From Open Review to Reproducible Review: FAIR Metrics Analysis of Open Peer Reviews for Brain Informatics Literature

Adam Craig and Carl Taswell

Brain Health Alliance, Ladera Ranch CA, USA

BHAVI Symposium online 9 October 2024

The challenge of assessing the quality of peer review

- Even experts in a field often disagree about the quality of a peer review.
- According to one study, the correlation in ratings of the same peer review among three editors was only 0.62 (95% confidence interval (CI) 0.50-0.71).
- The correlation between the author's rating and the mean of the editors' ratings was even lower, 0.28 (95% CI 0.14-0.41).
- Even the correlations between a first rating and second rating of the same peer review by the same editor was between 0.66 and 0.88.
- Source: Landkroon, A. P., Euser, A. M., Veeken, H., Hart, W., & Overbeke, A. J. P. (2006). Quality assessment of reviewers' reports using a simple instrument. *Obstetrics & Gynecology*, 108(4), 979-985.

Quantitative assessments of peer review

- We have only found two examples of quantitative, objective metrics of peer review quality.
- One counts the number of reviewers and number of editors.
- Optionally, one can weight them by some quantity, such as their Hirsch indices.
- Source: Etkin, A. (2014). A new method and metric to evaluate the peer review process of scholarly journals. *Publishing research quarterly*, 30, 23-38.
- The other metric measured amount of text and tone.
- Each comment is either positive or negative and either constructive or unconstructive.
- Source: Dobele, A. R. (2015). Assessing the quality of feedback in the peer-review process. *Higher Education Research & Development*, 34(5), 853-868.
- Both metrics are different ways of measuring the quantity of peer review received, not necessarily the quality.

What even defines good peer review?

- Studies have tested the effects of different approaches to approving peer review: concealing the identity of the author from the reviewer, providing a checklist or training to the reviewer, etc.
- However, the limited scale, variable design, and ambiguous results of such studies makes it difficult to draw conclusions.
- Source: Jefferson, T., Alderson, P., Wager, E., & Davidoff, F. (2002). Effects of editorial peer review: a systematic review. *JAMA*, 287(21), 2784-2786.
- Even the designs of questionnaires used to rate peer reviews in different studies vary widely.
- This suggests that researchers disagree over the standards peer review should meet.
- Source: Jefferson, T., Wager, E., & Davidoff, F. (2002). Measuring the quality of editorial peer review. *JAMA*, 287(21), 2786-2790.

Open peer review

- We cannot force authors, reviewers, and editors to agree on what constitutes a good peer review.
- We can ask for publication of reviews so that readers can decide for themselves.
- This is one aspect of open peer review, but it can have others, including...
- Identification of the reviewers and editors to the authors.
- Allowing correspondence between authors, reviewers, and editors.
- Publication of the names of authors alongside the reviews.
- Source: Wolfram, D., Wang, P., Hembree, A., & Park, H. (2020). Open peer review: promoting transparency in open science. *Scientometrics*, 125(2), 1033-1051.

Reproducible peer review

- When a journal fully implements open peer review, we know who reviewed an article and what they said.
- We still do not know on what basis they said it.
- At *Brainiacs Journal*, our goal is not just open but reproducible peer review.
- The standard: Can a second reviewer follow the reasoning and evidence the first reviewer used to arrive at their conclusion?
- Source: Craig, A., Lee, C., Bala, N., & Taswell, C. (2022). Motivating and maintaining ethics, equity, effectiveness, efficiency, and expertise in peer review. *Brainiacs Journal* 3 (1):I5B147D9D

The Fair Attribution to Indexed Reports (FAIR) Metrics

- To rate reproducibility of peer reviews, we need well-defined metrics.
- Our starting point is the design principles for the FAIR Metrics.
- The goal: Measure how well a scholarly work makes it possible to trace ideas back to their sources.
- Ideas are the same when they are equivalent in meaning, regardless of wording.
- The process of calculating the metrics must itself be reproducible, recorded in a systematic, preferably machine-readable format.
- Source: Craig, A., Ambati, A., Dutta, S., Kowshik, P., Nori, S., Taswell, S. K., ... & Taswell, C. (2019, June). DREAM Principles and FAIR Metrics from the PORTAL-DOORS Project for the Semantic Web. In 2019 *11th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)* (pp. 1-10). IEEE.

The 4 FAIR Metrics of appropriate use of citations

- We first introduced FAIR Metrics to measure the appropriateness of citations in scholarly works.
- We first categorize and count the substantive claims, using 4 counts:
  - A Number correctly attributed
  - M Number misattributed
  - N Number genuinely novel
  - P Number plagiarized
- We next compute ratio FAIR Metrics emphasizing different counts:
  - $F_A = \frac{A}{A+M+P}$ general appropriateness of attribution
  - $F_M = \frac{A-M}{A+M+P}$ misattribution focused
  - $F_N = \frac{A-N}{A+M+N+P}$ balance of novel to attributed (non-normative)
  - $F_P = \frac{A-P}{A+M+P}$ plagiarism focused
- Source: Craig, A., Athreya, A., & Taswell, C. (2023, October). Example evaluations of plagiarism cases using FAIR Metrics and the PDP-DREAM Ontology. In *2023 IEEE 19th International Conference on e-Science (e-Science)* (pp. 1-2). IEEE.

Design criteria for FAIR Metrics of peer reviews

- We next introduced FAIR Metrics for reproducibility of peer reviews.
- To be reproducible, a review must justify recommendations and opinions with factual claims.
- Does the reviewer correctly attribute the claim to a source (A)?
- Does the reviewer misattribute it or omit a source (M)?
- Reviews typically do not introduce novel ideas, so no N or P.
- We also categorize claims by the type of source:
  - T claims about the target work itself
  - V claims about the publication venue (conference or journal)
  - D claims based on domain-relevant outside knowledge
- Source: Craig, A., & Taswell, C. (2024, September). FAIR Metrics for Motivating Excellence in Peer Review. In *2024 IEEE 20th International Conference on e-Science (e-Science)* (pp. 1-2). IEEE.

The 4 FAIR Metrics for peer reviews

- This gives us 6 counts: A_T, M_T, A_V, M_V, A_D, M_D
- We then use these counts to define four ratio FAIR Metrics:
  - $f_T = (A_T - M_T)/(A_T + M_T)$
  - $f_V = (A_V - M_V)/(A_V + M_V)$
  - $f_D = (A_D - M_D)/(A_D + M_D$
  - $f_J = (A_T + A_V + A_D - M_T - M_V - M_D)/(A_T + A_V + A_D + M_T + M_V + M_D)$
- Source: Craig, A., & Taswell, C. (2024, September). FAIR Metrics for Motivating Excellence in Peer Review. In *2024 IEEE 20th International Conference on e-Science (e-Science)* (pp. 1-2). IEEE.

Initial test cases

- To demonstrate the FAIR Metrics, we chose 5 examples.
- A simple synthetic example peer review to illustrate the three categories of claims.
- 2 peer reviews of our rejected software grand challenge proposal submitted to ACM Multimedia 2023, a revised version of which we later published in *Brainiacs Journal*: Craig, A. & Taswell, C. (2024). The Multimedia FAIR Metrics Grand Challenge. *Brainiacs Journal* 5(1): G7ECAEAD9
- 2 published peer reviews of a published neuroscience article: Lu, G., Gong, C., Sun, Y., Qian, X., Rajendran Nair, D. S., Li, R., ... & Zhou, Q. (2024). Noninvasive imaging-guided ultrasonic neurostimulation with arbitrary 2D patterns and its application for high-quality vision restoration. *Nature Communications*, 15(1), 4481.

Results for initial test cases

Example	A_T	M_T	A_V	M_V	A_D	M_D	f_T	f_V	f_D	f_J
Synthetic example	1	0	1	0	0	1	1	1	-1	1/3
Craig & Taswell R1	2	2	0	0	0	0	0	0	0	0
Craig & Taswell R2	1	5	0	0	0	2	$-2/3$	0	-1	$-3/4$
Lu et al. R1	2	0	0	0	0	0	1	0	0	1
Lu et al. R2	7	0	0	0	0	0	1	0	0	1

Illustrating the evaluation with the synthetic example

- "This work is out of scope." the conclusion the reviewer must justify
- "It proposes a decision-tree-based expert system for retrieving drugs and drug targets relevant to a patient's symptoms."
  - This reviewer correctly attributes thia to the target manuscript.
- "The scope of this conference is biomedical applications of artificial intelligence (AI)."
  - The reviewer correctly attributes this to the conference website.
- "Human-curated decision trees are not AI (Wang, 2019)."
  - Source: Wang, P. (2019). On defining artificial intelligence. *Journal of Artificial General Intelligence*, 10(2), 1-37.
  - The source says, "To the larger community of computer science and information technology, AI is usually identified by the techniques grown from it, which at different periods may include theorem proving, heuristic search, game playing, expert systems[...]."
  - The reviewer misattributes their own narrow view of AI to the source.
- $A_T = 1$, $M_T = 0$, $A_V = 1$, $M_V = 0$, $A_D = 0$, $M_D = 1$
- $f_T = \frac{1-0}{1+0} = 1, f_V = \frac{1-0}{1+0} = 1, f_D = \frac{0-1}{0+1} = -1, f_J = \frac{1+1+0-0-0-1}{1+1+0+0+0+1} = \frac{1}{3}$

Selecting 10 new example cases from the brain informatics literature

- We searched 5 publishers known to practice open peer review:
  - BioMed Central and Nature Research (both owned by Springer Nature)
  - F1000Research
  - eLife Sciences Publications
  - MDPI
  - Frontiers
- We selected 10 articles, 2 from each publisher, concerning management of brain imaging data.
- Each article had at least 2 peer published reviews.
- We selected the first 2 reviews of each article.

Overview of results of FAIR Metrics analysis of the 10 new test cases

- In progress...

Common features of the brain informatics reviews

- Most comments are constructive.
- Most claims about content present in a work are accurate.
- Some reviewers overlook information authors provided and then complain that it is absent.
- Some treat common practices (e.g., organization of a paper into Introduction, Methods, Results, Discussion, and Conclusion sections) as absolute requirements for publication.
- Reviewers frequently invoke outside knowledge but rarely cite sources.
- Since journals only publish reviews of accepted articles, this may bias the sample toward more constructive reviews.

Recording FAIR Metrics analysis of peer review using the PDP-DREAM Ontology

- We previously designed a FAIR Metrics module of the PDP-DREAM Ontology with classes (e.g., "Document" and "Claim") and properties (e.g., "hasAttributionTo") to enable recording of a FAIR Metrics analysis in a resource description framework (RDF) document.
- We have now expanded it to include additional classes and properties useful for recording FAIR Metrics analyses of peer reviews.
- RDF record for synthetic example review: http://npds.portaldoors.net/nexus/fidentinus/Submission1Review1
- Also accessible via the NPDS curation app: https://portaldoors.net/NPDS/NexusService/AnonResreps/Diristry/Fidentinus/AnyAndAll/Nexus
- Source: Craig, A., & Taswell, C. (2024, September). FAIR Metrics for Motivating Excellence in Peer Review. In *2024 IEEE 20th International Conference on e-Science (e-Science)* (pp. 1-2). IEEE.

The Multimedia FAIR Metrics Grand Challenge

- FAIR Metrics analysis is labor-intensive compared to less methodical approaches to peer review and meta-review.
- We are organizing the Multimedia FAIR Metrics Grand Challenge to award cash prizes to developers who can automate some or all steps of the process, including but not limited to...
  - Extract the text of a work from different file formats, such as PDFs.
  - Separate the text into discrete statements.
  - Convert information in figures and tables into discrete statements.
  - Distinguish substantive claims from other statements.
  - Retrieve cited source documents of claims.
  - Search prior work for potential uncited sources of claims.
  - Distinguish whether two claims are equivalent in meaning.
- Craig, A. & Taswell, C. (2024). The Multimedia FAIR Metrics Grand Challenge. *Brainiacs Journal* 5(1): G7ECAEAD9

PORTAL-DOORS Project References

- Craig, A., Lee, C., Bala, N., & Taswell, C. (2022). Motivating and maintaining ethics, equity, effectiveness, efficiency, and expertise in peer review. *Brainiacs Journal* 3 (1):I5B147D9D
- Craig, A., Ambati, A., Dutta, S., Kowshik, P., Nori, S., Taswell, S. K., ... & Taswell, C. (2019, June). DREAM Principles and FAIR Metrics from the PORTAL-DOORS Project for the Semantic Web. In 2019 *11th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)* (pp. 1-10). IEEE.

PORTAL-DOORS Project References

- Craig, A., Athreya, A., & Taswell, C. (2023, October). Example evaluations of plagiarism cases using FAIR Metrics and the PDP-DREAM Ontology. In *2023 IEEE 19th International Conference on e-Science (e-Science)* (pp. 1-2). IEEE.
- Craig, A., & Taswell, C. (2024, September). FAIR Metrics for Motivating Excellence in Peer Review. In *2024 IEEE 20th International Conference on e-Science (e-Science)* (pp. 1-2).
- Craig, A. & Taswell, C. (2024). The Multimedia FAIR Metrics Grand Challenge. *Brainiacs Journal* 5(1): G7ECAEAD9

Other References

- Landkroon, A. P., Euser, A. M., Veeken, H., Hart, W., & Overbeke, A. J. P. (2006). Quality assessment of reviewers' reports using a simple instrument. *Obstetrics & Gynecology*, 108(4), 979-985.
- Dobele, A. R. (2015). Assessing the quality of feedback in the peer-review process. *Higher Education Research & Development*, 34(5), 853-868.
- Jefferson, T., Alderson, P., Wager, E., & Davidoff, F. (2002). Effects of editorial peer review: a systematic review. *JAMA*, 287(21), 2784-2786.
- Jefferson, T., Wager, E., & Davidoff, F. (2002). Measuring the quality of editorial peer review. *JAMA*, 287(21), 2786-2790.
- Wolfram, D., Wang, P., Hembree, A., & Park, H. (2020). Open peer review: promoting transparency in open science. *Scientometrics*, 125(2), 1033-1051.
- Wang, P. (2019). On defining artificial intelligence. *Journal of Artificial General Intelligence*, 10(2), 1-37.

Contact Info

- acraig@bhavi.us
- www.BHAVI.us
- www.BrainHealthAlliance.org