Reliability of Editors’ Subjective Quality Ratings of Peer Reviews of Manuscripts

Michael L. Callaham, MD; William G. Baxt, MD; Joseph F. Waeckerle, MD; Robert L. Wears, MD

Context.—Quality of reviewers is crucial to journal quality, but there are usually too many for editors to know them all personally. A reliable method of rating them (for education and monitoring) is needed.

Objective.—Whether editors’ quality ratings of peer reviewers are reliable and how they compare with other performance measures.

Design.—A 3.5-year prospective observational study.

Setting.—Peer-reviewed journal.

Participants.—All editors and peer reviewers who reviewed at least 3 manuscripts.

Main Outcome Measures.—Reviewer quality ratings, individual reviewer rate of recommendation for acceptance, congruence between reviewer recommendation and editorial decision (decision congruence), and accuracy in reporting flaws in a masked test manuscript.

Interventions.—Editors rated the quality of each review on a subjective 1 to 5 scale.

Results.—A total of 4161 reviews of 973 manuscripts by 395 reviewers were studied. The within-reviewer intraclass correlation was 0.44 (P<.001), indicating that 20% of the variance seen in the review ratings was attributable to the reviewer. Intraclass correlations for editor and manuscript were only 0.24 and 0.12, respectively. Reviewer average quality ratings correlated poorly with the rate of recommendation for acceptance (R = −0.34) and congruence with editorial decision (R = 0.26). Among 124 reviewers of the fictitious manuscript, the mean quality rating for each reviewer was modestly correlated with the number of flaws they reported (R = 0.53). Highly rated reviewers reported twice as many flaws as poorly rated reviewers.

Conclusions.—Subjective editor ratings of individual reviewers were moderately reliable and correlated with reviewer ability to report manuscript flaws. Individual reviewer rate of recommendation for acceptance and decision congruence might be thought to be markers of a discriminating (ie, high-quality) reviewer, but these variables were poorly correlated with editors’ ratings of review quality or the reviewer’s ability to detect flaws in a fictitious manuscript. Therefore, they cannot be substituted for actual quality ratings by editors.

LITTLE IS KNOWN about assessing the quality of peer review. Most journals do not have standardized methods of selecting reviewers, nor do they screen or train them. We conducted a study to determine if subjective ratings of review quality is a reliable measure; whether it can be easily replaced by more objective reviewer performance statistics, such as

individual rate of recommendation for acceptance; and whether ratings are correlated with reviewer accuracy in detecting errors in manuscripts.

METHODS

When making decisions about manuscripts, editors at Annals of Emergency Medicine routinely rate the quality of each peer review on a subjective ordinal scale developed by senior editors. The scale ranges from 1 (poor) to 5 (excellent) and is not specifically defined other than as “review quality.” No special training in use of the scale is provided, although all editors receive a detailed job description plus the reviewer orientation described below.

All manuscript peer reviews that had been rated by any editors between January 1, 1994, and June 1, 1997, were eligible for study. When recruited by the editorial board, reviewers receive a 4800-word, written orientation on journal expectations for reviews and, with each subsequent review, fill out a form rating 12 specific components of study design, originality, interest for readers, and manuscript quality, plus free text comments for editors and authors. Reviewers are routinely asked to recommend whether each manuscript should be accepted.

Routine performance statistics are calculated by the journal for each reviewer. They include the rate of recommending acceptance and rejection and decision congruence, defined as whether the reviewer recommendation for each manuscript matched the editor’s final decision regarding manuscript acceptance.

The quality ratings were examined for interreviewer agreement using intraclass correlation coefficients (ICCs) because reviews received only 1 editorial rating each, and the ICC takes into account the varying number of scores per reviewer while still weighting each “datapoint” (paper, editor, and the like) only once.

As part of a separate study, in fall 1994 all active reviewers were separately sent a fictitious manuscript, which contained 23 deliberate flaws and was masked to its true purpose, to review. The method is described in detail elsewhere. The performance of reviewers who reviewed the manuscript (and had at least 3 rated reviews) in reporting manuscript flaws was compared with the ratings above.

RESULTS

During the study, 4644 reviews (2922 [63%] of them rated) were performed by 756 reviewers on 1551 manuscripts. To focus on regular (not occasional guest) reviewers, reviewers with 2 or less reviews were excluded, leaving 4161 reviews by 395 reviewers. A total of 2686 (65%) of these were rated by 36 editors, involving 973 manuscripts. The 395 peer reviewers who were studied each reviewed an average of 10.5 manuscripts (95% confidence interval [CI], 9.8-11.2), 6.8 of which were rated (95% CI, 6.3-7.3), receiving an average score of 3.7 (95% CI, 3.5-3.8), they recommended acceptance in 26% of their reviews, rejection in 34%, and revision in 40%.
Spearman Rank Correlation of Individual Reviewer Characteristics to Number of Major Errors Detected

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean (SD)</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reviewer acceptance rate, %*</td>
<td>24 (17)</td>
<td>–0.19</td>
</tr>
<tr>
<td>Review rate, %*</td>
<td>26 (19)</td>
<td>0.18</td>
</tr>
<tr>
<td>Congruence rate with editor's decision, %*</td>
<td>39 (19)</td>
<td>0.22</td>
</tr>
<tr>
<td>Years since residency training</td>
<td>12 (5)</td>
<td>–0.14</td>
</tr>
<tr>
<td>Years of experience reviewing for Annals of Emergency Medicine</td>
<td>5.2 (3.5)</td>
<td>0.03</td>
</tr>
<tr>
<td>Annals of Emergency Medicine reviews in last 3.5 yrs, No.</td>
<td>17 (9)</td>
<td>0.06</td>
</tr>
<tr>
<td>Authored manuscripts (as lead author) No.</td>
<td>5.4 (3.5)</td>
<td>–0.03</td>
</tr>
<tr>
<td>Reviewer for how many other journals, No.</td>
<td>2.9 (1.8)</td>
<td>0.02</td>
</tr>
</tbody>
</table>

*Routine calculated mean performance for individual reviewers.

3.65 (SD, 0.7). These reviewers had an average of 12 (SD, 5.3) years of experience since residency; 40% were assistant professors, 38% were associate professors, and 17% were full professors.

The reviewers detected a mean 3.4 (SD, 1.6) of the 10 major flaws (those that invalidated or undermined the methods of the study) and 3.1 (SD, 1.7) of the 13 minor flaws (those not affecting study validity) in the fictitious manuscript. The mean editorial rating for each reviewer was modestly correlated with the number of flaws they detected (R = 0.45 for major errors, 0.45 for minor errors, and 0.55 for total errors). Reviewers with average ratings of 4 or more reported about twice as many total errors as those with average ratings of 3 or less. Individual reviewer rate of recommendation for acceptance, decision congruence, and various measures of experience were not associated with detection of errors, nor was academic rank (Table). Comment

Half of journal editors rely almost exclusively on reviewer recommendations when making acceptance decisions. Maintaining the quality of peer reviewers is therefore essential, but most journals have too many reviewers for editors to know their capabilities personally. Most journals do not train reviewers or assess their background in research methodology or critical literature review.

Editors need a simple rating system to monitor review quality, but this rating should require minimal additional work and should add unique value. Although the reliability of ratings of manuscripts by reviewers has been previously examined, there is only 1 prior study that examined the reliability of ratings of reviewers by editors. (An earlier study used a 9-axis scale of review quality, but did not test the scale’s reliability.) The study by Feurer et al examined 55 reviews of 25 manuscripts, rated by 3 editors. Editors used a scale of 7 axes with scores of varying complexity, guided by a detailed written explanation. The scale included timeliness standards, whether the reviewer wrote on the manuscript or provided supporting references, and other criteria not directly related to quality of scientific assessment. It demonstrated an excellent ICC of 0.84.

By contrast, we chose a scale that was simpler to use and provided a fair to good ICC of 0.44. It performed in a satisfactory fashion compared with scales used by a specially trained panel of experienced reviewers in rating manuscripts and exceeded the reliability of a multi-axis scale used by peer reviewers for the same purpose. It markedly exceeded the reliability of scoring systems used for awarding National Science Foundation grants. However, it correlated poorly with routine reviewer performance statistics such as rate of recommendation for acceptance and congruence rate. This is logical, since many factors other than just review and manuscript quality affect editorial decisions. It also means that these statistics cannot replace a subjective scale to assess review quality.

The rating scale also showed a modest correlation (R = 0.53) with the ability of reviewers to detect deliberate flaws in a fictitious manuscript. The more flaws they were able to detect and report, the more likely they were to produce reviews judged excellent by editors. Reviewers with high average ratings reported twice as many errors as those with low ratings. Reviewers reported only a minority of such flaws, similar to a prior study. There is no “gold standard” for judging a quality peer review, but this at least suggests that our subjective quality rating may be related to accuracy and/or thoroughness.

Our scale had only 1 axis and the definition of quality was not specific; previous research has suggested that multiple, more specific scales increase the reliability of ratings. Despite this, our ICC was higher than previous multi-axis scales used to rate manuscripts, although not as reliable as the much more complex scale of Feurer et al, which included objective data.

We have found this simple scale useful as part of a comprehensive evaluation to identify poor reviewers who should be dropped and excellent reviewers who deserve acknowledgment. Its reliability could be improved by defining the qualities of a good review more specifically and by formally training editors in its use.

In conclusion, this simple 5-point scale for editors to rate peer review quality is acceptably reliable and correlates moderately with reviewers’ reporting of flaws in a test manuscript. It provides a useful tool for monitoring reviewer performance and cannot be replaced by routine reviewer statistics such as individual rate of recommendation for acceptance and decision congruence.

We gratefully thank Larene Schiltz, Margaret Levenson, BA, Marian Wieseman, R.N., BA, McDonald, BA, and all the Annals of Emergency Medicine staff whose hard work made this study possible.

References

What Makes a Good Reviewer and a Good Review for a General Medical Journal?

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

Context.—Selecting peer reviewers who will provide high-quality reviews is a central task of editors of biomedical journals.

Objectives.—To determine the characteristics of reviewers for a general medical journal who produce high-quality reviews and to describe the characteristics of a good review, particularly in terms of the time spent reviewing and turnaround time.


Main Outcome Measures.—Review quality was assessed independently by 2 editors and by the corresponding author using a newly developed 7-item review quality instrument.

Results.—Of the 420 manuscripts, 345 (82%) had 2 reviews completed, for a total of 690 reviews. Authors’ assessments of review quality were available for 507 reviews. The characteristics of reviewers had little association with the quality of the review. In a logistic regression analysis, the only significant factor associated with higher-quality ratings by both editors and authors was reviewers trained in epidemiology or statistics. Younger age also was an independent predictor for editors’ quality assessments, while reviews performed by reviewers who were members of an editorial board were rated of poorer quality by authors. Review quality increased with time spent on a review, up to 3 hours but not beyond.

Conclusions.—The characteristics of reviewers we studied did not identify those who performed high-quality reviews. Readers might be advised that spending longer than 3 hours on a review on average did not appear to increase review quality as rated by editors and authors.

From the London School of Hygiene & Tropical Medicine (Drs Black and Evans) and BMJ (Ms Rooyen and Drs Godlee and Smith), London, England.

Presented at the International Congress on Peer Review in Biomedical Publication, Prague, Czech Republic, September 18, 1998.

Reprints: Nick Black, MD, Health Services Research Unit, London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT, England (e-mail: n.black@lshtm.ac.uk).

ALTHOUGH all editors would like to know how to select good reviewers, there have been only 3 attempts to identify their characteristics.1-3 Two of these studies found that the best-quality reports were provided by reviewers who were young,1,2 and, therefore, of junior academic status,1 particularly if they were working at a top academic institution or were known to the editors.4 The third study demonstrated that younger reviewers with considerable refereeing experience provided stricter assessments of manuscripts than other reviewers.5 None of the other characteristics examined (such as research training and postgraduate qualifications) were associated with review quality.

The process of peer reviewing has also received little attention. While 3 studies have reported on the time spent by reviewers on the task,4,6 none examined the relationship between time spent and review quality.

Our principal objective was to determine the characteristics of reviewers who produce high-quality reviews. In addition, we considered the characteristics of good reviews in terms of the time spent by the reviewer and time taken to deliver it to the journal.

METHODS

Consecutive manuscripts (research papers) submitted to BMJ and sent for review between January and June 1997 were eligible for inclusion. Each manuscript was sent to 2 reviewers (selected from the existing reviewer database) as having an interest in and knowledge of the subject matter of the manuscript) as part of a randomized trial of blinding and unmasking to coreviewer.7 Reviewers were supplied with the journal’s standard advice and asked to return their...