Rooyen S, Black N, Godlee F. Development of the review quality instrument (RQI) for assessing peer reviews of manuscripts. *J Clin Epidemiol* 1999; **52(7):** 625–629.
11. van Rooyen S, Godlee F, Evans S, *et al.* Effect of blinding and unmasking on the quality of peer review. *J Gen Intern Med* 1999; **14(10):** 622–624.
12. Callaham ML, Knopp RK, Gallagher EJ. Effect of written feedback by editors on quality of reviews: two randomized trials. *JAMA* 2002; **287(21):** 2781–2783.