Open peer review: a randomised controlled trial

ELIZABETH WALSH, MAEVE ROONEY, LOUIS APPLEBY and GREG WILKINSON

Background Most scientific journals practise anonymous peer review. There is no evidence, however, that this is any better than an open system.

Aims To evaluate the feasibility of an open peer review system.

Method Reviewers for the *British Journal of Psychiatry* were asked whether they would agree to have their name revealed to the authors whose papers they review; 408 manuscripts assigned to reviewers who agreed were randomised to signed or unsigned groups. We measured review quality, tone, recommendation for publication and time taken to complete each review.

Results A total of 245 reviewers (76%) agreed to sign. Signed reviews were of higher quality, were more courteous and took longer to complete than unsigned reviews. Reviewers who signed were more likely to recommend publication.

Conclusions This study supports the feasibility of an open peer review system and identifies such a system's potential drawbacks.

Declaration of interest G.W. is the Editor, L.A. an Assistant Editor and E.W. and M.R. Trainee Editors of the *British Journal of Psychiatry.*

Research into the peer review process has gathered pace over the past decade. The most notable recent development has been the opening-up of the process by the *British Medical Journal* (Smith, 1999). Reviewers had previously been reluctant to sign their reviews when asked (McNutt *et al*, 1990; van Rooyen *et al*, 1999), and no significant differences in review quality (McNutt *et al*, 1990; Godlee *et al*, 1998; van Rooyen *et al*, 1999), recommendation regarding publication or time taken to review (van Rooyen *et al*, 1999) have resulted from signing. As most research has been conducted in general medical journals, it has been suggested that similar studies be carried out in specialist journals in order to determine whether these results are generalisable.

The *British Journal of Psychiatry* currently operates a closed peer review system. For an open system to be practical, reviewers would have to be in favour of signing, and the quality of reviews produced ought not to be of inferior quality. In order to assess the feasibility of such a system, we conducted this study, which addresses the following questions:

(a) How many reviewers will agree to sign their review?

(b) Are signed reviews of higher quality than unsigned reviews?

(c) Are reviewers who sign their name more courteous?

(d) Are reviewers who sign their name more likely to recommend publication?

(e) Will reviewers who sign their name take longer to complete their reports?

METHOD

A postal questionnaire was sent by the Editor to the 322 reviewers on the *Journal*'s database; it asked whether they would participate in the trial, which would mean having their name revealed to the authors whose papers they reviewed.

Those who agreed to participate were included in the trial. Over an 18-month period, each time a participating reviewer was selected to review an unsolicited manuscript containing original research, the review was randomised to the signed or unsigned group. When randomised to the signed group the reviewer's name would be revealed to the author(s) of the manuscript. When randomised to the unsigned group the reviewer would remain anonymous. Individual referees' reviews could be randomised to different groups on different occasions, ensuring that only one key element differed between the groups, namely signing or not signing. Simple randomisation was performed using computer-generated random numbers.

On the basis of a previous study (Black *et al*, 1996) we decided that an editorially significant difference in review quality scores would be 0.4/4 (10%). In order to detect this difference it was calculated that we would need over 100 manuscripts (α=0.05, β=0.10, s.d.=1.5) in each arm of the study. It was, however, possible for randomisation to continue over the 18-month period; hence, a considerably greater number of reviews were randomised.

Throughout the trial the Editor continued the usual process of allocation of articles to reviewers. He was unaware which reviewers were participating in the trial. All reviewers were aware of the details of the author(s). Participating reviewers were specifically requested to include all their comments in their report and to refrain from placing additional confidential comments in a letter to the Editor.

On returning each review, participants were asked to complete a second questionnaire indicating their recommendation for publication (accept without revision, accept with revision, reconsider with revision, or reject) and the time taken to complete the review. Once reviews were received, secretarial staff made three copies of each review. One was passed to the Editor to aid Editorial decisions in the usual way. The remaining two were rendered anonymous and sent to two Trainee Editors (E.W., M.R.) who independently assessed review quality. The original reviews were filed.

Review quality was assessed by using a validated review quality instrument (Black *et al*, 1996), which consists of seven items, each scored on a five-point Likert scale. The instrument assesses whether the reviewer

has addressed the importance of the research question, the originality of the article, the methodology used including statistical analysis, and the organisation and writing style. It also assesses whether comments made by the reviewer were constructive, whether there was provision of examples from the paper to substantiate the comments and whether there were comments on the authors' interpretation of results. The Trainee Editors also rated the tone of each review (1, abusive; 5, excellent). Once all reviews were returned, the authors of papers randomised to the signed group were informed of the identity of their paper's reviewer.

The Trainee Editors and the Editor each rated 30 non-randomised reviews prior to the study in order to identify possible differences in ratings. Where disagreement arose, a discussion was held in order to clarify the instrument ratings.

Throughout the trial period, reviews returned by those who had declined to reveal their names to authors were rated on quality. These reviews were rated blind, interspersed with the other reviews, by the Trainee Editors. A mean quality score was obtained for each review from the review quality instrument using the mean of the two raters' scores. It was thought that the 'decliners' group may differ systematically from the participating group in terms of review quality. An attempt was made to estimate any differences.

Statistical analysis

Data were entered into the software package SPSS (SPSS, 1996) for analysis. The scores for each of the seven items and the overall quality score for the review quality instruments were based on the mean of the scores of the two Trainee Editors. The mean difference between scores in the signed and unsigned groups and the 95% confidence intervals of the difference were calculated for each item. Independent samples *t*-tests were used to compare mean scores. A between-group comparison of the time taken to complete the review and of the tone of the review was also carried out, using Student's *t*-tests. Recommendation for publication was compared between the groups using Pearson's χ^2 test.

Using the mean of the two Trainee Editors' scores was thought to improve the reliability of the method. Interrater reliability between the two Trainee Editors was assessed by using weighted κ statistics. Student's *t*-tests were used to compare the quality of reviews produced by decliners with the quality of those produced by both the signed and unsigned groups.

RESULTS

Agreement to sign

Of the 322 referees sent postal questionnaires, 279 (87%) replied. Of these 279, 245 (88%) agreed to participate in the trial while 34 (12%) declined. Hence, 245 (76%) of the 322 reviewers sent questionnaires agreed to participate, 34 (11%) refused and 43 (13%) failed to respond. Reminders were not sent to reviewers who failed to respond.

Randomisation

During the study period, 498 reviews were randomised – 222 (54%) to the signed group and 186 (46%) to the unsigned (see Fig. 1). A total of 358 (88%) reviews were returned, 194 (87%) in the signed group and 164 (88%) in the unsigned group. Although those who failed to complete reviews were not asked why, in the majority of cases an accompanying letter outlined difficulties due to heavy workloads and an inability to complete the review in the specified three-week period. In a number of cases reviewers returned manuscripts because they felt that it would be inappropriate to review a close colleague's work. One hundred and thirty-five reviewers completed reports, with the majority completing one (58 reviewers) two (16) or three (20) reports. Six reviewers completed more than 10 reports; randomisation ensured that these reports were evenly distributed between the groups.

Interrater reliability

The following weighted κ statistics were calculated from the scores of the two Trainee Editors on the seven review quality items. Importance of research (κ=0.52), originality (0.72), methodology (0.71), presentation (0.72), constructiveness of comments (0.71), substantiation of comments (0.68) and interpretation of results (0.67). A score of 0.4–0.75 represents fair to good agreement.

Review quality

When comparing the quality of reviews, the total mean score was significantly higher in the signed group than in the unsigned group (3.35 *v.* 3.14, P=0.02) (Table 1). Of the seven individual items rated, signed reviews scored higher than unsigned reviews on all items, and significantly so on three (methodology, presentation and constructiveness of comments). In tone, signed reviews were significantly more courteous and less abusive than unsigned reviews.

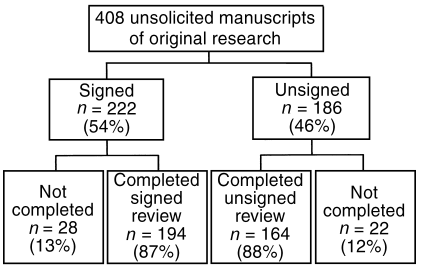

Fig. 1 Randomisation to signed and unsigned groups.

Time taken to complete review

The response rate regarding the time taken to complete reviews was low, with only 222 reviewers (54%) providing information, but this response rate was similar for the two groups. The mean time taken to complete a review was 2.05 hours in the signed group and 1.65 hours in the unsigned group (P=0.02). Of those who took more than four hours to complete the review, 69% were in the signed group and 31% in the unsigned group.

Recommendation for publication

Information regarding recommendation for publication was provided for 320 (89%) reviews (Table 2). Signed reviewers were significantly less likely to recommend rejection of manuscripts (18%) than unsigned reviewers (33%) (χ^2=10.81, d.f.=3, P<0.01). Similar numbers in the two groups recommended accepting papers without revision, but more signed reviewers recommended acceptance following revision and resubmission with revision.

Decliners

Fifty-seven reviews (non-randomised) were returned by the referees who refused to participate in the trial. The mean overall quality score in this group was 2.78. The mean difference compared with the signed group was 0.56 (95% CI 0.31–0.81) which

Table 1 Review quality ratings by item in signed and unsigned groups

Quality item	Quality rating (mean (s.d.))		
	Signed group	Unsigned group	Difference (95% CI)
Importance of research question	3.02 (1.07)	2.85 (1.12)	0.16 (−0.06 to 0.39)
Originality	2.85 (1.16)	2.71 (1.18)	0.15 (−0.09 to 0.39)
Methodology	3.63 (1.00)	3.40 (0.93)	0.23 (0.02 to 0.43)*
Presentation	3.30 (1.03)	3.04 (1.01)	0.26 (0.05 to 0.48)**
Constructiveness of comments	3.75 (0.94)	3.48 (0.93)	0.27 (0.07 to 0.47)**
Substantiation of comments	3.45 (1.05)	3.25 (1.03)	0.20 (−0.01 to 0.42)
Interpretation of results	4.43 (1.16)	3.25 (1.14)	0.18 (−0.06 to 0.41)
Mean score	3.35 (0.86)	3.14 (0.86)	0.21 (0.03 to 0.39)*
Tone of review	4.51 (0.65)	4.27 (0.91)	0.25 (0.09 to 0.42)**
Time taken	2.05 (1.25)	1.65 (1.33)	−0.39 (−0.74 to 0.06)*

*$P < 0.05$, **$P < 0.01$.

Table 2 Recommendation on publication in signed and unsigned groups

Recommendation for publication	Signed group (*n* (%))	Unsigned group (*n* (%))	Total (*n* (%))
Accept without revision	41 (25%)	36 (24%)	77 (24%)
Accept with revision	53 (31%)	35 (23%)	88 (28%)
Resubmit with revision	43 (26%)	31 (20%)	74 (23%)
Reject	30 (18%)	51 (33%)	81 (25%)
Total	167 (100%)	153 (100%)	320 (100%)

represented a significantly lower quality score ($P < 0.001$). The mean difference compared with the unsigned group was 0.35 (95% CI 0.1–0.62), again representing a significantly lower quality score ($P = 0.006$).

DISCUSSION

We wished to estimate how many reviewers would agree to sign their review, whether signed reviews were of higher quality and more courteous than unsigned reviews, and whether signing would influence recommendation for publication or the time taken to complete the review. Clarification of these issues should allow a judgement of the feasibility of opening up the review process to be made.

Of the referees approached, 76% agreed to sign their name. This figure compares with 43% (McNutt *et al*, 1990) and 70% (van Rooyen *et al*, 1999) in previous studies, and is the highest reported to date. When considering opening the review process on this basis, the results suggest that one may lose up to one-quarter of reviewers, making the editorial process more difficult and increasing the workload of the remaining referees. As 43 (13%) of those approached failed to reply, a higher agreement rate may have been possible.

With regard to review quality, at the outset of the trial a difference between the groups of 10% (0.4/4) was chosen as editorially significant. Although signed reviews were of statistically significantly higher quality than unsigned reviews, the mean difference was only 0.21, representing a percentage difference of only 5.5% (range 0.75–0.75%). This compares with a mean difference of 0.03, representing a percentage difference of 0.75%, in the largest randomised trial conducted before this one (van Rooyen *et al*, 1999). With such a wide range it is difficult to draw any firm conclusions about improved review quality. Another way of estimating the effect of signing is calculating the number needed to treat (Cook & Sackett, 1995). This represents the number of reviews that would on average need to be signed in order to produce one review of better quality. It is calculated by estimating the proportion of reviews in each group which were judged to be of low quality (scores on the review quality instrument of less than 4), subtracting the proportion in the signed group (0.74) from that in the unsigned group (0.85), and obtaining the reciprocal of the difference (9.09, 95% CI 9.01–9.17). Hence, on average, nine reviewers would need to sign their reports in order to produce one review of higher quality. What can be inferred from these results is that review quality did not suffer as a result of signing. Although we found a statistically significant difference in that named reviewers provided higher-quality reviews, it is doubtful whether this reflects an important difference in quality.

Signed reviews were found to be significantly more courteous. However, with mean scores of 4.51 and 4.27 in the signed and unsigned groups, respectively, it is clear that the majority of reviews were at the courteous end of the scale in both groups. It is therefore unlikely, on the present evidence, that authors in receipt of unsigned reports would suffer hostile or abusive comments.

Signed reviews took significantly longer to complete than unsigned reviews. Having to sign his or her name appears to make a reviewer spend longer on the review, possibly by checking references or reading about the statistical methods more carefully, for example. This extra time spent would be expected to enhance the quality of the report. This finding does, however, suggest that the time commitment involved in reviewing might be too arduous for referees if the peer review process were opened up, especially when one considers the increased workload resulting from the loss of reviewers who refuse to sign their names.

Although signing appears to make reviewers more likely to recommend publication and less likely to recommend rejection of papers, it is important to remember the role of the Editor in this process. At the *British Journal of Psychiatry* it is not unusual for a manuscript to be sent to four or more referees and for divergent opinions to be expressed regarding suitability for publication. The Editor frequently has to make difficult judgements in the light of these disagreements. If the process were opened up, the Editor might have to modify

his practice and become more autonomous in making decisions, perhaps necessitating rejection of more manuscripts than recommended by reviewers. After all, publication remains an editorial rather than a scientific decision, and the question asked of reviewers is whether there is any major impediment to publication on scientific grounds.

The quality of the 57 reviews completed by those who refused to have their names revealed to authors was significantly lower than that of both signed and unsigned reviews from participating reviewers. Only the signed reviews, however, were of editorially significantly higher quality (14%, range 8.3–20.5%). These findings may suggest that the loss of decliners from an open peer review process may be beneficial in terms of review quality. However, it is possible that the Hawthorne effect played a role, as both signed and unsigned groups were aware that they were participating in a study.

CLINICAL IMPLICATIONS

- A sufficient number of reviewers will agree to sign their names, making an open process feasible.
- Signed reviews are at least as good as unsigned reviews, and may be of better quality.
- Signed reviews take longer to complete, leading to a greater workload for reviewers.

LIMITATIONS

- The review quality instrument used, although validated, is open to subjective interpretation.
- Reasons for some reviewers not wishing to sign were not sought.
- We do not know why some randomised reviews were not returned.

ELIZABETH WALSH, MRCPsych, Institute of Psychiatry, London; MAEVE ROONEY, MRCPsych, Maudsley Hospital, London; LOUIS APPLEBY, MD, School of Psychiatry and Behavioural Sciences, University of Manchester; GREG WILKINSON, FRCPsych, University Department of Psychiatry, Royal Liverpool University Hospital, Liverpool

Correspondence: Dr Elizabeth Walsh, Department of Psychological Medicine, Institute of Psychiatry, De Crespigny Park, London SE5 8AF

(First received 12 July 1999, final revision 29 September 1999, accepted 30 September 1999)

Methodology

This study has many strengths. They include the large sample size, the randomised controlled design, the availability of a validated quality instrument, the use of blind ratings, and good interrater reliability as measured by weighted κ statistics. The study does, however, have several limitations. No reminders were sent to referees who failed to respond to the postal questionnaire. As referees give their valuable time free of charge, it was felt to be inappropriate to bother them with reminders. It is therefore possible that we could have gained a higher response rate. Those who refused to participate were not asked to provide a reason for their refusal. This number may have been too small to allow us to draw any meaningful conclusions. Identification of the reasons for 13% of reviews in the signed group not being returned would have been useful. Although these reviewers agreed to sign their names at the outset of the trial, when actually faced with doing so they may have decided against it. However, the fact that a similar percentage (12%) of reviews was not returned in the unsigned group suggests that this may be the normal rate of non-response for all reviews in general. Randomisation of reviews meant that individual reviewers could be allocated to either group on different occasions. It could therefore be argued that the samples were not entirely independent. However, the use of a paired analysis was inappropriate, as 58 reviewers completed only one review.

The study's final limitation concerns the measurement of quality. The quality review instrument, although validated, is open to some subjective interpretation. Although it can assess the quality of a review, it is unable to determine its accuracy (van Rooyen *et al*, 1999).

Feasibility of open peer review

Those opposed to open peer review put forward convincing arguments in favour of maintaining the status quo. Junior reviewers may hinder their career prospects by criticising the work of powerful senior colleagues or be intimidated into writing inappropriately favourable reviews. Unwanted, inappropriate or even acrimonious dialogue may occur between author and reviewer, and professional relationships may suffer. Reviewers may become less critical, and scientific standards may decline. Some people ask why we should interfere with a system which appears to be functioning adequately without good evidence that there is a better way (Hyams, 1996).

Increased accountability in the reviewing process is essential, however. This is because it has become so important to publish in good journals, not only for the careers of individuals but also for the funding of institutions through the Research Assessment Exercise. Reviewers give their valuable time free of charge and with little credit, yet they are performing an important job which plays a part in shaping our scientific future. It is critical that they do this job in the best possible way. By signing their name to a review they automatically become more accountable. Editors are forced to seek the best possible opinions for manuscripts and the editorial process is improved. Authors who are aware of the identity of their reviewer may also be less upset by hostile and discourteous comments (McNutt *et al*, 1990).

This study's findings certainly support the feasibility of open peer review. With

three-quarters of referees agreeing to sign their name, and with signed reviews being of higher quality than unsigned reviews (although not editorially so), a sufficient number of reviewers producing reviews of sufficient quality would probably be available. The drawbacks would be the loss of those reviewers opposed to signing, the increased time that a signed review appears to take, and the reduced reliance that an editor could place on the recommendation for publication from reviewers who may suggest acceptance of too many manuscripts. Before opening the system, a closer examination of the possible adverse effect that such a system may have on professional relationships in a fairly close-knit field is warranted – and such a study is planned for the current batch of signed reviews. Further research of the peer-review process should be encouraged not only in other journals but – as similar arguments could be directed at the process by which research funding is awarded – also by the research funding bodies.

ACKNOWLEDGEMENTS

We thank Susan van Rooyen and Fiona Godlee for their advice on the design of the project and use of the rating instrument; Glyn Lewis and Tony Johnson for background advice; Nick Black and Andrew Hutchings for advice; Zofia Ashmore and Sue Thakor for their major input into the practical aspects of the trial; David Jago for his support; and all the participating referees and the Editorial Board of the *British Journal of Psychiatry* for their helpful comments.

REFERENCES

Black, N., van Rooyen, S., Godlee, F., *et al* (1996) What makes a good reviewer and a good review for a general medical journal. *Journal of the American Medical Association*, **280**, 231–233.

Cook, R. J. & Sackett, D. L. (1995) The number needed to treat: a clinically useful measure of treatment effect. *British Medical Journal*, **310**, 452–454.

Godlee, F., Gale, C. R. & Martyn, C. N. (1998) Effect on the quality of peer review of blinding reviewers and asking them to sign their reports. A randomised controlled trial. *Journal of the American Medical Association*, **280**, 237–240.

Hyams, K. C. (1996) Letter. *Lancet* **34**, 132–133.

McNutt, R. A., Evans, A. T., Fletcher, R. H., *et al* (1990) The effects of blinding on the quality of peer review: a randomised trial. *Journal of the American Medical Association*, **263**, 1371–1376.

Smith, R. (1999) Opening up *BMJ* peer review. A beginning that should lead to complete transparency. *British Medical Journal*, **318**, 4–5.

SPSS (1996) *SPSS for Windows: Base System User's Guide. Release 7.5.1.* Chicago, IL: SPSS Inc.

van Rooyen, S., Godlee, F., Evans, S., *et al* (1999) Effect of open peer review on quality of reviews and on reviewers' recommendations: a randomised trial. *British Medical Journal*, **318**, 23–27.