# Chance, concurrence, and clustering. Analysis of reviewers' recommendations on 1,000 submissions to the Journal of Clinical Investigation.

B F Scharschmidt, … , P Bacchetti, M J Held

*J Clin Invest.* 1994;93(5):1877-1880. https://doi.org/10.1172/JCI117177.

Research Article



# Chance, Concurrence, and Clustering

## Analysis of Reviewers' Recommendations on 1,000 Submissions to *The Journal of Clinical Investigation*

**Bruce F. Scharschmidt,* Amy DeAmicis,* Peter Bacchetti,‡ and Michael J. Held***
*Departments of *Medicine, and ‡Epidemiology and Biostatistics, School of Medicine, University of California, San Francisco, San Francisco, California 94143*

Publication in a peer-reviewed journal is typically the culmination of an individual research project. Collectively, publications chronicle and facilitate the evolution of science, and the publications of an individual investigator represent both a permanent record of his or her contributions to a particular discipline and an important criterion for academic promotion (1). Editors of leading journals seek to publish papers that are innovative, important, and scientifically sound, and they typically rely upon the opinions of two or more independent reviewers to help them make such judgments.

Despite the central importance of peer review and the increasing interest in the peer review process, there is comparatively little information about grading tendencies of reviewers, the degree to which reviewers of an individual manuscript agree or disagree on technical and conceptual issues, and the possible relationship between the ways in which reviewers are selected and the judgments they render. To address these questions, we took advantage of the opportunity available to us during our stewardship of *The Journal of Clinical Investigation* (*JCI*), a rigorously reviewed publication of the American Society for Clinical Investigation, which ranks among the top biomedical research journals in citation frequency and overall impact (2–4). At the time of this study, the *JCI* had an acceptance rate of 25–30%.

### *Editorial handling of manuscripts at the time of this study and analysis of reviewers' responses*

*Editorial handing of manuscripts.* First-cycle submissions analyzed for this study needed to have received two reviews and have been submitted for regular publication, since manuscripts submitted for rapid publication or as "Perspectives" (i.e., Mini-Reviews) were reviewed in a less standardized fashion. Between June 1, 1989 and April 9, 1990, 1,000 consecutive submissions met these criteria and constituted the study group.

Submitting authors were requested to provide the names of several potential reviewers. The editor or associate or consulting editors were free to choose reviewers based upon the editor's own familiarity with the scientific discipline or from the list provided by the author. Editors were expected to choose reviewers thought to be best qualified and likely to provide a fair, timely, and comprehensive review. Editors provided a list of up to six reviewers ranked in order of preference, and the staff sent the manuscript to the first two "available" reviewers on the list (i.e., had not reviewed a manuscript for the JCI within 45 days). If no reviewers were available, the manuscript was returned to the editor for additional reviewer assignments. Reviewers were not chosen by the office staff.

Each of the two reviewers was asked to assign whole integer grades from 1 (highest) to 5 (lowest) in the categories of "originality," "experimental design," "data support the conclusions," and "overall priority" for publication. Reviewers also were asked to make a recommendation regarding publication by selecting among the four categories of "accept as is" (defined as a grade of 1), "accept with revision" (grade of 2), "reject with invitation to resubmit" (grade of 3), and "reject" (grade of 4).

*Analysis of reviewer agreement.* Overall reviewer agreement was analyzed for each of the above categories for all manuscripts. Reviewer agreement was also analyzed in relation to whether the initial editorial decision was to accept ("accept as is" or "accept with revision") or reject ("reject with an invitation to resubmit" or "reject") the manuscript, and whether the reviewers were identified by the editor or from the list provided by the author. Assignment of the same numerical grade by each member of the reviewer pair was defined as complete agreement, and assignment of numerical grades differing by one was defined as near agreement. Grade assignments which were not whole integers, but included decimal points, were rounded to the nearest integer (e.g., 1.3 rounded to 1.0) for the purposes of this analysis. If either of the reviewers failed to assign a grade in a category, that category for that manuscript was excluded from analysis. Other categories from that same manuscript with grade assignments from both reviewers were included.

*Statistical analysis.* For the purposes of calculating rates of agreement or near agreement expected by chance, it was assumed that each reviewer's grade was chosen independently of the other reviewer's grade. The probability of choosing a given grade was set to equal the overall rate at which that grade was assigned for all manuscripts by all reviewers in the study. It should be noted that the rate of agreement expected by chance is lower if the probabilities of different grades are equal (i.e., 20% chance for each of five grades) than if there is clustering of grades, as was actually observed (see below).

The statistical significance of each observed agreement rate was calculated from the binomial distribution with rate equal to the expected agreement rate. *P* values for differing agreement rates in different categories of reviewer pairs (e.g., those reviewing accepted versus rejected manuscripts) were calculated by Fisher's exact test. For the analysis of author-recommended versus editor-selected reviewers, the matched-pair *t*-test was used to compare grades; the statistical significance of

Address all correspondence to Bruce F. Scharschmidt, M.D., Gastroenterology Division, 1120 HSW, Box 0538, University of California, San Francisco, San Francisco, CA 94143-0538

*Received for publication 21 December 1993 and in revised form 31 January 1994.*

J. Clin. Invest.

0021-9738/94/05/1877/04 $2.00
Volume 93, May 1994, 1877–1880

the number of manuscripts with better grades from the author-recommended versus the editor-selected reviewer was calculated using the binomial distribution with rate equal to 0.5 and sample size equal to the number of pairs that did not show complete agreement.

### *Reviewer grading and recommendations*

*Did reviewers exhibit consistent grading tendencies?* Grades from both reviewers were available for 767–847 manuscripts of the 1,000 analyzed, depending on the category. The grades assigned by reviewers were not evenly distributed from highest to lowest (Fig. 1); rather, there was clustering in the middle. In categories (e.g., "originality") with five possible grades, 55–67% of grades were either 2 or 3, and only 7.6–15.8 and 3.7–7.9%, respectively, of grades were 1 or 5. The same was true for recommendation regarding publication, where only 6% of the reviewers recommended acceptance without revision.

*How good was agreement between reviewers?* The overall rate of complete or near agreement between reviewers for all manuscripts was high (67–80%). However, when the clustering of reviewer grades in the middle (see above) was taken into account, this rate of agreement was only marginally (4.4–7.5%), albeit significantly, better than predicted by chance (Table I). Among the various grading categories, complete agreement rates were lower for the category of "data support the conclusions" (26.0%) than for other categories (Table I).

*Did reviewers recommended by authors grade differently than those selected by editors?* Grades in all categories assigned by author-recommended reviewers were significantly ($P = 0.003$ to $P < 0.0001$ by matched-pairs $t$-test) better than grades assigned by editor-identified reviewers. This discrepancy in grades between differentially selected reviewers occurred at all grading levels, but it was particularly true for the highest grades (Fig. 2). That is, author-recommended reviewers were 2–3 times more likely to assign the highest possible grade or to make a recommendation to "publish as is."

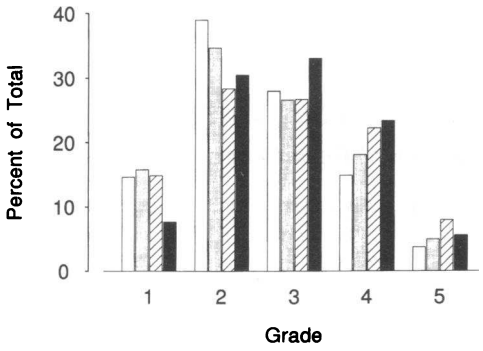


*Figure 1.* Distribution of grades from 1 (best) to 5 (worst) in the categories of "originality" (*open bar, n* = 1548), "experimental design" (*stippled bar, n* = 1542), "data support the conclusions" (*hashed bar, n* = 1536) and "overall priority" (*solid bar, n* = 1534).

When individual pairs of differentially selected reviewers were analyzed, 42–50% of grades assigned by the author-recommended reviewers were better ($P = 0.0051$ to $P < 0.0001$) than the grades assigned by the editor-identified reviewers, depending on the grading category. By contrast, the reverse was true for only 22–28% of the grades. For the category of "overall priority" (Fig. 2), the author-recommended reviewers' grades averaged 0.35 points better (95% CI 0.18–0.51, $P = 0.0001$) than the grades of editor-identified reviewers. The agreement between these differentially selected reviewer pairs ($n = 256$), with respect to recommendation regarding publication, was significantly ($P = 0.036$) poorer than among the

*Table I. Agreement between Reviewers as Compared With Chance: Relationship to Decision Category and how Reviewers Were Identified*

| Category | Agreement* | Agreement rates (Number of reviewer pairs) Expected by chance‡ | All reviewers | Discordant reviewers | Concordant reviewers |
|---|---|---|---|---|---|
| | | % | % | % | % |
| Originality | Complete | 27.4 | 34.2¶ | 32.2 | 35.0¶ |
| | Complete or near | 69.9 | 74.3‖ | 74.6 | 74.1§ |
| | | | (774) | (236) | (537) |
| Experimental design | Complete | 25.1 | 29.6§ | 27.7 | 30.5‖ |
| | Complete or near | 65.8 | 73.0¶ | 71.9§ | 73.5¶ |
| | | | (771) | (235) | (535) |
| Data support conclusions | Complete | 22.9 | 26.0§ | 23.1 | 27.4§ |
| | Complete or near | 61.8 | 67.2‖ | 64.1 | 68.5‖ |
| | | | (768) | (234) | (533) |
| Priority | Complete | 26.5 | 31.2‖ | 33.2§ | 30.4§ |
| | Complete or near | 69.3 | 75.9¶ | 75.1§ | 76.2‖ |
| | | | (767) | (229) | (537) |
| Recommendation regarding publication | Complete | 30.1 | 36.2¶ | 30.9** | 38.5¶ |
| | Complete or near | 72.0 | 79.5¶ | 79.7‖ | 79.3¶ |
| | | | (847) | (256) | (590) |

* Complete and near agreement, respectively, are defined as a numerical grade difference of 0 and 1. ‡ The agreement rate expected by chance is calculated from the observed distribution of reviewer scores for that category (Fig. 1). § $P < 0.05$ compared with chance. ‖ $P < 0.01$ compared with chance. ¶ $P < 0.0001$ compared with chance. ** $P = 0.036$ compared with concordant pairs.

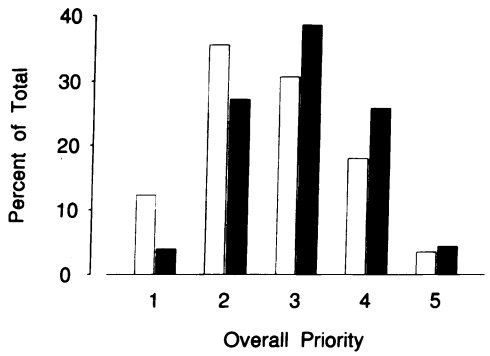


*Figure 2.* Grades assigned regarding overall priority, from 1 (highest) to 5 (lowest) by reviewer pairs for 228 manuscripts in which one member of the pair was recommended by the author (*open bar*), and one member of the pair was identified by the editor (*solid bar*).

other reviewer pairs ($n$ = 590), for which both members of the pair were identified by either the author ($n$ = 191) or editor ($n$ = 399) (Table I).

*What was the relationship between reviewers' grades and editorial decision?* Only manuscripts assigned top grades (1 or 2) by both reviewers had a high acceptance rate (Table II). Notably, even among manuscripts assigned an average grade of 2, the acceptance rate for manuscripts receiving a grade of 2 from both reviewers was more than double that for those manuscripts receiving grades of 1 and 3. As a corollary, in contrast to previous reports (see reference 11), complete or close agreement, collectively, was greater (82–92%) in all categories for manuscripts accepted outright or with revision than for those (65–78%) rejected outright or with an invitation to resubmit.

### *Summary and perspective*

Most analyses of the peer review process are based on personal observation and philosophy rather than systematic data gathering, and prior data-based analyses have led to different conclusions. Some reports have emphasized apparent problems with peer review, including its random and inconsistent nature and its inadequacy in recognizing innovative work (5–9). Others have reported reviewer agreement to be high and greater than predicted by chance (10–13). Reviewer agreement may differ depending on scientific discipline (14), and prior studies of journals in the social or behavioral sciences may not pertain to biomedical research journals such as *JCI*. Moreover, prior studies regarding reviewer agreement have not taken into account the grading tendencies of reviewers or the ways in which reviewers are selected. The present study is the largest comprehensive analysis of which we are aware of reviewer grading and agreement for a biomedical research journal.

Reviewers for the *JCI* tended to assign middle grades rather than grades at the extremes. With respect to editorial decision, they infrequently recommended outright acceptance or flat rejection. This likely reflects, on the one hand, the reviewers' perception of the JCI as a rigorous journal which anticipates rigorous reviews, and, on the other hand, the desire of most reviewers to be constructive and to identify worthwhile elements in even weak manuscripts.

*JCI* reviewers exhibited complete or close agreement in two thirds to three quarters of instances, and they tended to agree more often in the subjective category of "originality" than in the category of "data support the conclusions." While agreement between JCI reviewers was better than predicted by chance, it was only modestly so when the effects of grade clustering were taken into account (Table I). We found, in addition, that author-recommended reviewers graded more favorably than did editor-identified reviewers (Fig. 2). This discrepancy in grading, which occurred despite the screening of reviewers to eliminate those with potential conflicts which might cloud their objectivity (i.e., a collaborative arrangement with the author, as evidenced by prior coauthorship of publications, or appointment at the same institution), likely reflects the desire of authors to avoid reviewers regarded as antagonistic and to recommend reviewers who are familiar with and supportive of their field of research. The greater tendency of author-recommended reviewers to assign top grades (Fig. 2) or recommend "acceptance without revision" is particularly noteworthy. Because such grades and recommendations were uncommon, they tend to attract the attention of the editor.

The phenomena of grade clustering and disparate scoring by editor-identified versus author-recommended reviewers which we observed in this study tend to lessen the value of outside reviews in distinguishing among manuscripts which fall between the extremes of outstanding and bad; that is, most submissions. To the degree that grade clustering reflects inadequate communication, it may be improved by more explicit instructions to reviewers regarding the journal's priorities and publication standards. In fact, the current *JCI* instruction

*Table II. Percent Acceptance in Relation to Reviewer's Grades*

| | | Categories | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Originality | | Experimental design | | Data support conclusions | | Priority | | Recommendation regarding publication | |
| Average grade | Grade pairs | Percent acceptance* | No. of manuscripts | Percent acceptance* | No. of manuscripts | Percent acceptance* | No. of manuscripts | Percent acceptance* | No. of manuscripts | Percent acceptance* | No. of manuscripts |
| 1–1.5 | 1–1, 1–2 | 38.1 | 113 | 49.2 | 124 | 52.4 | 105 | 68.4 | 57 | 62.0 | 50 |
| 2.0 | 2–2 | 28.5 | 130 | 23.8 | 101 | 30.3 | 66 | 51.3 | 78 | 47.7 | 130 |
| 2.0 | 1–3 | 17.3‡ | 52 | 12.0‡ | 50 | 18.0‡ | 50 | 26.8§ | 41 | 19.0§ | 21 |
| 2.5 | 2–3, 1–4 | 7.4 | 190 | 6.4 | 171 | 8.2 | 146 | 5.6 | 162 | 7.2 | 195 |
| ≥3.0 | 3–3, etc. | 2.1 | 289 | 1.8 | 325 | 2.7 | 401 | 1.9 | 429 | 0.9 | 451 |

* Includes "accept as is" and "accept with revision." ‡ $0.05 < P < 0.1$, as compared with 2–2 grade pairs. § $P < 0.05$, as compared with 2–3 grade pairs.

forms have been restructured with the aims of achieving a more useful grade distribution and better understanding the rationale for reviewers' rankings. This sets the stage for a subsequent analysis to assess the impact and effectiveness of these changes in directives to reviewers.

Whatever the reasons for grade clustering, it helps explain the relationship we found between reviewers' grades and editorial decision. Editors of leading journals with a low acceptance, such as the *JCI*, receive far more submissions than they can publish. Therefore, they are most likely to accept manuscripts which represent "safe" choices, such as manuscripts given high grades by both reviewers. A negative or lukewarm response by one reviewer is likely to dampen the editor's enthusiasm, even for potentially innovative or novel work, and result in rejection. Our findings do not permit us to assess whether this practice of requiring consensus facilitates or hinders publication of innovative studies which propose new and potentially controversial paradigms and have received mixed reviews. However, it is in the evaluation of such work that the editor's input is most critical. Clearly, authors who submit manuscripts, as well as the editors and reviewers who judge them, need to be aware of the relationship between reviewer grading and editorial decision making and use this information to strengthen the peer review process.

## Acknowledgments

The authors gratefully acknowledge the contributions of those members of the American Society for Clinical Investigation who acted as Associate and Consulting Editors of the *JCI* during the time this study was conducted, as well as the contribution of the UCSF Liver Center Editorial Core Facility, and Lee Cantrell and Sadie MacFarlane for their assistance in preparing this manuscript.

## References

1. Stossel, T. P. 1987. Volume: papers and academic promotion. *Ann. Int. Med.* 106:146–149.

2. Science Citation Index: Journal Citation Reports. 1988. ISI Press, Philadelphia, PA.

3. Garfield, E. 1987. Fifty classics from The Journal of Clinical Investigation: Over 60 years of Nobel-class research. *Current Contents.* 8:3–8.

4. Scharschmidt, B. F. 1990. Something old, something new, something blue. *J. Clin. Invest.* 85:1.

5. Smith, R. 1988. Problems with peer review and alternatives. *Br. Med. J.* 296:774–777.

6. Horrobin, D. F. 1990. The philosophical basis of peer review and the suppression of innovation. *JAMA (J. Am. Med. Assoc.)* 263:1438–1441.

7. Yalow, R. S. 1982. Competency testing for reviewers and editors. *Behav. Brain Sci.* 5:244–245.

8. Cole, S., J. R. Cole, and G. A. Simon. 1991. Chance and consensus in peer review. *Science (Wash. DC).* 214:881–886.

9. Peters, D. P., and S. J. Ceci. 1982. Peer-review practices of psychological journals: the fate of published articles, submitted again. *Behav. Brain Sci.* 5:187–195.

10. Ingelfinger, F. J. 1974. Peer review in biomedical publication. *Am. J. Med.* 56:686–692.

11. Locke, S. 1985. A Difficult Balance: Editorial Peer Review in Medicine. The Nuffield Provincial Hospitals Trust, London.

12. Gallagher, E. B., and J. Ferrante. 1991. Agreement among peer reviewers for a middle-sized biomedical journal. *In* Peer Review in Scientific Publishing: Papers from the First International Congress on Peer Review in Biomedical Publication. Council of Biology Editors, Chicago, IL. 153–158.

13. Oxman, A. D., G. H. Guyatt, J. Singer, C. H. Goldsmith, B. G. Hutchison, R. A. Milner, and D. L. Streiner. 1991. Agreement among reviewers of review articles. *J. Clin. Epidemiol.* 44:91–98.

14. Zuckerman, H., and R. K. Merton. 1971. Patterns of evaluation in science: institutionalization, structure and functions of the referee system. *Minerva.* 9:66–100.

*Editors' Note.* Partly as a consequence of the data presented above, the *JCI* has changed its system for obtaining manuscript ratings. Reviewers are now asked to rate papers relative to others in the same field on a subdivided percentile scale (see below). This rating method seems to avoid the clustering near the middle seen with the prior system. In a sampling of reviewer responses using this method (see table below) the ratings appear skewed to the right (i.e., > 25% of the ratings are in the top 25% and higher categories). However, only ~ 65% of submitted manuscripts get two full reviews because of the possibility of early rejection or screening rejection (see accompanying Editorial in this issue for details). Thus, one should ideally see ~ 40% of the ratings in these upper categories—the actual values range from 45 to 55%. While these ratings are very useful, it should be emphasized that they represent only one form of feedback from the reviewers—the Editors consider many other factors in making a final decision on a given manuscript. The Board has also taken note of the finding in the above study that author-recommended reviewers tend to assign higher ratings than those picked by editors.

*Please rate this manuscript relative to other manuscripts in this field*

| (Check one box for each) | Lower 50% | Top 50% | Top 25% | Top 10% | Top 5% | Top 2% |
|---|---|---|---|---|---|---|
| Experimental design | 70 | 107 | 99 | 60 | 36 | 7 |
| Data quality | 65 | 102 | 89 | 67 | 42 | 10 |
| Originality | 43 | 102 | 96 | 88 | 35 | 11 |
| Overall priority for publication | 94 | 101 | 81 | 64 | 24 | 10 |

Summary of ratings for 200 fully reviewed *JCI* manuscripts by 400 reviewers. The actual rating matrix provided to reviewers on the response form is shown. The figures in italics shows the cumulative number of times each box was filled by the 200 reviewers. The boxes were left blank only ~ 5% of the time for each category.