Cite this article as: BMJ, doi:10.1136/bmj.38023.700775.AE (published 2 March 2004)

Papers

Effects of training on quality of peer review: randomised controlled trial

Sara Schroter, Nick Black, Stephen Evans, James Carpenter, Fiona Godlee, Richard Smith

Abstract

Objective To determine the effects of training on the quality of peer review.
Design Single blind randomised controlled trial with two intervention groups receiving different types of training plus a control group.
Setting and participants Reviewers at a general medical journal.
Interventions Attendance at a training workshop or reception of a self taught training package focusing on what editors want from reviewers and how to critically appraise randomised controlled trials.
Main outcome measures Quality of reviews of three manuscripts sent to reviewers at four to six monthly intervals, evaluated using the validated review quality instrument; number of deliberate major errors identified; time taken to review the manuscripts; proportion recommending rejection of the manuscripts.
Results Reviewers in the self taught group scored higher in review quality after training than did the control group (score 2.85 *v* 2.56; difference 0.29, 95% confidence interval 0.14 to 0.44; P = 0.001), but the difference was not of editorial significance and was not maintained in the long term. Both intervention groups identified significantly more major errors after training than did the control group (3.14 and 2.96 *v* 2.13; P < 0.001), and this remained significant after the reviewers' performance at baseline assessment was taken into account. The evidence for benefit of training was no longer apparent on further testing six months after the interventions. Training had no impact on the time taken to review the papers but was associated with an increased likelihood of recommending rejection (92% and 84% *v* 76%; P = 0.002).
Conclusions Short training packages have only a slight impact on the quality of peer review. The value of longer interventions needs to be assessed.

Introduction

Many studies have illustrated the inadequacies of peer review and its limitations in improving the quality of research papers.[1] However, few studies have evaluated interventions that try to improve peer review,[2] and no randomised controlled trials have examined the effects of training.[3] Training that would be feasible for reviewers to undergo and for a journal to provide would have to be short or provided at a distance. Although the effectiveness of short educational interventions is questionable, some brief interventions have been shown to be successful (depending on what is being taught and the methods used).[4 5]

We aimed to determine whether reviewers for the *BMJ* who underwent training would produce reviews of better quality than those who received no training; whether face to face training would be more beneficial than a self taught package; and whether any training effect would last at least six months.

Methods

Participants

We estimated that we needed 190 reviewers in each group to achieve a power of 0.9 to detect a difference in review quality between groups of 0.4 (one tenth of the maximum difference) on a scale of 1-5, with $\alpha = 0.05$ and standard deviation of difference = 1.2 on the review quality instrument.[6]

We randomised consenting reviewers into three groups: two intervention groups and a control group. We used a stratified permuted blocks randomisation method to ensure that the groups were similar in terms of factors known to influence the quality of reviews (age, current investigators in medical research projects, postgraduate training in epidemiology, postgraduate training in statistics, and editorial board members of a scientific or medical journal).[7 8]

Assessments and procedures

We selected three previously published papers, each describing a randomised controlled trial of alternative generic ways of organising and managing clinical work. We removed the names of the original authors and changed the titles of the manuscripts and any reference to study location (see bmj.com for test papers). We introduced 14 deliberate errors, classified as major (9) or minor (5) (see bmj.com for description). We asked all consenting reviewers to review the first paper. After this baseline assessment one intervention group received a full day of face to face training, and we mailed the other intervention group a self taught training package. Two to three months after the intervention we sent the second paper to reviewers who had completed the first review, and approximately six months later we sent the third paper to those who completed the second review.

We sent the manuscripts to the reviewers in a style similar to the standard *BMJ* review process, but we told them that these papers were part of the study, and we did not pay them. We asked reviewers to review the papers within three weeks and sent them the standard *BMJ* guidance for reviewers.

Outcome measures

Review quality—The review quality instrument version 3.2 is an eight item validated instrument (see bmj.com) developed spe-

P+ *Test papers, descriptions of deliberate errors, and review quality instrument are on bmj.com*

cifically for assessing the quality of reviews and has been used in several studies.[6 9–11] It includes items relating to the discussion of the methodological quality of the manuscript under review and items relating to the constructiveness and substantiation of the reviewers' comments. Two editors independently rated the quality of each review. We used the mean score of the items averaged over the two ratings.[6]

Number of deliberate major errors—Two researchers blind to the identity and study group of the reviewer independently assessed the number of major errors reported in each review. We used the total number of major errors identified averaged across the two raters. Inclusion of the minor errors in the total error score made no difference to the results.

Time taken and recommendation on publication—Reviewers recorded the time taken to review each paper and whether it should be published with no revision, published with minor revision, published with major revision, rejected, or other. Given the very poor quality of the papers, the most appropriate recommendation would have been rejection.

Interventions

Face to face training—The full day of training covered what *BMJ* editors require from reviewers and techniques of critical appraisal for randomised controlled trials. Participants were also given written instructions and a CD Rom, which included techniques on critical appraisal of randomised controlled trials, to use at home to help consolidate their learning.

Self taught training—We created a self taught training package based on the materials used in the training workshops, including the CD Rom. We asked reviewers to complete a questionnaire indicating the training exercises they had completed and to evaluate the training materials.

Statistical analysis

We examined differences between the groups in scores on the review quality instrument by using analysis of covariance, with baseline review score as a covariate, because scores followed a broadly normal distribution similar to our previous studies. We did an overall analysis comparing all three groups, and we report significant results only if the overall analysis was significant. Multiple comparison testing used a Bonferroni correction to set a significant cut-off. We give confidence intervals without correction for multiple testing. We used χ^2 to test for differences in proportions. Assessment of the impact of non-response used standard methods of multiple imputation that assume the data are missing at random.[12] We also investigated how much lower than the (observed) mean for responders the (unobserved) mean for non-responders would have to be, in order to remove any intervention effect.[13]

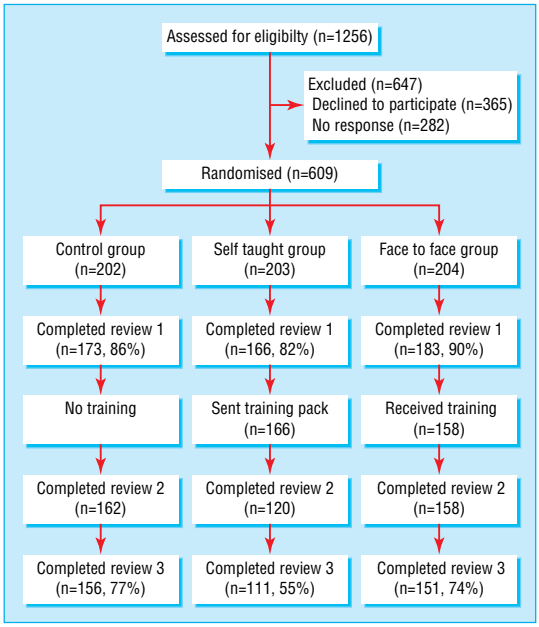
Results

Participants

Of 1256 reviewers assessed, 609 (48%) eligible reviewers agreed to take part (fig). The quality of the baseline reviews of those who did not complete follow up reviews was poorer than that of reviewers who did (review quality instrument score 2.60 *v* 2.73; P = 0.16), they detected fewer major errors (2.11 *v* 2.67; P = 0.01), and they recommended rejection less often (58% *v* 70%; P = 0.037) (table 1).

Evaluation of training interventions

One hundred and fifty eight reviewers attended training workshops, and 81% (114/141) anticipated that the quality of their reviews would improve. Most of the 120 recipients of the self taught package who completed review 2 reported having used the package (104 (87%) completed three of the five exercises, and 103 (86%) did all five), and 98 (82%) felt that the quality of their reviews would improve as a result.



Progress of participants through the trial

Outcome measures

Agreement was good between pairs of raters assessing the quality of reviews (intraclass correlation coefficient for total review quality instrument score 0.65) and the number of deliberate major errors identified (intraclass correlation coefficient 0.91).

Review quality instrument scores—The mean score for the whole sample was 2.71 (SD 0.73) for the first review and was similar across all three groups (table 2). For review 2, the self taught group scored significantly higher (2.85) than did the control group (2.56), a difference of 0.29 (95% confidence interval 0.14 to 0.44; P = 0.0002). The difference between the control group and the face to face group (2.72) was 0.16 (0.02 to 0.3; P = 0.025). We found no significant difference between any of the groups for the third review (P = 0.204), and the upper 95% confidence limit was at most 0.29. The participants in the control group who did a third review showed a small but significant rise in their score (0.17, 0.09 to 0.26; P = 0.0001), which reduced the difference between them and the intervention groups.

Errors identified—The number of errors detected in the baseline reviews was similar in the three groups (table 2). However, the difference between the control group and each of the intervention groups was significant for review 2 (2.13 *v* 3.14 and 2.96; self taught-control difference = 1.00, 0.65 to 1.37; face to face-control difference = 0.83, 0.47 to 1.19) and remained significant after adjustment of the scores for the number of errors reported in the first review (analysis of covariance, P < 0.0001). The differences observed for review 3 were slightly smaller but in a similar direction (2.71 *v* 3.37 and 3.18) and were not significantly different after adjustment for baseline and multiple testing.

Time taken to review and recommendation—Generally, the mean time taken to review papers did not differ significantly between the groups (table 2). All three groups spent less time doing the third review than the previous two reviews. The proportion of reviewers recommending rejection of paper 1 was similar across

Table 1 Comparison of characteristics and baseline scores for reviewers who completed the first review only and those who completed at least two reviews. Values are numbers (percentages) unless stated otherwise

	Review 1 only				At least two reviews			
	Total (n=82)	Control (n=11)	Self taught (n=46)	Face to face (n=25)	Total (n=440)	Control (n=162)	Self taught (n=120)	Face to face (n=158)
Characteristics								
Mean (SD) age in years	50.5 (8.1)	50.0 (10.6)	50.7 (6.5)	50.6 (9.7)	49.4 (8.4)	49.6 (8.3)	48.9 (9.2)	49.4 (7.7)
Male	59 (72)	7 (64)	32 (70)	20 (80)	320 (73)	118 (73)	88 (73)	114 (72)
Current investigator	68 (83)	8 (73)	38 (83)	22 (88)	364 (83)	132 (82)	102 (85)	130 (83)
Postgraduate training in epidemiology	27 (33)	6 (55)	8 (18)	13 (52)	166 (38)	63 (39)	51 (43)	52 (33)
Postgraduate training in statistics	39 (48)	7 (64)	17 (37)	15 (60)	241 (55)	89 (55)	69 (58)	83 (53)
Editorial board member	45 (55)	5 (46)	26 (57)	14 (56)	208 (47)	83 (51)	51 (43)	74 (47)
Baseline scores								
Mean (SD) RQI total score*	2.60 (0.7)	3.02 (0.5)	2.55 (0.8)	2.51 (0.7)	2.73 (0.7)	2.65 (0.8)	2.80 (0.6)	2.75 (0.7)
Mean (SD) No of major errors identified†	2.11 (1.6)	2.82 (2.0)	2.08 (1.5)	1.84 (1.5)	2.67 (1.9)	2.35 (2.0)	2.91 (1.8)	2.82 (1.9)
Mean (SD) time taken in minutes	138.8 (99.9)	143.2 (97.3)	137.1 (96.5)	139.8 (110.5)	136.0 (87.8)	129.1 (75.2)	141.4 (93.0)	138.8 (95.6)
Proportion recommending rejection	44/76 (58)	7/10 (70)	22/42 (52)	15/24 (63)	290/414 (70)	105/156 (67)	82/113 (73)	103/145 (71)

RQI=review quality instrument.
*Total scores range from 1 to 5; higher scores reflect higher review quality (average of two raters' scores).
†Number of nine major errors identified (average of two raters' scores).

the groups. The proportion recommending rejection of paper 2 was significantly lower for the control group than for the self taught group (76% *v* 92%; $P<0.0001$), and the same pattern occurred for paper 3 (74% *v* 91%; $P=0.001$).

Impact of non-responders

As the difference between responders and non-responders is unknown, the impact of non-response on the conclusions cannot be definitively determined. However, assuming that given the observed data from an individual reviewer his or her unseen response provides no additional information on the reason for non-response (the "missing at random" assumption), non-response has no effect on the statistical significance of the results. Alternatively, a more conservative approach is to assume that the mean for non-responders is shifted down from that of respond-

Table 2 Review quality, errors detected, time taken, and proportion recommending rejection (based on data from all participants). Values are means (SDs) unless stated otherwise

	Whole sample	Control group	Self taught group	Face to face group	P value for ANCOVA*	P value for χ^2
Review 1	(n=522)	(n=173)	(n=166)	(n=183)		
RQI total score†	2.71 (0.73)	2.67 (0.80)	2.73 (0.67)	2.72 (0.71)	—	—
No of major errors identified‡	2.58 (1.9)	2.38 (2.0)	2.68 (1.7)	2.68 (1.8)	—	—
Time (SD) (range) taken to review in minutes	136.4 (89.7) (25-720)	130.0 (76.6) (25-615)	140.2 (93.7) (30-720)	139.0 (97.5) (25-600)	—	—
Proportion (%) recommending rejection	334/490 (68)	112/166 (68)	104/155 (67)	118/169 (70)	—	—
Review 2	(n=440)	(n=162)	(n=120)	(n=158)		
RQI total score†	2.69 (0.65)	2.56 (0.64)	2.85 (0.64)	2.72 (0.63)	0.003§	—
No of major errors identified‡	2.71 (1.6)	2.13 (1.6)	3.14 (1.4)	2.96 (1.7)	<0.0001¶	—
Time (SD) (range) taken to review in minutes	130.9 (81.3) (10-720)	127.9 (76.5) (15-675)	144.4 (92.8) (20-720)	123.9 (76.3) (10-600)	0.024	—
Proportion (%) recommending rejection	346/417 (83)	114/151 (76)	104/113 (92)	128/153 (84)	—	0.002††
Review 3	(n=418)	(n=156)	(n=111)	(n=151)		
RQI total score†	2.79 (0.59)	2.74 (0.59)	2.89 (0.58)	2.76 (0.59)	0.204	—
No of major errors identified‡	3.05 (1.8)	2.71 (1.8)	3.37 (1.7)	3.18 (1.8)	0.125**	—
Time (SD) (range) taken to review in minutes	113.7 (63.8) (10-690)	108.5 (70.5) (30-690)	122.5 (65.8) (15-420)	112.7 (71.8) (10-600)	0.376	—
Proportion (%) recommending rejection	325/399 (82)	111/150 (74)	95/105 (91)	119/144 (83)	—	0.004‡‡

RQI= review quality instrument.
*Analysis of covariance (adjusting for baseline scores).
†Total scores range from 1 to 5; higher scores reflect higher review quality (average of two raters' scores).
‡Number of nine major errors identified (average of two raters' scores).
§Difference between control group and self taught group after Bonferroni correction for multiple comparisons.
¶Difference between control group and each intervention group after Bonferroni correction for multiple comparisons.
**Difference between control group and self taught group after Bonferroni correction for multiple comparisons.
††Difference between control group and self taught group P<0.0001 (two tailed Fisher's exact test).
‡‡Difference between control group and self taught group P=0.001 (two tailed Fisher's exact test).

ers, by the same amount in each intervention. Then, for the analysis of covariance comparison of review quality instrument scores between the self taught and control groups, we have to reduce the mean for the non-responders by 0.46 for the difference to become statistically insignificant.[14]

Discussion

This study has confirmed the limitations of peer review as witnessed by reviewers' failure to detect major methodological errors in three straightforward accounts of randomised controlled trials. Training led to some improvement in performance in terms of the detection of errors, the quality of the review, and the recommendations to the editor. With the exception of the recommendation, these improvements were slight and did not reach the a priori definition of editorial significance (review quality instrument score 0.4). The self taught package seemed to be more effective (and thus more cost effective for the journal) than the face to face training, although for the review quality instrument this result is only of borderline significance if non-responders are on average editorially significantly worse than responders. One possible reason for the differential response rate for the second review is that the non-responders in the self taught group had not used their training package and so did not respond to requests to review the second paper. The power of the study was sufficient to detect important differences had they existed.

The study was limited to peer reviewing of randomised controlled trials and cannot necessarily be extended to other study designs. The interventions, although involving actual transfer of knowledge through practical exercises, may have been inadequate to have a major effect. However, many of the participants valued the training and anticipated an improvement in their personal performance. Finally, the reviewers may not have been typical.

The validity of the data may have been affected in several ways. The artificiality of participating in a randomised controlled trial may have led reviewers to make less effort and underperform. Conversely, knowledge that performance was being scrutinised may have enhanced review quality. In addition, some reviewers may not have persisted in detecting all the errors after identifying enough to condemn a paper. All these potential influences are likely to have affected each of the three randomised groups of reviewers equally. Another potential threat to validity was that the papers were on topics outside the reviewers' area of expertise. We minimised this by limiting the papers to trials of medical records and communication activities that apply to all areas of health care.

Reviewers of the *BMJ* would not usually be sent papers of such poor quality. Some reviewers reported that they had became aware (through bibliographic searches) that the papers were derived from previously published articles and either assumed they must be good or realised they had been deliberately corrupted and therefore looked harder for errors. We believe that the overall effect of these various factors on the results was not sufficient to invalidate the conclusions.

The main implication of these results is that, as has been shown in areas outside the health sector, very short training has only a marginal impact. We cannot, therefore, recommend use of the intervention we studied. Similarly, previous studies have shown that voluntary attendance at a training session and written feedback by editors have no effect on quality of reviews.[4 5] In contrast, previous observational research has shown that extended training in epidemiology and statistics is associated with better reviewing.[8] This suggests that a simple, low cost educational approach to enhancing peer review may not be possible. However, an intermediate intervention (somewhere between a one day workshop and a one year postgraduate level training course) may be feasible for journals to provide, although this would need to be broad to reflect the wide range of methods and study designs that a journal needs to consider.

What is already known on this topic

Many studies have illustrated the inadequacies of peer review and its limitations in improving the quality of research articles

Although short educational interventions generally have limited effect, no major studies have been done in the field of peer reviewer training

What this study adds

Our short training package had only a slight impact on the quality of peer review in terms of quality of reviews and detection of deliberate major errors

The training did, however, influence reviewers' recommendations to editors

We thank all the reviewers who participated; the editors who assisted with the face to face training (Trish Groves, Sandy Goldbeck-Wood, Kamran Abbasi, Richard Smith, Fiona Godlee) and the Critical Appraisal Skills Programme (CASP) team; all the editors who rated the quality of reviews, especially Carole Mongin-Bulewski, Trish Groves, Rhona Macdonald, and Alison Tonks; Lyda Osorio for assessing the number of errors reported; and the authors of the original manuscripts for allowing us to use them.

Contributors: NB and RS initiated the study. SS, NB, RS, and FG designed the study. NB created the test papers. SS conducted the study; SS and SE did the data analysis; and SS, NB, and SE, interpreted the results. JC did the analysis of the impact of non-response. All authors assisted in writing the paper. SS, NB, and SE are guarantors for the paper.

Funding: This study was funded by the NHS London Regional Office Research and Development Directorate. The views and opinions expressed in this paper do not necessarily reflect those of NHSE (LRO) or the Department of Health.

Competing interests: RS is editor of the *BMJ*. SS, NB, and SE review for the *BMJ*.

Because members of *BMJ* staff were involved in the conduct of this research and writing the paper, assessment and peer review have been carried out entirely by external advisers. No member of *BMJ* staff has been involved in making the decision on the paper.

Ethical approval: The ethics committee of the London School of Hygiene and Tropical Medicine approved the study.

1 Rennie D. Editorial peer review: its development and rationale. In: Godlee F, Jefferson T, eds. *Peer review in health sciences*. London: BMJ Books, 1999.
2 Callaham ML, Knopp RK, Gallagher EJ. Effect of written feedback by editors on quality of reviews. *JAMA* 2002;287:2781-3.
3 Callaham ML, Wears RL, Waeckerle JF. Effect of attendance at a training session on peer reviewer quality and performance. *Ann Emerg Med* 1998;32:318-22.
4 Thomson O'Brien MA, Freemantle N, Oxman AD, Wolf F, David DA, Herrin J. Continuing education meetings and workshops: effects on professional practice and health care outcomes. *Cochrane Database Syst Rev* 2003;(4):CD003030.
5 Beaulieu M, Rivard M, Hudon E, Beaudoin C, Saucier D, Remondin M. Comparative trial of a short workshop designed to enhance appropriate use of screening tests by family physicians. *Can Med Assoc J* 2002;167:1241-6.
6 Van Rooyen S, Black N, Godlee F. Development of the review quality instrument (RQI) for assessing peer reviews of manuscripts. *J Clin Epidemiol* 1999;52:625-9.
7 Evans AT, McNutt RA, Fletcher SW, Fletcher RH. The characteristics of peer reviewers who produce good-quality reviews. *J Gen Intern Med* 1993;8:422-8.
8 Black N, van Rooyen S, Godlee F, Smith R, Evans S. What makes a good reviewer and a good review in a general medical journal. *JAMA* 1998;280:231-3.
9 Van Rooyen S, Godlee F, Smith R, Evans S, Black N. The effect of blinding and unmasking on the quality of peer review: a randomized trial. *JAMA* 1998;280:234-7.
10 Van Rooyen S, Godlee F, Evans S, Black N, Smith R. Effect of open peer review on quality of reviews and on reviewers' recommendations: a randomised trial. *BMJ* 1999;318:23-7.

11 Walsh E, Rooney M, Appleby L, Wilkinson G. Open peer review: a randomised controlled trial. *Br J Psychiatry* 2000;176:47-51.
12 Schafer JL. Multiple imputation: a primer. *Stat Methods Med Res* 1999;8:3-15.
13 Simon R. Bayesian subset analysis: applications to studying treatment-by-gender interactions. *Stat Med* 2002;21:2909-16.
14 White I, Carpenter J, Evans S, Schroter S. Eliciting and using expert opinions about non-response bias in randomised controlled trials. Technical report: email James.Carpenter@lshtm.ac.uk

(Accepted 30 November 2003)

doi 10.1136/bmj.38023.700775.AE

BMJ Editorial Office, BMA House, Tavistock Square, London WC1H 9JR
Sara Schroter *senior researcher*
Fiona Godlee *head of BMJ knowledge*
Richard Smith *editor*

London School of Hygiene and Tropical Medicine, London WC1E 7HT
Nick Black *professor of health services research*
Stephen Evans *professor of pharmacoepidemiology*
James Carpenter *senior lecturer in medical statistics*

Correspondence to: S Schroter sschroter@bmj.com