Effect of Blinding and Unmasking on the Quality of Peer Review

A Randomized Trial

Susan van Rooyen, BSc; Fiona Godlee, MRCP; Stephen Evans, MSc; Richard Smith, FRCP; Nick Black, MD

Context.—Little research has been conducted into the quality of peer review and, in particular, the effects of blinding peer reviewers to authors’ identities or masking peer reviewers’ identities.

Objective.—To determine whether concealing authors’ identities from reviewers (blinding) and/or revealing the reviewer’s identity to a coreviewer (unmasking) affects the quality of reviews, the time taken to carry out reviews, and the recommendation regarding publication.

Design and Setting.—Randomized trial of 527 consecutive manuscripts submitted to BMJ, which were randomized and each sent to 2 peer reviewers.

Interventions.—Manuscripts were randomized as to whether the reviewers were unmasked, masked, or uninformed that a study was taking place. Two reviewers for each manuscript were randomized to receive either a blinded or an unblinded version.

Main Outcome Measures.—Mean total quality score, time taken to carry out the review, and recommendation regarding publication.

Results.—Of the 527 manuscripts entered into the study, 467 (89%) were successfully randomized and followed up. The mean total quality score was 2.87. There was little or no difference in review quality between the masked and unmasked groups (scores of 2.82 and 2.96, respectively) and between the blinded and unblinded groups (scores of 2.87 and 2.90, respectively). There was no apparent Hawthorne effect. There was also no significant difference between groups in the recommendations regarding publication or time taken to review.

Conclusions.—Blinding and unmasking made no editorially significant difference to review quality, reviewers’ recommendations, or time taken to review. Other considerations should guide decisions as to the form of peer review adopted by a journal, and improvements in the quality of peer review should be sought via other means.

JAMA. 1998;280:234-237

PEER REVIEW has a key role in determining which original research is published and thus becomes part of the accepted body of scientific knowledge. Despite its central role, little research has been conducted into the relative benefits or effectiveness of different approaches to peer review. We decided to examine 2 questions: what are the effects of blinding reviewers to the identity of the authors of a manuscript and of unmasking (revealing the identity of) reviewers to their coreviewers?

There are several reasons for believing that blinding may be beneficial. First, blinded reviewers may provide less biased reviews. Second, some editors believe blinding improves the quality of reviews, a belief supported by one small randomized controlled trial. Finally, articles that appear in journals that use blinded review are more likely to be cited than those published in journals that use nonblinded review. Unmasking the identity of reviewers to one another has not previously been studied, although many journals already carry out the practice and believe it to result in higher-quality reviewing.

By means of a randomized trial, we set out to evaluate the effects of blinding (concealing the identity of authors from a reviewer), unmasking (revealing the identity of a reviewer to a coreviewer), and a combination of the 2 on the quality of reviews. The study also sought to establish the feasibility of successful blinding.

METHOD

Consecutive manuscripts in the categories of research articles, short reports, and research articles from general practice (also known as family medicine or primary care) received by BMJ and sent by editors for peer review between January and June 1997 were eligible for inclusion. Manuscripts were randomized (stratified by the 3 above-mentioned categories) into 1 of 3 groups: 2 intervention groups (masked and unmasked) and an uninformed group (Figure). The randomization process was undertaken by a researcher who was independent from the editorial decision-making process using a computerized minimization program with a random component. Each manuscript was sent to 2 paid clinical reviewers selected by whoever of the 11 editors was responsible for the particular manuscript. In both the masked and unmasked groups the reports of pairs of reviewers were exchanged. Reviewers in the unmasked group were asked to consent to their identity being revealed to their coreviewer. Ethical approval was granted by a university ethics committee.

Having randomly allocated the manuscripts, the reviewers in the masked and unmasked groups were randomized to receive either a blinded or an unblinded version of the manuscript. Blinding consisted of removing authors’ details from the title page and acknowledgments. No attempt was made to remove authors’ details from within the text of the manuscript, the illustrations, or the references. Blinded reviewers were asked whether they thought they knew the identity of the author(s), and if so, to detail the name(s) and/or the institution and to explain why they thought they could tell. All reviewers in the intervention groups were also asked to record with whom they interacted as a result of blinding and whether the interaction had any effect on their recommendation regarding publication of the manuscript. If 1 of the 2 reviewers of a manuscript in the unmasked group withheld consent, the manuscript was transferred into a pref-
Since awareness of being in a study might affect the reviewers’ behavior, an un-informed group was included that allowed us to test for a Hawthorne effect. Manuscripts in the un-informed group were sent to 2 reviewers who were not informed that a study was taking place. Care was taken that those who had reviewed manuscripts in the masked or un-masked group were not subsequently selected to review manuscripts in the un-informed group. At no stage were editors or authors aware of the group to which a manuscript had been allocated.

On receipt of both reviews of a manuscript, the reviews, with authors’ details removed from them, were passed together with the manuscript to the responsible editor, who was asked to assess the quality of the reviews. All the documents were subsequently returned to the researcher, who passed the manuscript to a second editor randomly selected from the remaining 10 editors taking part in the study for a second, independent evaluation. The quality of the reviews was assessed using a validated review quality instrument developed from an instrument used in a previous study. A decision on whether to publish the article was made in the journal’s usual manner. At least 10 days after the decision had been communicated to the authors, the corresponding author was asked to evaluate the 2 reviews using the review quality instrument. The authors were also asked whether they thought each reviewer had been blinded to their identity.

The review quality instrument consisted of 7 items (importance of the research question, originality, methodology, presentation, constructiveness of comments, substantiation of comments, interpretation of results), each scored on a 5-point Likert scale (1 = poor, 5 = excellent). A total score was based on the mean of the 7 item scores. A full version of the instrument has been reported on elsewhere. In addition, a global item seeking an overall assessment of the quality of the review was included. The quality of each review was based on the means of the 2 editors’ scores for each item and total score and on the corresponding author’s scores. This article considers only the editors’ assessments of review quality. The means of 2 editors’ scores were used to improve the reliability of the method. Data collection from authors is still continuing and will be reported later. Two additional outcome measures were used: the time taken to carry out the review and the editorial decision (accept, revise, reject).

It was calculated in advance that in order to detect an editorially significant difference in review quality scores of 0.4 ($\alpha = .05$, $\beta = .10$, SD = 1.5), 148 manuscripts would be required in each of the masked, un-masked, and un-informed groups. Recruitment of manuscript was continued until we were certain of retaining at least 148 in each group after taking account of exclusions and losses after randomization.

Analysis used independent comparisons of outcome measures between masked and un-masked reviewers (excluding manuscripts for which one reviewer had withheld consent for unmasking), paired comparisons between blinded and un-blinded reviewers and independent comparisons between masked un-blinded reviewers and un-informed reviewers, using $t$ tests. Two-way analysis of variance was used to compare the 2 factors in the 4 intervention arms of the study.

## RESULTS

### Recruitment and Randomization

Between January and June 1997, an estimated 570 eligible manuscripts were sent for peer review. Of these, 43 were not entered into the study, either as a result of an administrative error or because, in the case of 5 pairs of articles by the same authors, a decision was made that only the first article would be included. The 527 manuscripts (92%) included consisted of 393 research articles, 74 short reports, and 60 general practice research articles. Of these 527 manuscripts, 60 were excluded after randomization, either because it proved impossible to obtain 2 suitable reviews without causing an unacceptable delay in the editorial decision-making process or because a reviewer who was randomized to receive a blinded manuscript had the authors’ identity revealed in error. The distribution of short reports, research articles, and general practice articles was similar for the exclusions and for the total sample.

The remaining 467 manuscripts were randomized to the masked group ($n = 149$), the un-masked group ($n = 160$), and the un-informed group ($n = 158$). Of the 160 manuscripts in the un-masked group, 10 of the 320 reviewers did not give consent to their identity being revealed. These 10 manuscripts were included in the preference arm (Figure). Successful follow-up was achieved for all 467 manuscripts.

In order to assess the success of randomization, we compared characteristics of the manuscripts (geographic origin) and the reviewers (mean age, residence in North America, postgraduate training in epidemiology or statistics, involved in medical research). There were no striking differences between groups. Exclusions did not introduce any bias.

### Success of Blinding

Of the 309 blinded reviewers, 293 (95%) replied to the question concerning whether they could identify the authors of the manuscript. With successful blinding defined as either author not identified or author identified incorrectly, 170 reviewers (58%) were successfully blinded (Table 1). The main reasons given for being able to identify the author included self-referencing, clues contained within the text of the manuscript, and a small research field. If successful blinding is extended to include those who were only partially successful in identifying authorship (for example, named one author correctly but others incorrectly), then 196 reviewers (67%) were successfully blinded.
Extent of a Hawthorne Effect

There was no evidence of any difference between masked unblinded and uninformd reviewers and therefore no detectable Hawthorne effect (Table 2).

Effect of Blinding and Unmasking on Review Quality

The mean total quality score was 2.87. There was little or no difference in total or item scores between blinded and unblinded reviewers or between masked and unmasked reviewers (Table 2). The largest difference in mean total score was only 0.14. Although some of the differences were statistically significant ($P < .05$), showing that unmasking tended to produce higher-quality reviews, in absolute terms these differences were not editorially significant. Although 2-factor analysis of variance revealed a statistically significant difference between the masked and unmasked groups ($P = .04$), absolute differences were of no editorial significance (blinded/unblinded $P = 26$, interaction $P = .41$, overall $P = .05$). Analyses based only on those successfully blinded (170 reviewers) led to similar results.

Table 1.—Assessment of the Success of Blinding

<table>
<thead>
<tr>
<th>No. of Reviewers (%)</th>
<th>Author correctly identified</th>
<th>Author partially identified</th>
<th>Author identified incorrectly</th>
<th>Author not identified</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>97 (33.1)</td>
<td>26 (8.9)</td>
<td>20 (6.8)</td>
<td>150 (51.2)</td>
<td>293 (100.0)</td>
<td></td>
</tr>
</tbody>
</table>

*Percentage of those who answered the question (n = 293); 16 (5.2%) of the 309 reviewers did not respond.

Effect of Blinding and Unmasking on Editorial Decision and Review Time

Although 2-factor analysis of variance revealed a statistically significant difference between the blinded and unblinded groups ($P = .04$), absolute differences were of no editorial significance (blinded/unblinded $P = 26$, interaction $P = .14$, overall $P = .05$). Analyses based only on those successfully blinded (170 reviewers) led to similar results.

Table 2.—Comparison of Review Quality and Review Time*

<table>
<thead>
<tr>
<th>Masked vs Unmasked Reviewers</th>
<th>Mean (SD) Score</th>
<th>Difference, Mean (95% CI)</th>
<th>Blinded vs Unblinded Reviewers</th>
<th>Mean (SD) Score</th>
<th>Difference, Mean (95% CI)</th>
<th>Masked vs Uninformed Reviewers</th>
<th>Mean (SD) Score</th>
<th>Difference, Mean (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Importance</td>
<td>2.44 (0.92)</td>
<td>−0.09 (−0.22 to 0.05)</td>
<td>2.53 (0.95)</td>
<td>2.42 (0.95)</td>
<td>−0.16 (−0.31 to −0.01)</td>
<td>2.58 (0.91)</td>
<td>2.39 (0.92)</td>
<td>−0.16 (−0.34 to 0.01)</td>
</tr>
<tr>
<td>Originality</td>
<td>2.29 (1.08)</td>
<td>−0.02 (−0.18 to 0.15)</td>
<td>2.31 (1.20)</td>
<td>2.20 (1.10)</td>
<td>−0.20 (−0.38 to −0.01)</td>
<td>2.40 (1.16)</td>
<td>2.17 (1.12)</td>
<td>−0.17 (−0.39 to 0.05)</td>
</tr>
<tr>
<td>Methodology</td>
<td>3.21 (0.98)</td>
<td>0.03 (−0.11 to 0.16)</td>
<td>3.18 (1.00)</td>
<td>3.07 (1.98)</td>
<td>−0.23 (−0.39 to −0.08)</td>
<td>3.31 (0.99)</td>
<td>3.09 (1.00)</td>
<td>−0.17 (−0.37 to 0.03)</td>
</tr>
<tr>
<td>Presentation</td>
<td>2.72 (1.01)</td>
<td>−0.12 (−0.26 to 0.02)</td>
<td>2.84 (0.98)</td>
<td>2.72 (0.98)</td>
<td>−0.11 (−0.27 to 0.06)</td>
<td>2.83 (1.10)</td>
<td>2.75 (0.95)</td>
<td>0.02 (−0.19 to 0.21)</td>
</tr>
<tr>
<td>Constructiveness of comments</td>
<td>3.30 (0.80)</td>
<td>−0.01 (−0.13 to 0.10)</td>
<td>3.31 (0.84)</td>
<td>3.26 (0.81)</td>
<td>−0.01 (−0.24 to 0.03)</td>
<td>3.36 (0.83)</td>
<td>3.23 (0.86)</td>
<td>−0.02 (−0.20 to 0.15)</td>
</tr>
<tr>
<td>Substantiation of comments</td>
<td>3.08 (0.93)</td>
<td>−0.02 (−0.14 to 0.10)</td>
<td>3.11 (0.90)</td>
<td>3.04 (0.93)</td>
<td>−0.10 (−0.24 to 0.05)</td>
<td>3.14 (0.92)</td>
<td>3.01 (0.93)</td>
<td>−0.03 (−0.17 to 0.22)</td>
</tr>
<tr>
<td>Interpretation of results</td>
<td>3.08 (0.97)</td>
<td>0.07 (−0.07 to 0.21)</td>
<td>3.01 (0.99)</td>
<td>3.02 (0.96)</td>
<td>−0.09 (−0.24 to 0.07)</td>
<td>3.10 (1.00)</td>
<td>2.94 (0.97)</td>
<td>−0.05 (−0.24 to 0.15)</td>
</tr>
<tr>
<td>Mean total score (range, 1-5)</td>
<td>2.87 (0.68)</td>
<td>−0.02 (−0.12 to 0.07)</td>
<td>2.90 (0.69)</td>
<td>2.82 (0.67)</td>
<td>−0.14 (−0.25 to −0.03)</td>
<td>2.96 (0.71)</td>
<td>2.79 (0.66)</td>
<td>−0.08 (−0.22 to 0.06)</td>
</tr>
<tr>
<td>Overall quality</td>
<td>3.26 (0.87)</td>
<td>0.02 (−0.11 to 0.14)</td>
<td>3.25 (0.84)</td>
<td>3.18 (0.86)</td>
<td>−0.16 (−0.29 to −0.02)</td>
<td>3.33 (0.86)</td>
<td>3.12 (0.86)</td>
<td>−0.10 (−0.28 to 0.08)</td>
</tr>
<tr>
<td>Review time h†</td>
<td>2.05 (1.51)</td>
<td>−0.13 (−0.37 to 0.11)</td>
<td>2.18 (1.64)</td>
<td>2.03 (1.52)</td>
<td>−0.13 (−0.37 to 0.11)</td>
<td>2.14 (1.58)</td>
<td>2.02 (1.40)</td>
<td>NA</td>
</tr>
</tbody>
</table>

*Each item was scored on a 5-point Likert scale, with 1 indicating poor and 5 indicating excellent. CI indicates confidence interval. NA, not applicable.

†Analysis of blinded vs unblinded groups includes all cases for which data were available for 2 reviewers (n = 282).

COMMENT

Blinding and unmasking have little effect on the quality of reviews of manuscripts. Any differences that have statistical significance are too small to be of any practical significance in editorial decision making. The only previous randomized trial reported higher-quality reviews when reviewers were blinded. This difference may have arisen either because the previous study was based on a more specialized journal, in which reviewers and authors would be more likely to be familiar with one another's work, or because of differences in the way review quality was assessed (the psychometric properties of the instrument used in the earlier study are unknown).

Before discussing the implications of these findings, potential methodologic shortcomings need to be considered. First, can the sample of manuscripts and their reviewers be considered truly random? There is no evidence of bias at any stage, although difficulties in finding suitable reviewers in the uninformd arm during the latter stages of recruiting due to the large number of reviewers already recruited to the intervention arms could have been one reason why we failed to find a Hawthorne effect. Of eligible manuscripts, 92% were recruited, and 89% of those were successfully followed up. The distribution of the manuscripts excluded and unavailable for follow-up was similar to that of those followed up.

Second, the results concerning review quality are completely dependent on the review quality instrument. This has been validated and has good internal consistency and interrater and intrarater reliability, and we believe it to be sufficiently accurate and robust to discriminate between reviews of differing quality for the purposes of this study. Full details of its development and validation will be reported elsewhere.

Third, the success rate for blinding is within the range found in previous studies. Although we were successful with only 58% of reviewers, analyses based on those actually blinded produced similar results to analyses based on the intention to blind.

Fourth, the views of authors, which have yet to be analyzed since the data are still incomplete, may differ from the views of editors. These will be reported in a subsequent article.
Effect on the Quality of Peer Review of Blinding Reviewers and Asking Them to Sign Their Reports

A Randomized Controlled Trial

Fiona Godlee, BSc, MRCP; Catharine R. Gale, BSc; Christopher N. Martyn, DPhil, FRCP

Context.—Anxiety about bias, lack of accountability, and poor quality of peer review has led to questions about the imbalance in anonymity between reviewers and authors.

Objective.—To evaluate the effect on the quality of peer review of blinding reviewers to the authors’ identities and requiring reviewers to sign their reports.

Design.—Randomized controlled trial.

Setting.—A general medical journal.

Participants.—A total of 420 reviewers from the journal’s database.

Intervention.—We modified a paper accepted for publication introducing 8 areas of weakness. Reviewers were randomly allocated to 5 groups. Groups 1 and 2 received manuscripts from which the authors’ names and affiliations had been removed, while groups 3 and 4 were aware of the authors’ identities. Groups 1 and 3 were asked to sign their reports, while groups 2 and 4 were asked to return their reports unsigned. The fifth group was sent the paper in the usual manner of the journal, with authors’ identities revealed and a request to comment anonymously.

Main Outcome Measure.—The number of weaknesses in the paper that were commented on by the reviewers.

Results.—Reports were received from 221 reviewers (53%). The mean number of weaknesses commented on was 2 (1.7, 2.1, 1.8, and 1.9 for groups 1, 2, 3, and 4 and 5 combined, respectively). There were no statistically significant differences between groups in their performance. Reviewers who were blinded to authors’ identities were less likely to recommend rejection than those who were aware of the authors’ identities (odds ratio, 0.5; 95% confidence interval, 0.3-1.0).

Conclusions.—Neither blinding reviewers to the authors and origin of the paper nor requiring them to sign their reports had any effect on rate of detection of errors. Such measures are unlikely to improve the quality of peer review reports.

JAMA. 1998;279:237-240

PEER REVIEW, as usually practiced by biomedical journals, protects the identity of reviewers but not of authors. Concerns about bias, lack of accountability, and poor quality of peer review have brought this practice into question. Two interventions, which together would reverse the balance of anonymity, have been suggested as possible solutions: removing authors’ identities from the manuscript (blinding) and asking reviewers to sign their reports (signing).

We performed a randomized controlled trial to examine the effect on peer review of blinding reviewers and asking them to sign their reports.

METHODS

With the authors’ consent, a paper already peer reviewed and accepted for publication by BMJ was altered to introduce 8 weaknesses in design, analysis, or interpretation. All reviewers whose spec...