Paper quality assessment based on individual wisdom metrics from open peer review

Andrii Zahorodnii[1,2*], Jasper J.F. van den Bosch[3], Ian Charest[4], Christopher Summerfield[5], and Ila R. Fiete[2*]

[1]**Department of Electrical Engineering and Computer Science, MIT**
[2]**Department of Brain and Cognitive Sciences & McGovern Institute, MIT**
[3]**School of Psychology, University of Leeds, United Kingdom**
[4]**Département de Psychologie, Université de Montréal, Canada**
[5]**Department of Experimental Psychology, University of Oxford, United Kingdom**
[*]**Correspondence: zaho@mit.edu (AZ), fiete@mit.edu (IRF)**

ABSTRACT

This study proposes a data-driven framework for enhancing the accuracy and efficiency of scientific peer review through an open, bottom-up process that estimates reviewer quality. Traditional closed peer review systems, while essential for quality control, are often slow, costly, and subject to biases that can impede scientific progress. Here, we introduce a method that evaluates individual reviewer reliability by quantifying agreement with community consensus scores and applying Bayesian weighting to refine paper quality assessments. We analyze open peer review data from two major scientific conferences, and demonstrate that reviewer-specific quality scores significantly improve the reliability of paper quality estimation. Perhaps surprisingly, we find that reviewer quality scores are unrelated to authorship quality. Our model incorporates incentive structures to recognize high-quality reviewers and encourage broader coverage of submitted papers, thereby mitigating the common "rich-get-richer" pitfall of social media. These findings suggest that open peer review, with mechanisms for estimating and incentivizing reviewer quality, offers a scalable and equitable alternative for scientific publishing, with potential to enhance the speed, fairness, and transparency of the peer review process.

Keywords: Open peer review, post-publication review, paper quality assessment, Bayesian weighting, community consensus, scientific incentive structures

1 INTRODUCTION

Modern technologies have empowered the sharing of information at scale, as well as commentary and feedback on the shared content. However, distilling this collective feedback into a reliable collective assessment of shared information has remained a thorny challenge. The well-known and ubiquitous problem of mis- and dis-information on social media platforms is a testament to this difficulty (Vosoughi et al., 2018; Kitchens et al., 2020).

Scientific communication, the primary focus of the present work, suffers from related problems. Partially for historical reasons (Birukou et al., 2011) and partly to head off the kinds of misinformation problems found in open media, scientific review is a closed, slow, top-down, expensive process with incentives for high-profile publishing that might also distort science (Noorden, 2013; Young et al., 2008; Buranyi, 2017; da Silva and Dobránszki, 2014). A scientific report is submitted to one journal, whose editors select whether to desk-reject it or send it out for review to a small number of hand-picked reviewers. The resulting reviews emanate from a small number of reviewers and are thus highly stochastic. If the paper is rejected, the process repeats until the paper is accepted at some journal. This friction-filled process slows scientific progress in a way that can impede junior scientists' careers. The exponentially growing volume of scientific output makes these challenges more daunting. The concept of "publish then review", with bottom-up review, offers a potential antidote (Eisen et al., 2020; Ginsparg, 1997; Eisen, 2016; Kravitz and Baker, 2011; Kriegeskorte et al., 2012; Nosek and Bar-Anan, 2012; Stern and

O'Shea, 2019; da Silva, 2013; Walther and van den Bosch, 2012). Submitted papers are immediately considered as published, and review is an ongoing process that begins after publication. This process requires making reviewing open, equitable, and bottom-up, so that any scientist can review any paper, by self-selection (LeCun, 2013). However, how to make such an open system reliable, and with the right incentive mechanisms, is an open problem (Sandewall, 2012; Birukou et al., 2011; Pöschl, 2012; Goldberg et al., 2024).

In this paper, we take a data-driven approach to the question of how to extract a reliable signal from noisy and flawed data provided by the scientific community. We first quantify the level of agreement between reviewers in peer review data, and show that it is surprisingly small. Next, we propose a method for assessing paper quality based on an estimate of reviewer quality for each individual. Interestingly, reviewership quality is independent of authorship quality. We show that this metric can support an open, democratic peer-review solution that tackles the following challenges: It allows reviewers to self-organize by picking the works that they want to review. It provides recognition and thus incentives to reviewers for consistently writing high-quality reviews. In addition, it allows an effective extraction of the community belief about the quality of papers from scarce and noisy reviews.

2 RESULTS

2.1 High variability: A core challenge for collective paper review

We consider data from two conferences that employed a transparent review process and used explicit scoring for determining paper acceptance decisions: ICLR 2023 and the Conference on Cognitive Computational Neuroscience (CCN) 2023. In the CCN2023 peer review process, each of the 589 reviewers assigned "Impact" and "Clarity" numerical scores to the 527 submitted abstracts. All reviewers were authors of submitted abstracts. Each abstract received an average of 9.1 reviews. In ICLR2023, each of the 3798 submitted papers received anywhere from 2 to 9 reviews by reviewers selected through a separate track, with an average of 3.8 reviews per submission. Each review consisted of multiple dimensions of scores, with the overall rating and confidence scores (used in analyses in this paper).

In both datasets, the correlation in scores given by pairs of reviewers to the same paper was surprisingly low (Figure 1). The correlation was very weak in the community-based review process of CCN ($r = 0.161 \pm 0.014$; Figure 1A). Though somewhat higher for ICLR2023 ($r = 0.361 \pm 0.014$; Figure 1C), the correlations were still small despite the fact that the submissions being assessed were full-length standalone papers. Common possible causes of low correlation, such as variations in what range of scores each reviewer actually utilizes, are sometimes addressed by score normalization techniques (e.g., z-scoring, ranking, mean removal, or distribution inversion). Applying any of these methods only marginally increased the correlation ($r < 0.19$ for CCN2023; Figure 1B). Although the correlation increased with the self-reported confidence scores in ICLR2023 data (from $r = 0.24$ for confidence=2 up to $r = 0.49$ for confidence=5; Figure 1D), only 10.0% of reviewers reported having this high confidence. Moreover, the confidence of ratings was inversely correlated with the ratings scores (Supplementary Figure S1), implying that on average, better papers receive less confident reviews. In sum, our analysis reveals a striking lack of agreement among reviewers when assessing the same paper.

2.2 Bayesian paper quality estimation using reviewer quality estimatates

We first consider a very simple model (Figure 2; a slightly richer model is presented below). Assume that each submission (paper) j has a hidden ground truth quality q_j. Reviewers are assumed to give scores to that paper drawn from a distribution centered at q_j, with a user-specific standard deviation σ_i, where i is the index of the reviewer (Figure 2A). We can think of the inverse standard deviation of a reviewer as a ground-truth reviewer quality measure. The goal is to obtain good estimates of q_j without knowledge of the paper ground truth values. This simple model leaves out possible systematic reviewer biases (see Shah et al. (2017); Helmer et al. (2017); Ross (2017); Stelmakh et al. (2019); Goldberg et al. (2024); Stelmakh et al. (2023) for discussion of biases in the review process), assumes there are no bots or bad actors (we add bots below), and doesn't constrain the scores to lie in a bounded interval, however it suffices to demonstrate the core idea. Traditionally, the scores of all reviewers are averaged to obtain a final estimated quality score of the paper, on which decisions are made. This calculation produces an estimator of the paper quality score with mean squared deviation (MSD) equal to

$$\text{MSD(simple mean)} = \frac{\sum \sigma_i^2}{n^2}. \tag{1}$$

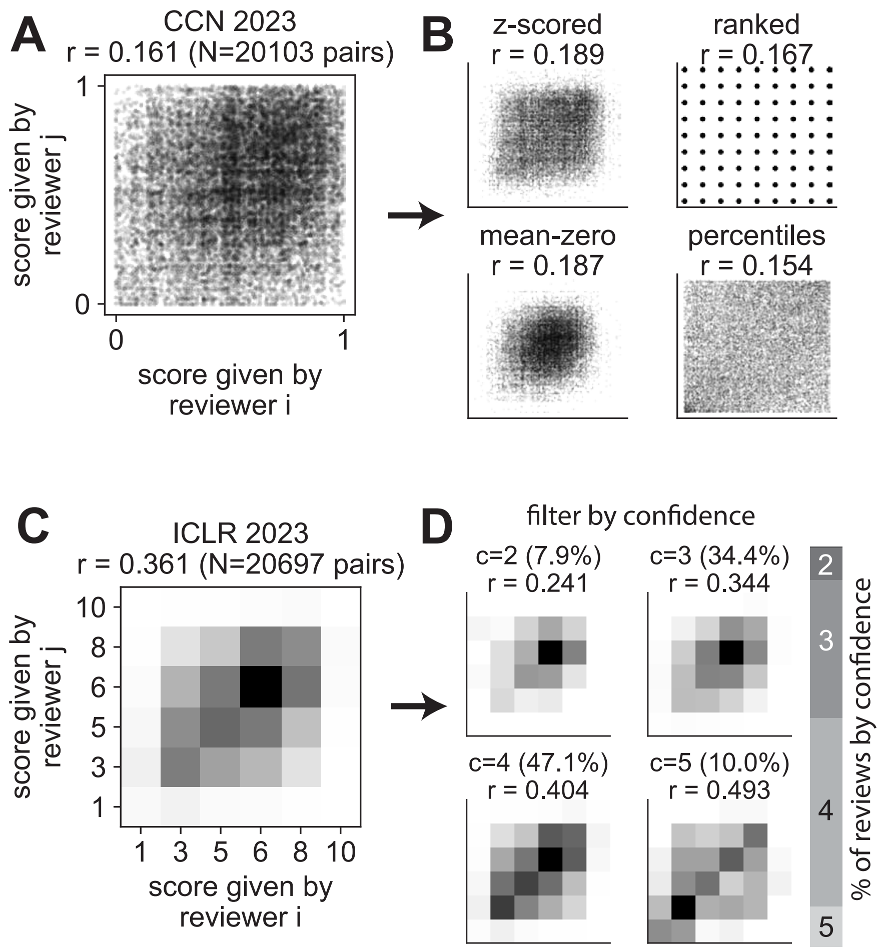


Figure 1. **High variability in paper scores and low correlation across reviewers are challenges for collective paper review.** (A) Correlation of scores between pairs of reviewers reviewing the same submission in the conference Cognitive Computational Neuroscience 2023 ("Impact" scores). (B) Common score-normalizing techniques – z-scoring, ranking, removing the mean, and inverting the distribution, applied to the CCN2023 data: the correlations remain low. (C) Corresponding to (A) analysis for the International Conference on Learning Representations 2023. (D) In ICLR2023 data, pairs of ratings with higher self-reported confidence scores have a higher correlation, but only a minority of reviews is tagged as high-confidence. For all panels showing correlation, $p < 0.001$.

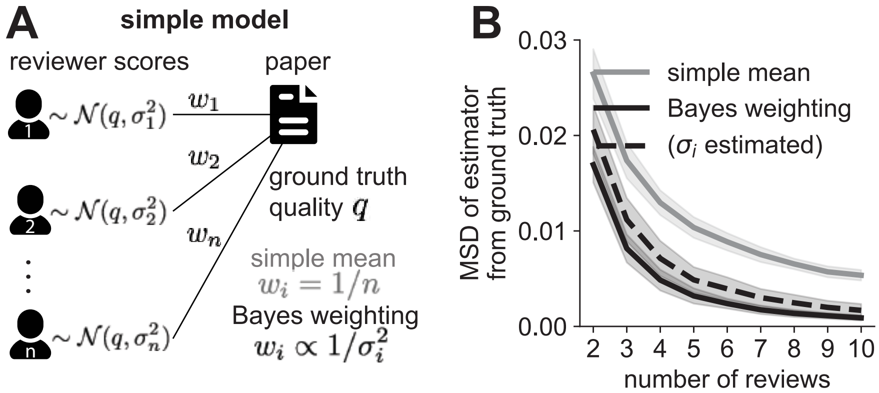


Figure 2. In a simple model, weighting the reviews using a reviewer quality metric results in a better estimation of paper quality. (A) The simple model: different reviewers all assign scores to the same paper according to a distribution centered at the ground truth quality (hidden), but with different standard deviations, corresponding to different reviewer accuracies. (B) Bayes-optimal weighting of review scores leads to a tighter estimation of the ground truth paper quality scores (reviewer accuracies either taken from ground truth or estimated from five previous reviews). Shaded regions denote the s.d.

If reviewer quality is known, the optimal weighting technique, a Bayesian approach, assigns weights to reviewers' scores that are proportional to the inverse square of the estimated standard deviation of that reviewer (for derivation of these results, see Appendix 1): $w_i \propto 1/\sigma_i^2$. This approach leads to a paper quality estimator with

$$\mathrm{MSD(Bayesian)} = \frac{1}{\sum 1/\sigma_i^2}. \tag{2}$$

This Bayesian paper quality estimator based on known reviewer quality much more tightly follows the ground truth quality of the paper (Figure 2B, black) than the standard mean (Figure 2B, gray).

In the real setting, reviewers' standard deviations are not known a priori, requiring that they are estimated from the data. The Bayesian MSD using an empirically estimated reviewer quality still significantly outperforms the standard mean (Figure 2B, black dashed line), suggesting that using estimates of individual reviewer quality can improve estimates of paper quality.

2.3 Assessing the Bayesian approach in CCN 2023 review data

The above result motivates applying a Bayesian weighting of paper reviews by estimated reviewer quality in a real world review process. To assess outcomes, we again turn to the CCN2023 process (Figure 3). In the experiments that follow, we used the CCN2023 ”Impact” scores.

Unfortunately, in such real world datasets we lack ground truth knowledge of paper quality to assess the robustness of different measures. To construct a proxy, we assume that the full community average score (CAS) for a paper is the ground truth. The CAS represents the score that would be assigned if many scientists in the community were to review the paper. With an estimate for individual reviewer quality, we can assess the efficacy of different metrics on subsampled sets of reviews for each paper.

CCN2023 data contained unique reviewer IDs for each reviewer, across papers, thus we were able to compute reviewer standard deviations relative to the CAS and generate an estimate of reviewer quality. Unlike in our simple model, scientists in CCN2023 play two roles: they are reviewers and authors. A quality value can be assigned to each role. Given the sparsity of the data (each individual writes relatively few reviews and each paper is reviewed by relatively few reviewers), a natural question is whether quality as an author is related to quality as a reviewer and thus whether it might be possible to arrive at a better per-individual reviewership score by combining both.

Is authorship quality predictive of reviewer quality? We considered the mean squared deviation of an individual's scores from the CAS each paper reviewed by that individual (i.e., their quality as a reviewer), and compared it with the CAS they received for their own submissions (i.e., their quality as an author).

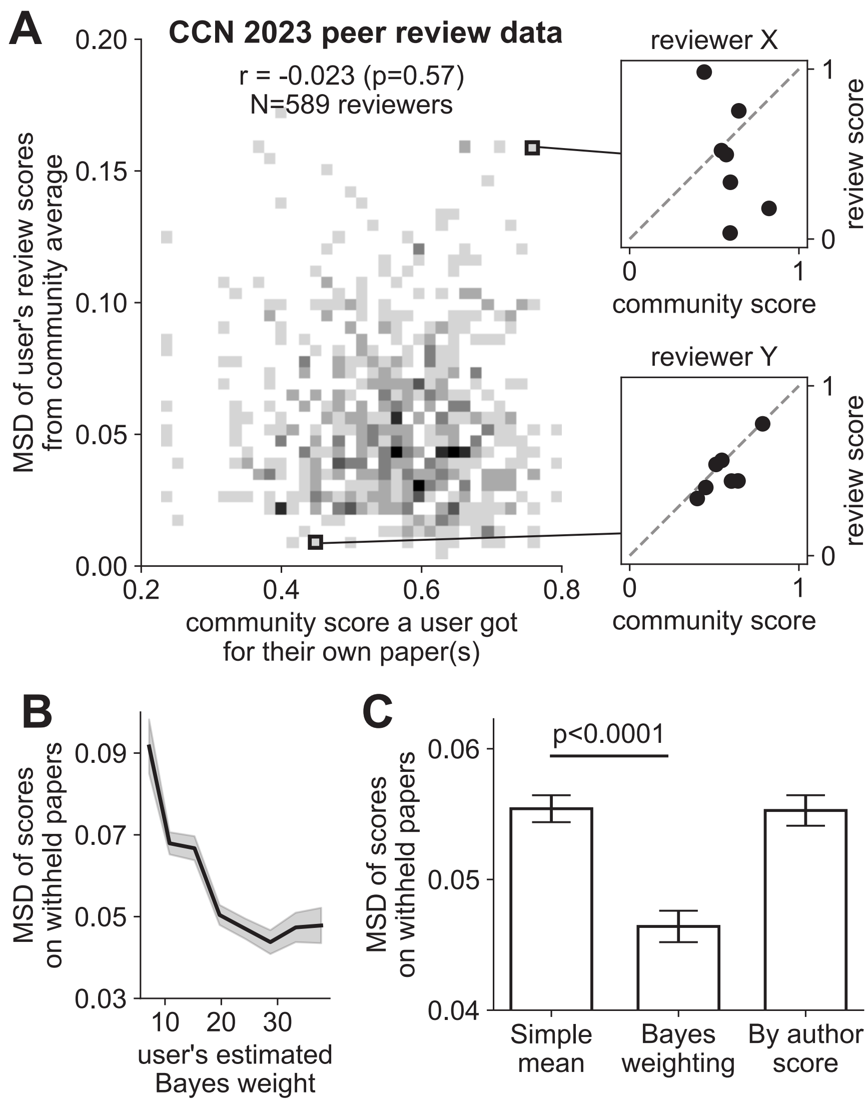


Figure 3. **The reviewer quality metric, unrelated to authorship quality, improved community score estimation of CCN2023 submissions.** (A) In the CCN2023 review process, reviewer quality (measured as MSD of user's review scores from community average score across all reviewed papers) and author quality (measured as average score the user got for their paper) were unrelated. To preserve reviewer privacy in this visualization, some outlier points were removed from the plot and reviewers were binned together in small bins. (B) Users with higher estimated reviewer quality (Bayes weight), when estimating paper quality, consistently had smaller deviations from community score on withheld papers. Shaded regions denote the s.e.m. (C) Weighting by the inferred Bayes-optimal coefficient resulted in a significantly reduced mean squared deviation from the community score on withheld papers. By contrast, weighting by the authorship quality ("author score") did not produce benefits. Error bars denote the s.e.m.

This resulted in the surprising discovery that authorship quality is unrelated to reviewership quality (Figure 3A; $r = -0.023, p = 0.57, N = 539$ reviewers). The finding challenges the assumption that high-scoring authors are necessarily good at predicting the community judgment of a paper's quality. Instead, there were reviewers who received very high scores for their own submissions yet were remarkably poor at predicting the community judgment of papers they reviewed (Figure 3A, inset, upper right). Conversely, some reviewers who received much lower scores for their own submissions were able to estimate the CAS quality of reviewed papers much more accurately (Figure 3A, inset, lower right). In fact, the group of reviewers that was able to estimate the CAS with the highest accuracy was the group with intermediate authorship scores (Figure 3, Supplementary Figure S2A-B), and this group was also had the highest level of agreement between reviewers reviewing the same paper (Figure 3, Supplementary Figure S3; $r = 0.352$ for the middle percentiles, N=795 pairs). This discrepancy between the authorship score and the reviewer quality might be attributable to various factors, including a lack of rewards and recognition for putting effort into conducting a thorough review.

In sum, the reviewer quality metric is independent of the author quality metric, and author quality likely cannot be used as a way to augment reviewer quality scores.

Improving paper assessments through empirical reviewer quality estimates Individuals with a higher reviewer quality score consistently demonstrated a better ability to predict the CAS score. This is a cross-validated result, shown for withheld papers – those not included in the estimation of the users' reviewer quality scores, thus avoiding information leakage (Figure 3B).

All reviewers can be combined to obtain a better prediction of the CAS than simple averaging or averaging with traditional normalizers, even when working with a limited number of reviews (Figure 3C; from $MSD = 0.0554 \pm 0.001$ for the simple mean estimator to $MSD = 0.0464 \pm 0.001$ for the Bayes weighting procedure). On the contrary, using the author quality score instead of the reviewer quality score does not lead to a significant improvement ($MSD = 0.0553 \pm 0.001$).

This improved estimation technique has the potential to enhance the overall quality and fairness of the peer review process, particularly when the number of reviews per paper is limited. We were only able to test this hypothesis in the CCN 2023 dataset, because the ICLR open process does not provide a consistent reviewer identifier across the user's reviews and submissions, which made it impossible to aggregate data on the performance of the reviewer across more than just one review.

2.4 How things could be: an open framework for peer review

Based on the findings above, we propose a new open framework for scientific peer review (Figure 4) and assess its soundness in a computational model. In this framework, all papers are immediately published (Eisen et al., 2020). Post-publication, users on the platform self-select themselves to review the paper. In this framework, users may give quality ratings not only to papers but also to others' reviews (Walther and van den Bosch, 2012). The quality of the given publication is estimated from these reviews and the ratings of the reviews. Authors of new submissions, reviewers, and raters of reviewers all come from the same user pool.

Generative model We implement a generative model of this process (Figure 4A): The model consists of a set of papers and a set of reviewers. Each generated paper i has a hidden ground truth quality $q_i \in (0,1)$. Each reviewer has a ground truth reviewer quality $p_i \in (0,1)$. Whenever reviewer i writes a review for paper j, they assign it a score $\sim N(q_j, \alpha/p_i)$: in other words, the score is based on the intrinsic paper quality, with a standard deviation that is inversely proportional to reviewer quality. Whenever user i rates a review of another user j, they assign a score to them which is $\sim N(p_j, \alpha/p_i)$. We set $\alpha = 0.18$ to match the reviewer correlation levels to those we found in CCN2023 data. We set the standard deviations of the review scores and paper scores to be equal, based on the finding from a NeurIPS 2022 randomized control trial (Goldberg et al., 2024) that disagreement between reviewers assessing review quality are comparable to the disagreement rates of paper reviewers. Varying α does not change the results qualitatively. In addition, we consider the existence of consistently low-quality, unreliable reviewers or "bots". These unreliable users are modeled as assigning paper review scores and ratings of other reviews uniformly at random in$[0,1]$, regardless of paper or review quality.

All assigned scores must lie in the interval $[0,1]$, and user-assigned scores are the only data available for the system to derive its quality estimators. As a worst-case conservative scenario, we consider a batch/offnline process, in which users have provided reviews of papers and other users, and we must

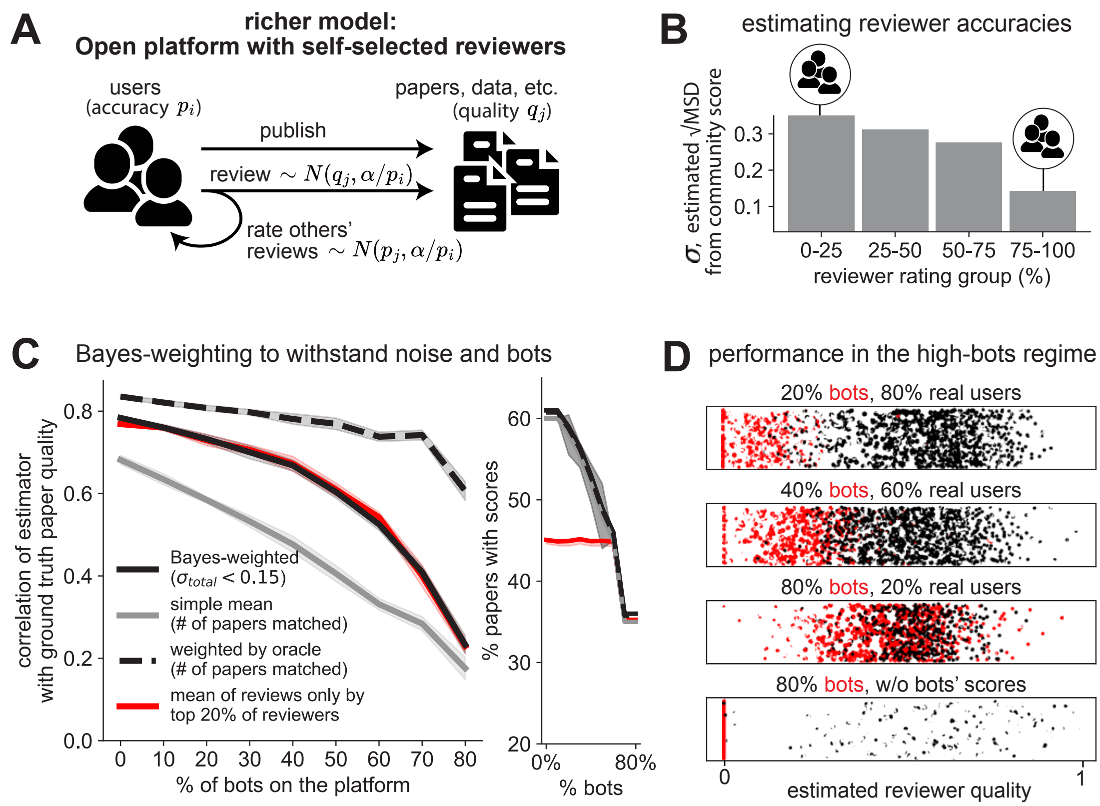


Figure 4. **Benefits of reviewer quality metrics in a richer ecosystem.** (A) The conception of an open platform, where content is published immediately, and scored by reviewers that are self-selected from the user pool, and assessed on the review quality by the same community. (B) Binning reviewers by rating scores allows the estimation of the accuracy of any group in assessing the community belief about the quality of any given submission. (C) Bayes weighting of review scores based on the estimated reviewer accuracy metric from (B) leads to a better estimate of the ground truth paper quality scores (left; solid black line) than simple averaging of all review scores (left; gray line). An alternative heuristic of simply thresholding out all users below a certain rating percentile (here, below 80%, red line) results in a similar correlation to Bayes weighting but lower coverage of papers on the platform (right). (D) Dsitributions of estimated reviewer quality scores for bots and real users in the simulated platform. In high-noise regime with most of the users being bots (up to 80%), reliability of the reviewer quality estimation breaks down, suggesting that alternative ways of excluding bots might be more effective.

estimate all quantities (paper quality and reveiwer quality) at the same time: scoring is based on all available papers and reviews available at the instantaneous time. (A history-dependent process, which we discuss below, may enable further accuracy gains.) We further assume, as a conservative poor-case scenario and consistent with the CCN 2023 dataset, that we are in a regime where assessing quality is hardest: the low-review regime in which each user publishes one paper a year, reviews ≤ 3 papers and rates ≤ 10 other users, meaning that each paper receives 3 reviews on average.

Scoring Scoring proceeds as follows: a rating is calculated for each reviewer by averaging all scores that this reviewer received for their reviews from other users. We then bin reviewers by their reviewer rating, and estimate how closely each bin predicts the CAS for any submission (Figure 4B). We also estimate the mean squared deviation (MSD) of each bin's scores around the CAS, on the subset of papers on the platform for which the CAS is known with high certainty (in this example we use the top 20% reviewed papers in terms of the number of reviews received, when only including reviews by reviewers which have a rating in the top 20% percentile).

The MSD of the bin is used as the MSD for each reviewer in the bin. The binning procedure helps to mitigate the problem that each reviewer generally scores very few papers and each paper generally receives very few reviews, for accurate assessment of MSD. In addition, the binning removes the incentive to copy other reviewers' assessments, preventing reviewers from artificially inflating the measure of others' agreement with their scores. With this estimated MSD for each reviewer, we can now use Bayes weighting, as in the simple model of Figure 2, to generate a quality score for each article on the platform based on the reviews received.

Computing the MSD of reviewers allows for the generation of an estimation certainty score for each paper, based on the equation (derived in Appendix 1)

$$1/\sigma_{\text{total}}^2 = \sum 1/\sigma_i^2. \tag{3}$$

This weighting procedure, combined with only committing to publish the scores that pass a certain threshold of certainty (here – $\sigma_{\text{total}} < 0.15$) results in an estimator of paper quality (Figure 4C; solid black line) that tracks the (hidden) ground truth quality of papers on the platform more closely than the simple mean estimator (Figure 4C; solid gray line).

Unreliable reviewers Unreliable reviewers, defined for our purposes as users who assign random reviews that are uncorrelated with user and paper quality, may consist of different populations: they could be scientists who do not invest time or effort in the review process, or they could be adversarial human or even non-human ("bot") actors.

The standard way to reduce the influence of non-human bots is to require all users who deposit papers or review papers to be verified scientists (Walther and van den Bosch, 2012). Such a step would be essential because the number of non-human bots can overwhelm the number of human users by orders of magnitude. After this step, in the low-review regime with the possibility of unreliable human users, imposing a constraint on minimum certainty before publishing a score results in substantial robustness to contamination with noise and unreliable reviewers, as done in Figure 4C.

In addition, given that there are two qualitatively distinct types of reviewers in the system (those who issue scores with some correlation to paper quality and those whose outputs are uncorrelated), we consider entirely thresholding out (removing from consideration) the bottom percentiles of reviewers instead of weighting their reviews by the Bayesian MSD review quality factor, reasoning that this bin will contain mostly the uncorrelated reviewers. When we compute the correlation of the resulting paper quality scores with ground-truth paper quality, we find that, surprisingly, thresholding reviewers by their reviewer quality does not yield a prediction improvement over the full Bayesian weighting (Figure 4C, solid red line). As a sanity check, if we directly use a reviewer quality "oracle" estimator, based on directly estimating the correlation of paper scores assigned by the reviewer to ground truth paper scores, and threshold reviewers out based on this oracular reviewer quality estimate, our reasoning holds and thresholding significantly improves paper quality estimation (Figure 4C, dashed black line).

The cause of the discrepancy between the oracular review quality thresholding and the normally estimated review quality thresholding result becomes clear when we examine the distributions of estimated reviewer quality, Figure 4D: through the unreliable reviewers have a ground-truth quality of 0, as they become a larger fraction of the reviewing population, their reviews lead other unreliable reviewers to have a broad distribution of reviewer quality scores, meaning they are no longer confined to the lower quality bins. Thresholding out the lowest reviewer quality bins does not target only or the majority of the actual low-quality reviewers. This result suggests that it might be possible to further improve the fidelity of review with a "warm start" or online reviewing process, in which the platform begins with high-quality reviewers. New reviewers can join, but they must work their way up (in the sense of obtaining above-threshold review quality scores from estabilished users) for their reviews to influence paper quality scores. Future work will examine a model that includes such a warm-start process.

Reviewer Quality Incentives In addition to improving estimates of paper quality on the platform, we expect the quantification of reviewer accuracies to incentivize reviewers to write high-quality reviews, because it would place better reviewers in a higher reviewer quality bin, and weigh their decisions about future papers more heavily (Figure 5A). The inverse of the reviewer MSD is also a reviewer quality score that can be used to recognize not only the work of the reviewer but also the quality of their work. The reviewer quality score could be used to provide reviewer credit in different ways, including for promotion or tenure decisions, for public display via a quality certificate issued to the reviewer, and so on. This recognition and accountability would provide a second incentive for writing higher-quality reviews (Pascual-Ezama et al., 2015).

Paper Coverage Incentives The estimation certainty measure (Eq. 3) can be used to incentivize reviewers to provide better coverage for papers whose quality certainty is low. In a simple experiment, we consider that reviewers select papers to review based on the Chinese restaurant process prior (CRP), a stochastic process where the probability of a paper being selected is proportional to how many reviews the paper already has, in other words its popularity (a typical rich-get-richer process; Figure 5B, left). Next, we define a reward signal, which we set to be equal to $\Delta\sigma^2_{\text{total}}$ - the *amount of reduction* in the uncertainty in the paper's score that can be provided by this reviewer. Reviewers are rewarded for reviewing papers for which they can provide a strong boost in score estimation certainty. We assume that through offering platform incentives and rewards, the Chinese restaurant process probability of a reviewer choosing to review a certain paper can be modulated by this reward score. We find that such a process can avoid the typical snowballing dynamics of popularity that are evident in many social media systems (Narayanan, 2023), and lead to a broader coverage of reviewed papers across the whole simulated platform (Figure 5B, right).

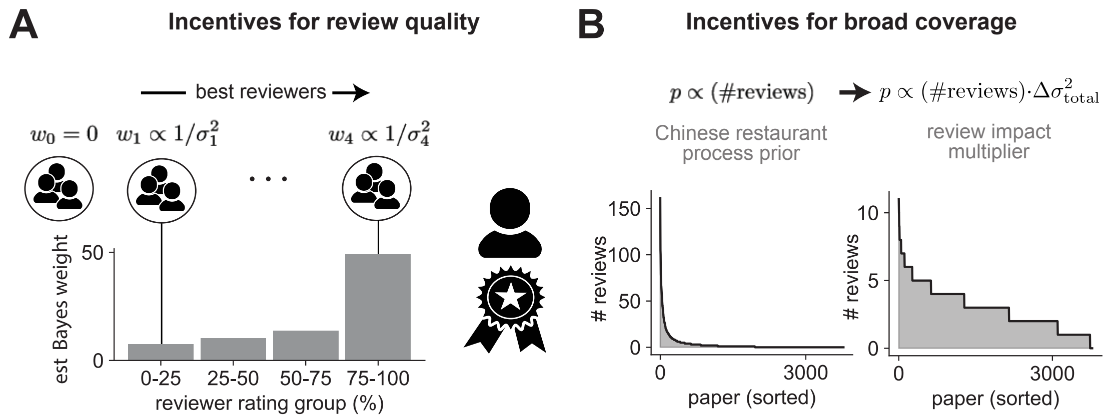


Figure 5. Embedding incentives for consistently writing high-quality reviews and ensuring broad coverage within the open platform framework. (A) *Left:* Bayes weighting allows reviewers with higher ratings have more influence on the final paper score estimation, which provides one type of incentive for providing quality reviews. *Right:* Users can publicly advertise and use for promotion and tenure their reviewer ratings, another incentive for high-quality review. (B) *Left:* Distribution of reviews per paper, assuming papers accrue reviews according to a Chinese restaurant process (CRP). *Right:* Rewarding reviewers based on their contribution to reducing the uncertainty in the reviewed paper's score estimate can provide , typical snowballing dynamics of popularity, evident in any form of social media, may be softened or avoided.

3 DISCUSSION

Summary We evaluated two real-world peer-review datasets to determine how well reviewer scores correlate with each other, and found very low levels of agreement between reviewers. In one of these datasets, we made an estimate of paper quality for a subset of papers that had a large number of reviews based on the community average score, and used this as a metric for assessing both authorship quality and reviewer quality for each individual reviewer. We found that the authorship quality of an individual is not predictive of their reviewer quality.

Given these quantified challenges in extracting meaningful paper quality scores from small numbers of noisy reviewers, we proposed and assessed a Bayesian weighting of reviewer scores based on empirically estimating reviewer quality. In the low number of reviews per paper regime, the proposed measure significantly outperforms standard averaging methods. We mitigated the problem of estimating reviewer quality given the small numbers of reviews written by each individual with a method for binned estimation of reviewer quality.

Finally, we showed that generating reviewer scores can incentivize reviewers to produce high-quality reviews in two ways: a desire for impact (a reviewer's assessment of a paper is weighted more heavily if they have a higher reviewer score), and a desire for recognition (a reviewer's work can be visibly recognized in the form of a reviewer score that can be used for promotion and tenure decisions and as a badge of honor). On top of such an incentive structure, we explored the possibility of incentivizing reviewers to cover a larger set of papers, avoiding the rich-get-richer phenomenon in which a few papers get heavily reviewed while most are neglected.

Related work Open peer review frameworks for scientific publishing, where reviews can be made public and evaluated, have been widely proposed over the past decade (Eisen et al., 2020; Ginsparg, 1997; Eisen, 2016; Kravitz and Baker, 2011; Kriegeskorte et al., 2012; Nosek and Bar-Anan, 2012; Stern and O'Shea, 2019; da Silva, 2013; LeCun, 2013; Walther and van den Bosch, 2012). The general idea of users generating commentary and then allowing users to rate the commentary of others has been implemented to considerable success in platforms like StackOverflow and Reddit, with its drawbacks investigated (Davis and Graham, 2021; Mazloomzadeh et al., 2024; Melnikov et al., 2018; Movshovitz-Attias et al., 2013; Wang et al., 2021). Specifically in the field of scientific peer review, Goldberg et al. (2024) presented a randomized control trial, carried out at NeurIPS 2022, which showcases the measures of inter-evaluator disagreement, miscalibration, subjectivity, and biases in peer reviews of peer reviews. Rating user comments is considered a low-bandwidth and simpler exercise than generating commentary. We find that it provides a useful signal for open peer review that simultaneously enables positive incentives as well as more accurate community belief estimation. Recommendation systems in general and specifically for disseminating scientific research are likewise being investigated (Isinkaye et al., 2015; Bai et al., 2020; Achakulvisut et al., 2016; Ko et al., 2022; Pinedo et al., 2024; Putra Utama et al., 2023; Shani and Gunawardana, 2011). In general settings, crowd-based metrics have been proposed to extract the wisdom of the crowd from the noisy and sometimes on average incorrect signal (Prelec et al., 2017; Kameda et al., 2022; Lee, 2024; Palley and Soll, 2019).

Outlook The volume of scientific publication has been increasing exponentially, and the difficulty of securing accurate high-quality paper assessments, in a timely way, has grown in tandem (Shah et al., 2017; Kuznetsov et al., 2024). The scientific community has shied away from bottom-up peer review to mitigate the potential for low-quality assessments. We conceive of the process as a dynamical system, in which the right mixture of estimation and incenitives can produce self-organized high-quality reviewing ecosystems. Here, we have taken small steps to explore possible metrics and incentives. Implementing and testing these ideas in open publishing platforms should help to provide real-world tests and improvements of such methods, possibly leading to a new process for scientific publishing.

ACKNOWLEDGEMENTS

All authors would like to thank Laurence Hunt, Mariya Toneva, Steven Scholte, and the organizing committee of the 2023 Conference on Cognitive Computational Neuroscience for their support and feedback throughout the project. A.Z. expresses gratitude to Professors James DiCarlo and Rebecca Saxe from the Department of Brain and Cognitive Sciences at MIT for their feedback, guidance and inspiration in early stages of this project.

REFERENCES

Achakulvisut, T., Acuna, D. E., Ruangrong, T., and Kording, K. (2016). Science Concierge: A Fast Content-Based Recommendation System for Scientific Publications. *PLOS ONE*, 11(7):e0158423.

Bai, X., Wang, M., Lee, I., Yang, Z., Kong, X., and Xia, F. (2020). Scientific paper recommendation: A survey.

Birukou, A., Wakeling, J. R., Bartolini, C., Casati, F., Marchese, M., Mirylenka, K., Osman, N., Ragone, A., Sierra, C., and Wassef, A. (2011). Alternatives to peer review: Novel approaches for research evaluation. *Front. Comput. Neurosci.*, 5.

Buranyi, S. (2017). *Is the staggeringly profitable business of scientific publishing bad for science?* Guardian.

da Silva, J. A. T. (2013). The need for post-publication peer review in plant science publishing. *Frontiers in Plant Science*, 4.

da Silva, J. A. T. and Dobránszki, J. (2014). Problems with traditional science publishing and finding a wider niche for post-publication peer review. *Accountability in Research*, 22(1):22–40.

Davis, J. L. and Graham, T. (2021). Emotional consequences and attention rewards: The social effects of ratings on Reddit. *Information, Communication & Society*, 24(5):649–666.

Eisen, M. B. (2016). *Coupling pre-prints and post-publication peer review for fast, cheap, fair, and effective science publishing.* http://www.michaeleisen.org/blog/?p=1820.

Eisen, M. B., Akhmanova, A., Behrens, T. E., Harper, D. M., Weigel, D., and Zaidi, M. (2020). Implementing a "publish, then review" model of publishing. *eLife*, 9:e64910.

Ginsparg, P. (1997). Winners and losers in the global research village. *The Serials Librarian*, 30(3-4):83–95.

Goldberg, A., Stelmakh, I., Cho, K., Oh, A., Agarwal, A., Belgrave, D., and Shah, N. B. (2024). Peer reviews of peer reviews: A randomized controlled trial and other experiments.

Helmer, M., Schottdorf, M., Neef, A., and Battaglia, D. (2017). Gender bias in scholarly peer review. *eLife*, 6.

Isinkaye, F., Folajimi, Y., and Ojokoh, B. (2015). Recommendation systems: Principles, methods and evaluation. *Egyptian Informatics Journal*, 16(3):261–273.

Kameda, T., Toyokawa, W., and Tindale, R. S. (2022). Information aggregation and collective intelligence beyond the wisdom of crowds. *Nature Reviews Psychology*, 1(6):345–357.

Kitchens, B., Johnson, S. L., and Gray, P. (2020). UNDERSTANDING ECHO CHAMBERS AND FILTER BUBBLES: THE IMPACT OF SOCIAL MEDIA ON DIVERSIFICATION AND PARTISAN SHIFTS IN NEWS CONSUMPTION. *Miss. Q.*, 44(4):1619–1649.

Ko, H., Lee, S., Park, Y., and Choi, A. (2022). A Survey of Recommendation Systems: Recommendation Models, Techniques, and Application Fields. *Electronics*, 11(1):141.

Kravitz, D. J. and Baker, C. I. (2011). Toward a new model of scientific publishing: Discussion and a proposal. *Frontiers in Computational Neuroscience*, 5.

Kriegeskorte, N., Walther, A., and Deca, D. (2012). An emerging consensus for open evaluation: 18 visions for the future of scientific publishing. *Frontiers in Computational Neuroscience*, 6.

Kuznetsov, I., Afzal, O. M., Dercksen, K., Dycke, N., Goldberg, A., Hope, T., Hovy, D., Kummerfeld, J. K., Lauscher, A., Leyton-Brown, K., Lu, S., Mausam, Mieskes, M., Névéol, A., Pruthi, D., Qu, L., Schwartz, R., Smith, N. A., Solorio, T., Wang, J., Zhu, X., Rogers, A., Shah, N. B., and Gurevych, I. (2024). What can natural language processing do for peer review?

LeCun, Y. (2013). *A New Publishing Model in Computer Science*. http://yann.lecun.com/ex/pamphlets/publishing-models.html.

Lee, M. D. (2024). Using Cognitive Models to Improve the Wisdom of the Crowd. *Current Directions in Psychological Science*, 33(5):308–316.

Mazloomzadeh, I., Uddin, G., Khomh, F., and Sami, A. (2024). Reputation Gaming in Stack Overflow.

Melnikov, A., Lee, J., Rivera, V., Mazzara, M., and Longo, L. (2018). Towards Dynamic Interaction-Based Reputation Models. In *2018 IEEE 32nd International Conference on Advanced Information Networking and Applications (AINA)*, pages 422–428.

Movshovitz-Attias, D., Movshovitz-Attias, Y., Steenkiste, P., and Faloutsos, C. (2013). Analysis of the reputation system and user contributions on a question answering website: StackOverflow. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, pages 886–893, New York, NY, USA. Association for Computing Machinery.

Narayanan, A. (2023). Understanding social media recommendation algorithms.

Noorden, R. V. (2013). Open access: The true cost of science publishing. *Nature*, 495(7442):426–429.

Nosek, B. A. and Bar-Anan, Y. (2012). Scientific utopia: I. opening scientific communication. *Psychological Inquiry*, 23(3):217–243.

Palley, A. B. and Soll, J. B. (2019). Extracting the Wisdom of Crowds When Information Is Shared. *Management Science*, 65(5):2291–2309.

Pascual-Ezama, D., Dunfield, D., de Liaño, B. G.-G., and Prelec, D. (2015). Peer Effects in Unethical Behavior: Standing or Reputation? *PLOS ONE*, 10(4):e0122305.

Pinedo, I., Larrañaga, M., and Arruarte, A. (2024). ArZiGo: A recommendation system for scientific articles. *Information Systems*, 122:102367.

Prelec, D., Seung, H. S., and McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, 541(7638):532–535.

Putra Utama, F., Mardiansyah, T., Faurina, R., and Vatresia, A. (2023). SCIENTIFIC ARTICLES RECOMMENDATION SYSTEM BASED ON USER'S RELATEDNESS USING ITEM-BASED

COLLABORATIVE FILTERING METHOD. *Jurnal Teknik Informatika (Jutif)*, 4(3):467–475.
Pöschl, U. (2012). Multi-stage open peer review: Scientific evaluation integrating the strengths of traditional peer review with the virtues of transparency and self-regulation. *Front. Comput. Neurosci.*, 6.
Ross, E. (2017). Gender bias distorts peer review across fields. *Nature*.
Sandewall, E. (2012). Maintaining live discussion in two-stage open peer review. *Front. Comput. Neurosci.*, 6.
Shah, N. B., Tabibian, B., Muandet, K., Guyon, I., and von Luxburg, U. (2017). Design and analysis of the nips 2016 review process.
Shani, G. and Gunawardana, A. (2011). Evaluating Recommendation Systems. In Ricci, F., Rokach, L., Shapira, B., and Kantor, P. B., editors, *Recommender Systems Handbook*, pages 257–297. Springer US, Boston, MA.
Stelmakh, I., Rastogi, C., Liu, R., Chawla, S., Echenique, F., and Shah, N. B. (2023). Cite-seeing and reviewing: A study on citation bias in peer review. *PLoS ONE*, 18(7):e0283980.
Stelmakh, I., Shah, N. B., and Singh, A. (2019). On testing for biases in peer review.
Stern, B. M. and O'Shea, E. K. (2019). A proposal for the future of scientific publishing in the life sciences. *PLOS Biology*, 17(2):e3000116.
Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380):1146–1151.
Walther, A. and van den Bosch, J. J. (2012). Fose: a framework for open science evaluation. *Frontiers in Computational Neuroscience*, 6.
Wang, S., German, D. M., Chen, T.-H., Tian, Y., and Hassan, A. E. (2021). Is reputation on Stack Overflow always a good indicator for users' expertise? No! In *2021 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pages 614–618.
Young, N. S., Ioannidis, J. P. A., and Al-Ubaydli, O. (2008). Why current publication practices may distort science. *PLoS Medicine*, 5(10):e201.

CODE AVAILABILITY

All code associated with analyses in this paper will be made available upon publication.

APPENDIX 1. MSD DERIVATIONS

MSD of the Simple Mean

Reviewer i's score for paper j is given by: $s_{ij} = q_j + \varepsilon_{ij}$, where $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_i^2)$. The simple mean estimator for the paper's quality is:

$$\hat{q}_j^{\text{simple}} = \frac{1}{n} \sum_{i=1}^{n} s_{ij}.$$

The mean squared deviation (MSD) is the expected squared difference between the estimator and the true quality:

$$\text{MSD(simple mean)} = \mathbb{E}\left[\left(\frac{1}{n} \sum_{i=1}^{n} \varepsilon_{ij} \right)^2 \right].$$

Since ε_{ij}'s are independent, the variance of the estimator is:

$$\text{MSD(simple mean)} = \frac{1}{n^2} \sum_{i=1}^{n} \sigma_i^2$$

MSD of the Bayes-Optimal Estimator

The Bayes-optimal estimator weights reviewers by $w_i \propto 1/\sigma_i^2$, so:

$$w_i = \frac{1/\sigma_i^2}{\sum_{k=1}^{n} 1/\sigma_k^2}.$$

The Bayes-optimal estimator is given by:

$$\hat{q}_j^{\text{Bayes}} = \sum_{i=1}^{n} w_i s_{ij}.$$

The MSD is the expected squared difference:

$$\text{MSD(Bayes-optimal)} = \mathbb{E}\left[\left(\sum_{i=1}^{n} w_i \varepsilon_{ij}\right)^2\right].$$

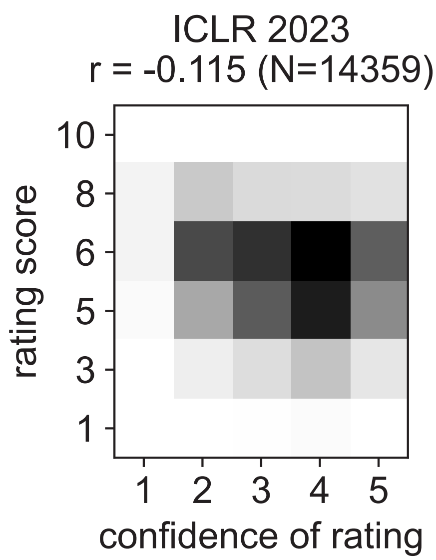
The variance of this estimator is:

$$\text{MSD(Bayes-optimal)} = \sum_{i=1}^{n} w_i^2 \sigma_i^2.$$

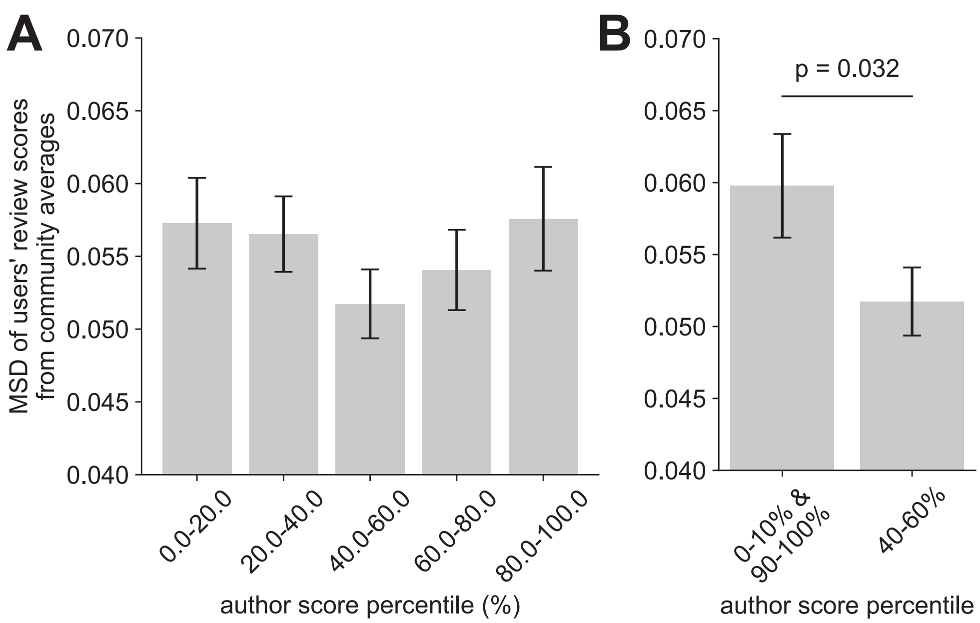
Substituting $w_i = \frac{1/\sigma_i^2}{\sum_{k=1}^{n} 1/\sigma_k^2}$ gives:

$$\text{MSD(Bayes-optimal)} = \frac{1}{\sum_{i=1}^{n} 1/\sigma_i^2}.$$

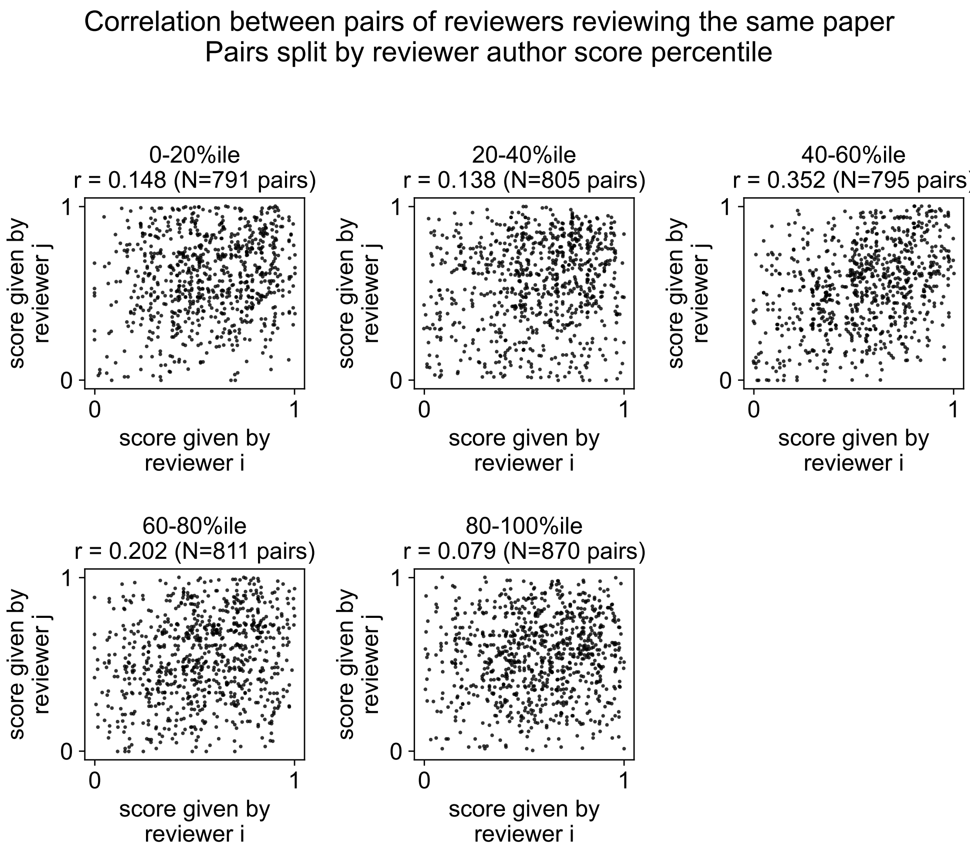
SUPPLEMENTARY MATERIALS



Supplementary Figure S1. Confidence of ratings is inversely correlated to the assigned scores. In the ICLR2023 open peer review data, confidence of assigned ratings was inversely correlated to the assigned scores ($r = -0.115, p < 0.001, N = 14359$). For every review of every submission in ICLR2023, this plot shows the overall paper score given by the reviewer on the y-axis and the corresponding confidence assigned by the reviewer on the x-axis.



Supplementary Figure S2. Reviewers with intermediate author scores are the best at predicting the community average paper quality score. (A) Distribution of mean standard deviations from community average score across the different reviewer groups, grouped by the authorship score of the reviewer. (B) Reviewers with the author scores in the range between 40th and 60th percentiles have significantly lower mean squared deviation from the community average score than the reviewers from the ends of the spectrum of author scores (0-10th and 90-100th percentiles; $*p < 0.05$).



Supplementary Figure S3. Reviewers with intermediate author scores correlate the most with other reviewers within their group. Correlation between pairs of reviewers reviewing the same paper, split by the authorship score of the reviewers. The correlation is the highest for the reviewers with intermediate authorship scores (in the range between 40th and 60th percentiles; $N > 90$ pairs, all panels).