Neuroendocrinology Letters 1999; **20**:256–260
pii: NEL200599E03

Guest Editorials

Reprinted with permission from the British Journal of Psychiatry 1998; **173**:110–113

Peer review and editorial decision-making

Louise Howard and Greg Wilkinson

Introduction. This paper describes and analyses the editor's decision-making process at the *British Journal of Psychiatry* (BJP), and investigates the association between reviewers' assessments and editorial decisions.
Method. Four hundred consecutive manuscripts submitted over a six-month period to the BJP were examined prospectively for assessors' comments and editorial decisions on acceptance or rejection. Interrater reliability of assessments was calculated and a logistic regression analysis investigated the effect of the rank allocated by'assessors and the comprehensiveness of the assessments on the editor's decision.
Results. The editor sent 248/400 (62%) manuscripts to assessors for peer review. Kappa for reliability of assessors' rankings was 0.1 indicating poor interrater reliability. Assessors agreed best on whether to reject a paper. A ranking of five (indicating rejection) had the greatest association with editor's rejection ($P < 0.001$, odds ratio 0.079), and the mean ranking of assessments was also significantly associated with editorial acceptance or rejection ($P=0.004$, odds ratio 0.24).
Conclusion. Assessors and editors tend to agree on what is clearly not acceptable for publication but there is less agreement on what is suitable.
Declaration of interest. The second author is Editor of the BJP.

There is little published on how editorial decisions are made, despite increasing scientific interest in the peer review process (Smith, 1997). Editors' requests of assessors vary substantially (Frank, 1996) and editorial peer review practices differ between journals (Weller, 1990). While some journals give readers detailed information on the editorial process (Smith, 1990) this is the exception rather than the rule. Editorial decisions may appear shrouded in mystery. This article aims to describe and analyse the editorial process at the *British Journal of Psychiatry* (BJP) in an attempt to make the process more open and available for scrutiny.

There are many factors which will influence the editor, including the number of manuscripts submitted and the nature of the journal in question. Editors may need to override the recommendations of assessors and it has been suggested that editors can be 'super reviewers' in view of their exposure to papers submitted (Crandall, 1991). It should also be recognised that editors decide which assessors to ask for assessments and this decision itself might be biased (Tyrer, 1991), though there are no objective criteria for this process. We aimed to investigate how assessors' recommendations influenced editorial decisions and whether the comprehensiveness of the assessment made by the assessor also had an influence on outcome.

METHOD

Editorial decisions and assessors' comments were examined prospectively on 400 consecutive manuscripts submitted to the BJP between February 1996 and September 1996. The editor indicated his reasons for rejecting papers without assessment on a questionnaire which listed possible reasons for rejection. Assessors were rated for the presence or absence of comments requested by the covering letter sent to all assessors.

The covering letter asked reviewers to comment on whether the study was reasonable or flawed, whether it was readable and logical, to comment on the introduction and discussion, address length, state whether all sections should be included and discuss layout. Items were rated and totalled, together with a point for giving a score to the paper (see below), giving a total 'comprehensiveness of assessment' score (the maximum possible score was eight).

Rankings

The covering letter to assessors also asked for a score to be allocated to the paper (one=should be published without amendment, two=a useful piece of work of high rank, three=as 'two' but of lower rank, four=not suitable for publication in its present form but would be likely to fall into categories 'one' or 'two' after revision, five=to be rejected, six=does not fall into these categories). These rankings were recorded and the researcher (L.H.) scored manuscripts using assessors' covering comments where the assessor had not given a rank to a manuscript but had stated an opinion on whether it should be published.

Statistical analysis

SPSS for Windows 6.1 was used for statistical analysis. Kappa coefficients were calculated to assess interrater reliability of assessors. A Pearson correlation coefficient was calculated to investigate the correlation between rankings and comprehensiveness. A logistic regression (forward stepwise) analysis investigated the effect of rankings allocated to papers by reviewers and the comprehensiveness of the assessment on the editor's decision. The outcome measure (dependent variable) was acceptance for publication by the editor.

Hypotheses

We hypothesised that assessors would show most agreement on which papers should be rejected, and that the editor would be more likely to agree with an assessor's recommendation of rejection and reject the paper, than agree with other assessments of a paper. We also hypothesised that the editor might be most influenced by the mean of two assessors' scores, or by a score of one or five by either assessor. We investigated whether the comprehensiveness of assessment influenced the editor's decisions by analysing the impact of the mean, best and worst comprehensiveness scores on acceptance.

RESULTS

Initial decisions

The editor sent 248 of 400 (62%) manuscripts to assessors for peer review. Three editorials were accepted without assessment and the remaining 149 manuscripts (37%) were rejected without being sent for external assessment. The most common main reason for rejection was that the paper was too specialised (n=59), 32 papers were considered to be unoriginal, 31 were poor methodologically, 11 were rejected because of their subject matter, 10 were case reports and six were written in an inappropriate format.

Decisions after external assessment

Two or more assessments were carried out on 191 of the 248 (77%) manuscripts sent for assessment. One assessment only was returned in 57 (33%) cases, usually because the second assessor was unable to find time to assess the paper. A rank (1–6) was given to the paper by 335 assessors out of a total of 439 assessments. Where rankings were omitted accompanying comments were usually clear indications of the assessors' views. Numerical rankings were allocated by L.H. for 103 manuscripts where assessors' comments only were provided (e.g. "this should be condensed into a short paper" was given a ranking of four; "certainly deserves publication" given a ranking of one; "does not reach the standard for publication" given a five. It was not possible to estimate rankings for three assessments. The rankings given for all 436 assessments are shown in Table 1.

One hundred and thirty-three of the 248 (54%) manuscripts sent for assessment were subsequently rejected. Six (2%) papers were accepted with no further revision requested and 108 (44%) were accepted after revision. No papers were rejected after being revised. Six papers had been sent for further assessment after the initial peer review. One manuscript was withdrawn.

Comprehensiveness of peer review

The mean comprehensiveness of assessment score was four, range 0–8 (maximum eight), mode four, s.d. 1.54. Comprehensiveness was not significantly correlated with rankings (r=0.0411, P=0.575).

Interrater reliability of assessors

The overall kappa for reliability of assessors' rankings (i.e. whether it should be accepted or rejected using categories 1–6) was 0.1. The reliability for rank one v. others was 0. 11, ranks one or two v. three, four or five was 0.16, one, two or three v. four or five was 0.19, and for 1-4 v. five was 0.27. (Kappa values <0.2 indicate poor agreement and kappa values of 0.21–0.40 indicate fair agreement (Altman, 1991).) There was therefore an apparent

trend for assessors to agree best on whether to reject a paper (see Table 1).

Table 1. Assessors' rankings

		Rankings by assessor one					
		1	2	3	4	5	Total
Rankings by assessor two	1	3	3	1	7	3	17
	2	6	4	4	12	5	31
	3	1	4	3	4	8	20
	4	4	11	5	18	22	60
	5	1	4	4	16	35	60
No second assessment		5	10	2	19	20	56
Total		**20**	**36**	**19**	**76**	**93**	

Effect of peer review on editorial decision

Rankings.
A logistic regression model analysed the effect on editorial decisions of the mean rank given by two reviewers, rankings of five, and rankings of one. A rank of five had the greatest influence on the editor's decision to reject a manuscript ($P<0.001$, $B=-2.5359$, odds ratio=0.079, 95% CI 0.023–0.273). The mean rank also had a very significant influence on the editor's decision with the editor most likely to reject manuscripts with a lower mean score ($P=0.004$, $B=-1.4205$, odds ratio=0.24, 95% CI 0.093–0.631). A best score of one did not have a significant influence after adjusting for mean score and a worst score of five ($P=0.2$).

Comprehensiveness.
A separate logistic regression analysis was carried out to analyse the effect of comprehensiveness on editorial decisions. The highest and lowest of the two comprehensiveness scores for each pair of assessments did not have a significant influence on decisions but the mean comprehensiveness score of pairs of assessments showed a significant effect ($P=0.027$, $B=0.214$, odds ratio=1.239, 95% CI 1.049–1.429). This latter variable was therefore entered into the first logistic regression model above but when the model was adjusted for rankings of five and mean comprehensiveness score, mean comprehensiveness no longer had a significant effect ($P=0.76$).

Summary of logistic regression model.
Variables significantly associated with editor's acceptance of an article for publication therefore remained the mean rank and rankings of five from one or both assessors.

DISCUSSION

Peer review is a time-consuming task, with assessors usually working for more than one journal, reviewing approximately 12 manuscripts per year and spending one to two hours on each (Lock & Smith, 1990) with little recognition of their work. The assessors studied here have produced impressive reviews of manuscripts which help the editor greatly in assessing submitted papers. However, the assessment process itself should be subject to review and scientifically evaluated, and while there has been increasing scientific interest in peer review in recent years (Lock, 1985; Wessely, 1996) this has usually focused on the assessments rather than the whole editorial decision-making process.

Several methods of assessing peer review have been developed (e.g. McNutt et al, 1990; Feurer et al, 1994). The checklists used by researchers usually try to measure the quality of the review but this involves further subjective assessment by the researcher. This study focused on whether the assessor complied with the requests made by the covering letter from the journal (i.e. comprehensiveness), the interrater reliability for the overall recommendation on acceptance or rejection and the effect of these on editorial decisions as measured by a logistic regression model.

Main findings

Comprehensiveness.
The assessors did not attain high comprehensiveness scores, as the mean comprehensiveness score of the assessment was 4/8.
However, the assessors may consider that not all the requests made of them are relevant in all cases; for example, when the paper is outstanding and a high ranking for acceptance is given.

Interrater reliability.
Kappa coefficients indicated that assessors interrater reliability is low ranging from 0.1–0.3. Raters appear to agree on whether to reject a paper, but opinions on the value of other papers are more mixed. This is in line with other research in the area—most studies find assessors' interrater reliabilities to be between 0.2 and 0.4 and that assessors show greater agreement when recommending rejection than when recommending acceptance (Cicchetti, 1991). However, some authors have argued that editors do not need reliable comments from

assessors but rather need a variety of different perspectives to help them make editorial decisions (Kiesler, 1991). It is difficult to be certain whether an increase in the number of assessments would make the editor's job easier or more difficult.

Factors associated with editorial decisions. When the editor received the assessors' rankings he appeared to be most influenced by a clear rejection (five) from either assessor. The editor is also affected by the average score of the two assessors but is less influenced by a clear recommendation for acceptance (one) or by the comprehensiveness of the assessment. This probably reflects scepticism of solitary high rankings. The rankings appear more important in influencing the editor than the comprehensiveness of the assessment, but a comprehensive assessment can be very useful for the authors of the manuscript.

Limitations

The omission of rankings by nearly one quarter of assessors meant that we decided to estimate these rankings using raters' comments. This may have biased the data but it is unlikely to have led to differential misclassification. However, some assessors may not give an indication as to the suitability of a paper for publication because they do not feel this is the function of the referee, despite specific requests for rankings from journals.

Another problem also became clear from these covering comments—it appears that different assessors attach different criteria to the scoring system used despite the guidelines given by the BJP. For example, 18 assessors gave a rank of four (i.e. the paper needs revision) while stating that the paper was excellent and needed minor changes only, while three assessors used four to mean that a paper needs extensive revision. This difference in interpretation has led to changes in the criteria now given to assessors for ranking manuscripts for the BJP. The main analyses here have used the rankings only and may therefore not have fully represented the effect of the assessors' comments on editorial decisions.

Assessments of assessor reliability, comprehensiveness and their influence on editorial decisions may be less important than assessments of the validity of the peer review process (Bornstein, 1991). Validity is difficult to measure; a paper's impact factor is one measure of how useful the work has been to other researchers but impact factors have a number of limitations (Howard & Wilkinson, 1997; Seglen, 1997; Smith, 1998). However, there is evidence that the peer review process leads to significant improvements in articles, at least as evaluated by readers in a blind assessment of papers before and after peer review and editing (Pierie *et al,* 1996). Early work found that many papers rejected by one journal are published in other equally prestigious journals (Wilson, 1978); the *British Medical journal* reported a 68% publication rate for manuscripts in the 1970s (Lock, 1985). However, recent evidence demonstrates that nearly two-thirds of manuscripts rejected by peer review are not published in other indexed medical journals (Abby et al, 1994). This may be an indication that the peer review process has improved.

Publication of the peer review process in journals and on the Internet may facilitate further improvements by making peer review more open to the reader (Smith, 1997). Editors should make explicit the judgements that are required of the assessor and ensure that, when appropriate, authors have the opportunity to respond to an assessor's comments (Persaud, 1995). The success or failure of the peer review process ultimately will be judged by the readers of the Journal.

The Editor's decision is final…

Because of space constraints less than a quarter of the papers submitted to the BJP are accepted for publication. In such circumstances, the editor needs to choose which papers to accept, and a variety of factors influence his decisions. Fairness, openness, accountability and transparency compete with hubris and human error. Constructive proposals for improvement can contribute to the debate arising from the findings presented here.

Acknowledgements

We thank the two referees for their helpful comments; Dr Pak Sham for statistical advice; Zofia Ashmore, BJP, for administrative support; and Dr. Vimal Sharma for commenting on an earlier draft.

CLINICAL IMPLICATIONS

- Assessors agree best on whether to reject papers and recommendations of rejection from assessors are strongly associated with rejection by the editor.
- Assessors have a low interrater reliability and the editor therefore plays a very significant role in the decision-making process.
- Further research is needed to investigate the validity of the editorial decision-making process, though this is limited as it is not possible to use a 'gold standard'.

LIMITATIONS

- Estimation of missing rankings may have led to bias, but this is unlikely to have led to differential misclassification.
- Assessors had different interpretations of the criteria for rankings which may have led to bias.
- Quality of assessments was not measured and this is the subject of ongoing research at the British Journal of Psychiatry.

LOUISE HOWARD, MRCPsych. Institute of Psychiatry, De Crespigny Park, London; GREG WILKINSON, FRCPsych, University Department of Psychiatry, Royal Liverpool University Hospital, Liverpool.
Correspondence: Dr Louise Howard, Institute of Psychiatry, De Crespigny Park, London SE5 8AF
(First received 20 January 1998, final revision 9 April 1998, accepted 21 April 1998)

REFERENCES

Abby, M., Massey, M. D., Gallanduk, S., et al (1994) Peer review is an effective screening process to evaluate medical manuscripts. *Journal of the American Medical Association,* **272**, 105–107.
Altman, D. G. (1991) *Practical Statistics for Medical Research*. London: Chapman and Hall.
Bornstein, R. F. (1991) The predictive validity of peer review: a neglected issue. *Behavioral and Brain Sciences,* **14**, 138–139.
Cicchetti, D. V. (1991) The reliability of peer review for manuscript and grant submissions: a cross disciplinary investigation. *Behavioral and Brain Sciences,* **14**, 119–186.
Crandall, R. (1991) What should be done to improve reviewing? *Behavioral and Brain Sciences*. **14**, 143.
Feurer, I. D., Becker, G. J., Picus, D., et al (1994) Evaluating peer reviews. *Journal of the American Medical Association,* **272**, 98–100.
Frank, E. (1996) Editors' requests of peer reviewers: a study and a proposal. *Preventive Medicine,* **25**, 102–104.
Howard, L. M. & Wilkinson, G. (1997) Impact factors of psychiatric journals. *British Journal of Psychiatry,* **170**, 109–112.
Kiesler, C. (1991) Confusion between reviewer reliability and bad/wise editorial and funding decisions. *Behavioral and Brain Sciences,* **14**, 151–152.
Lock, D. & Smith, J. (1990) What do peer reviewers do? *Journal of the American Medical Association,* **263**, 1341–1343.
Lock, S. (198S) *A Difficult Balance: Editorial Peer Review in Medicine*. London: Nuffield Provincial Hospitals Trust.
McNutt, R., Evans, A., Fletcher, R., 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.
Persaud, R. (1995) Peering into peer review. *Psychiatric Bulletin,* **19**, 529–531.
Pierie, J. P., Walvoort, H. C. & Overbeke, A. J. (1996) Readers' evaluation of effect of peer review and editing of articles in the Nederlands Tijdschrift voor Geneeskunde. *Lancet,* **348**, 1480–1483.
Seglen, P. O. (1997) Why the impact factor should not be used for evaluating research. *British Medical Journal,* **314**, 498–502.
Smith, J. (1990) Journalology—or what editors do. *British Medical Journal,* **301**, 756–759.
Smith, R. (1997) Peer review: reform or revolution? *British Medical Journal,* **315**, 759–760.
——**(1998)** Unscientific practice flourishes in medicine. *British Medical Journal,* **316**, 1036.
Tyrer, P. (1991) Chairman's action: the importance of executive decisions in peer review. *Behavioral and Brain Sciences,* **14**, 164–165.
Weller, A. C. (1990) Editorial peer review in US medical journals. *Journal of the American Medical Association,* **263**, 1344–1347.
Wessely, S. (1996) What do we know about peer review? *Psychological Medicine,* **26**, 883-886.
Wilson, E. B. (1978) Peer review and publication. *Journal of Clinical Investigation,* **61**, 1697–1701.