ORIGINAL ARTICLE

A randomized controlled study of reviewer bias against an unconventional therapy

K I Resch MD PhD E Ernst MD PhD[1] J Garrow MD

J R Soc Med 2000;**93**:164–167

SUMMARY

A study was designed to test the hypothesis that experts who review papers for publication are prejudiced against an unconventional form of therapy. Two versions were produced (A and B) of a 'short report' that related to treatments of obesity, identical except for the nature of the intervention. Version A related to an orthodox treatment, version B to an unconventional treatment. 398 reviewers were randomized to receive one or the other version for peer review. The primary outcomes were the reviewers' rating of 'importance' on a scale of 1–5 and their verdict regarding rejection or acceptance of the paper. Reviewers were unaware that they were taking part in a study.

The overall response rate was 41.7%, and 141 assessment forms were suitable for statistical evaluation. After dichotomization of the rating scale, a significant difference in favour of the orthodox version with an odds ratio of 3.01 (95% confidence interval, 1.03 to 8.25), was found. This observation mirrored that of the visual analogue scale for which the respective medians and interquartile ranges were 67% (51% to 78.5%) for version A and 57% (29.7% to 72.6%) for version B.

Reviewers showed a wide range of responses to both versions of the paper, with a significant bias in favour of the orthodox version. Authors of technically good unconventional papers may therefore be at a disadvantage in the peer review process. Yet the effect is probably too small to preclude publication of their work in peer-reviewed orthodox journals.

INTRODUCTION

The processes by which medical journals evaluate scientific papers have come under critical examination and some of the proposed reforms can be expected to increase the internal validity of the review process. However, one potentially important source of bias, reviewers' preconceptions[2], has been little investigated. Reviewers may, for instance, be biased in favour of submissions from their own country of origin[3,4]. A language bias seems to exist, with statistically significant results more likely to gain a place in English-language journals[5]. Commonly used evaluation instruments have been shown to be unreliable[6]. Creative ideas may be penalized in grant applications[7].

Much thought went into ameliorating the peer review process[8,9]. Blinding of reviewers to authorship, or having them sign their comments, does not seem to improve the quality of their reports[10]. Lack of open-mindedness in the peer review process could affect the introduction of unconventional concepts into medicine. We conducted a randomized, controlled, double-blind trial to test the hypothesis that peer review favours an orthodox form of treatment over an unconventional therapy.

METHODS

The protocol was developed jointly with members of LOCKNET, an international network for research into peer review organized by the *British Medical Journal*. To test the hypothesis, we prepared two versions ('orthodox'=version A *versus* 'unconventional'=version B) of an invented 'short report' describing a randomized, placebo-controlled, trial of appetite suppressants. These were sent for peer review to experts on obesity.

The choice of the 'orthodox' drug, *hydroxycitrate*, was based on a Medline search (1966–1996) and expert consultations. Homoeopathic *sulphur* was identified as an adequate remedy for the purpose of version B, from information in Kent's repertory[11] (a classic homoeopathic reference book) and the advice of two experienced homoeopaths.

Forschungsinstitut für Balneologie und Kurortwissenschaft, Lindenstrasse 5, 08645 Bad Elster, Germany; [1]Department of Complementary Medicine, School of Postgraduate Medicine & Health Sciences, University of Exeter, UK; [2]*European Journal of Clinical Nutrition*, Herts, UK

Correspondence to: Professor E Ernst, Department of Complementary Medicine, School of Postgraduate Medicine and Health Sciences, University of Exeter, 25 Victoria Park Road, Exeter EX2 4NT, UK
E-mail: E.Ernst@exeter.ac.uk

The two versions of our 'short report' were identical except for the first few lines citing a key reference of the respective orthodox or unconventional drug. The remainder of the text was phrased such that the name of either drug could be used interchangeably. Both versions had design features necessary to make the report methodologically acceptable (e.g. randomization, blinding, placebo controls, run-in period, compliance test, approval by an ethics committee). Both were fabrications of trials that were, in fact, never conducted; copies are obtainable on request from EE.

Vienna, Austria, was chosen as the origin of the fictitious authors, because two of us (KLR, EE) were familiar with the local situation. To avoid difficulties, we 'invented' a *Ludwig Boltzmann Institute for Metabolic and Nutritional Diseases* (numerous Boltzmann institutes exist in Austria, so a reviewer would be highly unlikely to know, for sure, that such an institution did not exist). 'Real' authors' names relating to Vienna were chosen by identifying, through Medline, at least two individuals with the same name in Vienna; in our 'short report' the initials after the first name were switched.

We scanned Medline, January 1993 to June 1996, for articles dealing with treatment of obesity (strategy: APPETITE DEPRESSANTS or DIET, REDUCING or OBESITY/DH or OBESITY/DT or OBESITY/PC or OBESITY/SU or OBESITY/TH) and retrieved 1137 records. All suitable addresses for first authors were collected. Researchers who had previously reviewed manuscripts for the *European Journal of Clinical Nutrition* (*EJCN*) were removed from the list. If more than one paper from one institution was identified (whether from the same or different authors), only the latest was retained. Thus 396 addresses were obtained, sorted by country and randomized into two groups (block randomization with blocks of four).

The evaluation sheets of the *EJCN* were used for peer review assessment. They consist of dichotomous questions on 8 items (title, summary, methods, results, discussion, references, reliability, and ethics), where 'yes' indicates a positive and 'no' a negative vote, and a summarizing question on 'importance' (1=trivial even if true, 5=major contribution to knowledge in the field). For the purpose of our study, it was complemented by a visual analogue scale (VAS) relating to a recommendation to accept or reject the paper. On this VAS, 0=reject outright, 100=must accept. This and the intergroup differences(s) in rating of importance were predefined as the primary outcome measures. The proportions of reviewers voting against or in favour of the above 8 items were used as explanatory variables. The review sheet is obtainable from EE on request.

Randomization was performed in Germany (by KLR), and two address lists were sent to Exeter, UK, where the letters were prepared for mailing (by EE) on *EJCN* stationery (provided by JG). Letters for referees of version A contained an evaluation sheet with the title underlined, whereas for version B the title was not underlined; this allowed reliable identification of group allocation of otherwise anonymous evaluation sheets but did not allow the retrospective identification of those referees who had not responded. The mailing was done by JG who also received the responses. Returned evaluation sheets and comments were checked for hints on group allocation (by JG and EE). If present, such hints were blacked out. Subsequently, all evaluation sheets were sent to KLR for evaluation, where they were checked a third time by a person not previously involved in the study. A debriefing letter was sent by EE to all recipients of the original mailing, summarizing the main goals of the study and the main findings.

A power calculation had revealed that a sample size of 72 per group would be required to identify a two-tailed difference of 10% (SD 15%) at the 5% level between groups on a VAS. A non-parametric test for independent samples (Mann–Whitney U-test) was chosen to test the hypothesis (intergroup difference on VAS), and a χ^2 analysis to analyse the rating of importance.

RESULTS

166 responses were received (response rate 41.7%), 141 of which were suitable for evaluation (35.4% of total). 79 responses were related to version A and 62 to version B ($P > 0.2$, χ^2-test).

Recommendations to accept or reject the paper covered the entire range of the VAS (Figure 1). Medians were 67% (interquartile range [IQR] 51% to 78.5%) for version A and 57% (29.7% to 72.6%) for version B (P=0.052).

The *rating of importance* was condensed to 1/2 ='negative', 3='undecided', 4/5='positive'. These ratings significantly favoured the acceptance of version A (P=0.05, χ^2-test). This tendency was even more pronounced when only 'positive' and 'negative' ratings were considered (P=0.028, χ^2) (Table 1), resulting in an odds ratio of 3.01 (95% CI 1.025 to 8.25) in favour of the orthodox treatment.

Table 1 **Ratings on the importance of the paper**

Rating	***Orthodox (valid information n=78)***	***Unconventional (valid information n=55)***
Negative (1/2)	11	13
Undecided (3)	33	29
Positive (4/5)	34	13

*1=trivial even if true, 5=major contribution to knowledge
P=0.05 for 2 × 3 table, P=0.028 for 2 × 2 table of positive *vs* negative ratings

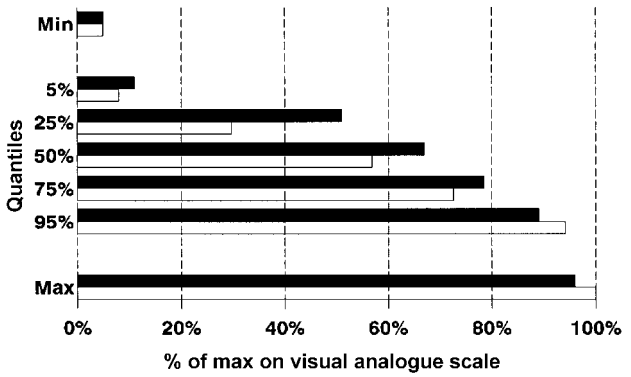

Figure 1 **Responses from reviewers to two different versions of a short report. □ Orthodox version (*n*=79); ■ unconventional version (*n*=62)**

For dichotomous questions (explanatory variables), odds ratios were calculated to compare answers. There were no significant differences between versions A and B (Table 2), the pooled odds ratio being 1.04 (95% CI 0.79 to 1.38).

DISCUSSION

These results suggest that, in this particular setting, a bias against publication of unconventional therapies exists. They support the conclusion of others[12] that studies incongruent with *a priori* beliefs tend to be rated by outside reviewers as incompetently conducted.

In designing this study, much effort was put into minimizing bias and ethical concerns. To avoid a bias due to a difference in credibility of the therapies involved, we compared a 'questionable' orthodox drug with a questionable unconventional drug. Identification of reviewers through Medline searches followed widespread practice and minimized selection bias. At the planning stage there was much discussion of whether or not we needed formal approval from an ethics committee. After weighing the pros and cons, the consensus was that we did not; however, the initial protocol was changed in several respects to lessen ethical concerns. One major worry was that our 'short report' might generate misinformation. We therefore decided to debrief all referees by means of a short note explaining the main goal of the study. None of the referees responded adversely, or otherwise, to this debriefing.

Table 2 **Odds ratios of positive and negative statements on 8 items of the peer review evaluation sheet, ORs above 1 favouring the orthodox version**

Item	*Odds ratio*	*95% confidence interval*
Title	0.741	0.352 to 1.596
Summary	1.399	0.519 to 3.800
Methods	0.757	0.372 to 1.558
Results	1.004	0.484 to 2.089
Discussion	1.237	0.597 to 2.545
References	1.160	0.531 to 2.539
Reliability	1.750	0.729 to 4.113
Ethics	0.871	0.324 to 2.441
Pooled ratings	1.044	0.790 to 1.381

Clearly, informed consent by reviewers would have invalidated the study. Others have stressed that informed consent may not be mandatory for investigating the peer review process[13]. To maximize the power of the study, preference was given to a continuous variable over a categorical one, and a categorical one over a dichotomous one (an 80% chance of identifying a two-tailed difference of 10% [SD 15%] at the 5% level would require a sample size of 72, and >1000 for a dichotomous variable). General agreement was reached during the planning period that primary outcome measures would be desirable. A structured evaluation sheet was therefore deemed superior to an unstructured one.

The response rate was disappointing but resembles that in similar investigations[14–16]. Rates would have been higher if we had been able to send a reminder, but this would have jeopardized confidentiality and thus increased our ethical dilemma. A low response rate does not necessarily affect the validity of the data collected[17].

Although the difference in ratings on the VAS (one of the main outcome measures) was just short of statistical significance (P=0.052), the medians showed a potentially relevant and meaningful difference of 10 percentage points (which could be expressed as an 18% better rating for version A). Interquartile ranges also indicate that more reviewers gave version B a poor rating. When ratings of importance (our other primary endpoint) were taken into account, version B was significantly less favoured by reviewers. Interestingly, none of the 8 items that dealt with defined aspects of the 'short reports' reflected those tendencies. This suggests that reviewers' verdicts were related less to definable aspects of version B than to the fact that it was not conventional, mainstream or plausible.

In Figure 2 the VAS ratings are plotted against the rating of importance. The latter variable was not a reliable predictor of the reviewers' recommendation regarding acceptance or rejection. Editors who use instruments for peer review that do not explicitly ask for a suggestion might therefore interpret 'surrogate endpoints' incorrectly. This aspect seems worthy of further investigation.

If we accept the existence of reviewer bias against the unconventional, how might it be minimized? Of several recent suggestions for improvements of the peer review

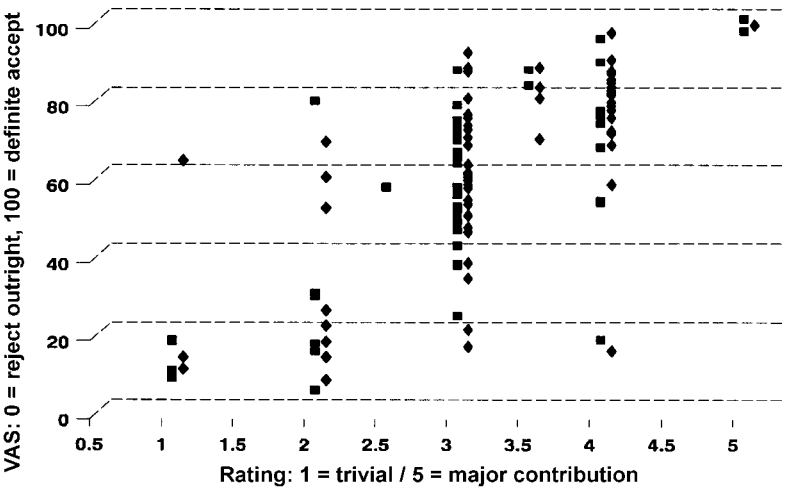

Figure 2 **Rating of importance for two different versions of a short report. ■ Unconventional version ; ◆ orthodox version**

process[8,9,18] none would affect the type of bias discussed here. Perhaps the best way of dealing with it is simply to make editors aware of its existence. They can then use their common sense to counterbalance it.

We conclude that, in this particular setting, a reviewer bias against publication of unconventional therapies exists. This hypothesis should now be retested in other settings and by independent investigators. The bias that we detected may put authors of unconventional papers at a disadvantage—but not, we think, a large enough disadvantage to preclude publication in peer-reviewed orthodox journals. Thus it does not explain the scarcity of methodologically sound papers on unconventional treatments in peer reviewed journals.

Acknowledgment We thank the LOCKNET group for help in the planning of this study.

REFERENCES

1 Begg C, Cho M, Eastwood S, *et al*. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *JAMA* 1996;**276**:637–9

2 Ernst E, Resch KL, Uher EM. Reviewer bias. *Ann Intern Med* 1992;**116**:958

3 Link AM. US and non-US submissions. An analysis of reviewer bias. *JAMA* 1998;**280**:246–1

4 Nieminen P, Isohanni M. Bias against European journals in medical publication databases. *Lancet* 1999;**353**:1592

5 Egger M, Zellweger-Zähner T, Schneider M, *et al*. Language bias in randomised controlled trials published in English and German. *Lancet* 1997;**350**:326–9

6 Ernst E, Saradeth T, Resch KL. Drawbacks of peer review. *Nature* 1993;**363**:296

7 Durso TW. Researchers disagree on NIH plan to improve its peer-review process. *Scientist* 9 Dec 1996

8 Davidoff F. Masking, blinding, and peer review: the blind leading the blinded. *Ann Intern Med* 1998;**128**:66–8

9 Goldbeck-Wood S. Evidence on peer review—scientific quality control or smokescreen? *BMJ* 1999;**318**:44–5

10 Godlee F, Gale CR, Martyn CN. Effect on the quality of peer review of blinding reviewers and asking them to sign their reports. *JAMA* 1998;**280**:237–40

11 Kent JT. *Repertorium der homöopathischen Arzneimittellehre*. Stuttgart: Hippokrates Verlag, 1981

12 Roe CA. Critical thinking and belief in the paranormal: a re-evaluation. *Br J Psychol* 1999;**90**:85–98

13 Feinstein AR. Construction, consent, and condemnation in research of peer review. *J Clin Epidemiol* 1991;**44**:339–41

14 Deehan A, Templeton L, Taylor C, Drummond C, Strang J. The effect of cash and other financial inducements on the response rate of general practitioners in a national postal study. *Br J Gen Pract* 1997;**47**:87–90

15 Atkinson M, el Guebaly N. Research productivity among PhD faculty members and affiliates responding to the Canadian Association of Professors of Psychiatry and Canadian Psychiatric Association survey. *Can J Psychiatry* 1996;**41**:509–12

16 Sanderson M. Survey of candidates taking the MRCP(UK) Part 2 examination. *J R Coll Physicians Lond* 1996;**30**:523–6

17 Templeton L, Deehan A, Taylor C, Drummond C, Strang J. Surveying general practitioners: does a low response rate matter? *Br J Gen Pract* 1997;**47**:91–4

18 Wessely S. Peer review of grant applications: what do we know? *Lancet* 1998;**352**:301–5