https://doi.org/10.51408/issi2025_189

AI-Powered Evaluation of Peer Review Quality: A Case Study of eLIBRARY.RU

Dmitry Kochetkov[1], Denis Kosyakov[2], Irina Lakizo[3], Viktor Glukhov[4], Andrey Guskov[5]

[1]*kochetkov@elibrary.ru, d.kochetkov@cwts.leidenuniv.nl*
Scientific Electronic Library LLC, Nauchny Proezd 14A block 3, 117246 Moscow (Russian Federation)
Centre for Science and Technology Studies, Leiden University, Kolffpad 1, 2333 BN Leiden (The Netherlands)

[2]*kosyakov@sscc.ru, d.kosyakov@riep.ru,* [5]*guskov@sscc.ru, a.guskov@riep.ru*
Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Ac. Lavrentieva ave. 6, Novosibirsk (Russian Federation)
Russian Research Institute of Economics, Policy and Law in Science and Technology, Dobrolubova Str. 20A, Moscow (Russian Federation)

[3]*i.lakizo@riep.ru*
Russian Research Institute of Economics, Policy and Law in Science and Technology, Dobrolubova Str. 20A, Moscow (Russian Federation)

[4]*olunid@elibrary.ru*
Scientific Electronic Library LLC, Nauchny Proezd 14A block 3, 117246 Moscow (Russian Federation)

Introduction

eLIBRARY.RU is the largest Russian electronic library of scientific publications and home to the Russian Index of Science Citation (RISC) and highly selective Russian Science Citation Index (RSCI). One of the challenges we face in the expert evaluation of review quality and journal policies is the shortage of qualified experts. A potential solution to this problem is the use of *generative artificial intelligence* (GenAI) technologies to assess the quality of reviews. Recent studies cautiously evaluate the potential of GenAI in scientific peer review. For example, AI tools can assist in the initial screening of articles, plagiarism detection, and reviewer matching, potentially saving millions of working hours (Checco et al., 2021). However, concerns remain about biases and ethical implications (Shcherbiak et al., 2024). Seghier (2025) advocates for the gradual integration of AI into the peer review process under human oversight, emphasizing the importance of transparency, accountability, and robust safeguards. At the same time, *the potential of AI technologies for evaluating review quality remains largely unexplored*.

The goal of this study is to address the question of *whether AI-based evaluation of journal review quality is feasible at the current level of technological development.* This report presents preliminary findings based on a test sample of 240 reviews.

Data and Methods

To assess peer review quality, we created a test sample by selecting four diverse disciplines (*Economics & Business, Information & Computer Science, Physics & Mathematics, and Medicine*) to test AI versatility across different research types. Within each discipline, we chose two journals representing high-impact (top 1-500) and mid-tier (1501-2000) rankings in the *Science Index*[1], randomly selecting 30 review reports

[1] Science Index is a composite journal ranking on eLIBRARY.RU.

from each journal. This approach ensured a diverse sample spanning methodological approaches and journal prestige levels.

The selected reviews were evaluated using two sets of criteria. The first set, based on Russian Science Citation Index parameters, assessed *depth, usefulness, rigor, and clarity*. The second set adapted the *Review Quality Instrument (RQI)* (van Rooyen et al., 1999), evaluating eight aspects: *research importance, originality, methods, presentation, comment constructiveness/substantiation, result interpretation, and overall quality*. Each criterion was scored on a detailed 5-point Likert scale. GPT-4 was employed via API to assign scores and provide justifications, specifically referencing the review text for the RQI criteria. The process ensured no disclosure of personally identifiable information.

Results

The results of the analysis based on *Criterion Set 1* are presented in Figure 1. Journals are categorized by subject area: Economics and business (eco), Information and computer science (info), Physics and mathematics (phys), and Medicine (med), as well as by their ranking range in the Science Index (SI) – 1-500 (index 1) or 1501-2000 (index 2).

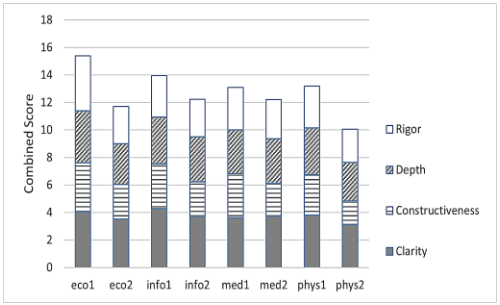

Figure 1. Average scores by journals' categories according to Criterion Set 1.

The quality of reviews in journals across all disciplines was higher for those in the SI 1-500 range compared to those in the 1501-2000 range. This finding indirectly supports the hypothesis of a correlation between bibliometric indicators and the quality of editorial policies, particularly peer review.

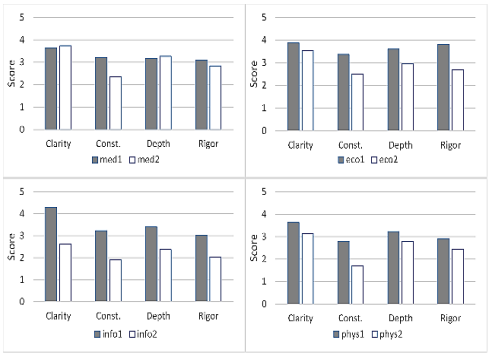

Figure 2. Comparative analysis of the average scores by criteria and journals' category according to Criterion Set 1.

For *Clarity* and *Depth* criteria, we see a superiority of mid-tier journal scores over high-impact journal scores in *Medicine* (*Figure 2)*. In other disciplines, review scores for all four criteria are weaker for mid-tier journals.

The application of *Criterion Set 2* yielded slightly different results (*Figure 3*). In this case, the difference between journals in the two ranking tiers was less pronounced in economics and business. Moreover, the medical journal in the 1501-2000 range performed slightly better than its counterpart in the 1-500 range. In contrast, the advantage of high-impact journals is more pronounced in the other two areas.

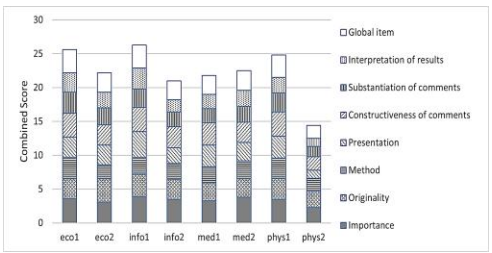

Figure 3. Journals' scores according to Criterion Set 2.

To analyze in detail the results that do not fit the intended picture, we compared the scores of the journals for each criterion (*Figure 4*). The mid-tier medical journal outperformed the high-impact journal in all but two criteria: *Presentation*, and *Constructiveness of comments*. The superiority of mid-tier journal is also observed in the field of *Economics and business* in terms of *Originality* and to a lesser extent in terms of *Presentation*. The most

significant difference was observed for criterion *Importance*.

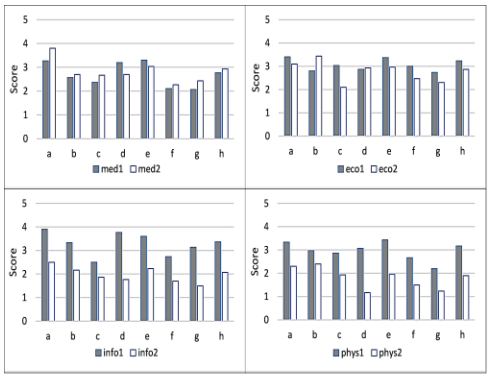

Figure 4. Comparative analysis of the average scores by criteria and journals' category according to Criterion Set 2. Criteria: a – importance, b – originality, c – method, d – presentation, e - constructiveness of comments, f - substantiation of comments, g - interpretation of results, h- global item.

Competing Interests

Dmitry Kochetkov and Viktor Glukhov are Deputy Directors of Scientific Electronic Library LLC, the operator of eLIBRARY.RU, RISC, and RSCI.

References

Checco, A., Bracciale, L., Loreti, P., Pinfield, S., & Bianchi, G. (2021). AI-assisted peer review. *Humanities and Social Sciences Communications*, 8(1), 25. https://doi.org/10.1057/s41599-020-00703-8

Seghier, M. L. (2025). AI-powered peer review needs human supervision. *Journal of Information, Communication and Ethics in Society*, 23(1), 104–116. https://doi.org/10.1108/JICES-09-2024-0132

Shcherbiak, A., Habibnia, H., Böhm, R., & Fiedler, S. (2024). Evaluating science: A comparison of human and AI reviewers. *Judgment and Decision Making*, 19, e21. https://doi.org/10.1017/jdm.2024.24

van Rooyen, S., Black, N., & Godlee, F. (1999). Development of the Review Quality Instrument (RQI) for Assessing Peer Reviews of Manuscripts. *Journal of Clinical Epidemiology*, 52(7), 625–629. https://doi.org/10.1016/S0895-4356(99)00047-5