Evaluation of Crowdsourced Peer Review using Synthetic Data and Simulations

Michael Soprano[1,*], Eddy Maddalena[1], Francesca Da Ros[1], Maria Elena Zuliani[1] and Stefano Mizzaro[1]

[1]Department of Mathematics, Computer Science and Physics, University of Udine, Udine, Friuli-Venezia Giulia, Italy

Abstract

The scholarly publishing process relies on peer review to uphold the quality of scientific knowledge. However, challenges such as increasing submission volumes and potential malicious behavior undermine its effectiveness. In this study, we evaluate Readersourcing, an alternative peer review approach that leverages community-driven judgments. Using simulations with synthetic data based on a probabilistic model and a publicly available implementation, we assess six quantities and examine the impact of each component on the outcomes. Our findings show that the co-determination algorithm captures distinct aspects of manuscript judgments compared to simpler aggregation strategies. Key simulation parameters consistently influence the computed quantities across different settings. We also publicly release the data, code, and simulation runs.

Keywords

Scholarly Publishing, Peer Review, Evaluation, Readersourcing

1. Introduction

The primary method for disseminating scientific knowledge is the *scholarly publishing* process, which relies on *peer review*. In this process, a scientific article authored by individuals is assessed and evaluated by peers with equivalent expertise. Although peer review is a well-established method for ensuring the quality of scientific publications, it is not without drawbacks [1]. These include challenges in managing the increasing volume of submissions and the potential for malicious behavior by some stakeholders [2]. Several approaches to addressing these limitations have been discussed in the literature, including outsourcing the review process to the broader scientific community itself [3, 4, 5].

While recent advancements in Artificial Intelligence (AI) technologies make automating peer review in the scholarly publishing process an appealing prospect, several issues and concerns warrant further investigation. For instance, existing tools struggle to understand and interpret manuscripts within the broader context of scientific literature [6], and AI-based approaches to peer review are prone to systematic biases [7]. Some researchers suggest that a hybrid approach could involve cooperation between humans and AI in the peer review process [8]. However, the nuanced judgment and contextual understanding that human reviewers provide remain crucial for ensuring the integrity and reliability of scientific evaluation.

In light of this, we specifically focus on the Readersourcing model (RSM), originally introduced by Mizzaro [4], which provides a framework for enhancing the peer review process in scholarly publishing through community-driven numerical judgments. RSM quantifies both the overall quality of an article and the reputation of a scholar, considered as a reader and as an author, using a co-determination algorithm. The primary challenge lies in aggregating these ratings into quality and reputation indices

IRCDL 2025: 21st Conference on Information and Research Science Connecting to Digital and Library Science, February 20-21 2025, Udine, Italy

*Corresponding Author.

michael.soprano@uniud.it (M. Soprano); eddy.maddalena@uniud.it (E. Maddalena); francesca.daros@uniud.it (F. Da Ros); zuliani.mariaelena@spes.uniud.it (M. E. Zuliani); stefano.mizzaro@uniud.it (S. Mizzaro)

https://michaelsoprano.com/ (M. Soprano); https://users.dimi.uniud.it/~eddy.maddalena/ (E. Maddalena); https://users.dimi.uniud.it/~stefano.mizzaro/ (S. Mizzaro)

0000-0002-7337-7592 (M. Soprano); 0000-0002-5423-8669 (E. Maddalena); 0000-0001-7026-4165 (F. Da Ros); 0009-0006-6374-261X (M. E. Zuliani); 0000-0002-2852-168X (S. Mizzaro)

and ultimately deriving a single index for each measure. The model has been implemented in a system accessible to the research community [9, 10]. Although RSM has been evaluated using social network metrics [11], the co-determination algorithm and the influence of its components on the computed quantities remain insufficiently investigated.

1.1. Aims

In this work, we evaluate RSM through simulations on synthetic data generated using a probabilistic model. First, we show that the model effectively captures distinct and meaningful aspects of judgments, providing strong evidence for its adoption. Second, we validate its structural properties, confirming that the model's design aligns with the foundational principles outlined by Mizzaro [4]. Through a detailed analysis of these properties, we highlight the model's potential to enable the outsourcing of the peer review process to the broader community of readers.

Specifically, we investigate the following Research Questions (RQs):

RQ1 How does the probabilistic approach used in our simulations influence the computed quantities of the model? What improvements could enhance its outcomes?

RQ2 How effectively does RSM capture meaningful and distinct aspects of judgments made by readers? How do the insights generated by RSM compare with those from simpler aggregation strategies?

RQ3 What is the impact of each component of the RSM model on the computed quantities? How do the different components influence the overall results? How does their interaction contribute to the model's outcomes?

The data, code, and all supplementary materials related to our study are publicly available to the research community at: https://osf.io/kwv47/.

1.2. Contributions

Our contributions are as follows: (i) We show that the co-determination algorithm in RSM provides meaningful differentiation compared to traditional, widely adopted aggregation strategies. (ii) We identify key simulation parameters that significantly affect the model's outputs, highlighting those that play a prominent role in computing critical quantities. (iii) We confirm that these key parameters consistently shape the co-determination process across various simulation settings. (iv) We publicly release our simulation dataset to support further research and analysis.

1.3. Outline

The remainder of this paper is structured as follows: Section 2 provides an overview of the related literature, Section 3 describes the methodology, Section 4 presents the results, and Section 5 discusses the impact and outlines the limitations of our approach. Finally, Section 6 presents the conclusions and indicates directions for future research.

2. Related Work

Scholarly publishing, which relies on peer review, is the primary method for disseminating scientific knowledge. In this process, scientific articles authored by researchers are evaluated by peers with comparable expertise and, if deemed of sufficient quality, are made available to the broader community [12]. Although peer review is the cornerstone of evaluating the quality of scientific publications, it has shortcomings [13]. First, the system is under strain due to the large volume of submissions [14, 15] and the time required to process them efficiently. In this regard, reviewers have been described as a *scarce resource* [3]. Second, the peer review system is prone to bias [16, 17] and inconsistencies [18].

To address these limitations, several solutions have been proposed [19], focusing on the transparency, efficiency, quality, and equity of the process [20]. One example is the concept of open peer review, in which the publication is accompanied by anonymous reviews [21, 22].

More recently, efforts have been made to incorporate automated tools and AI into the peer review process [23]. For instance, Checco et al. [24] developed an AI tool trained on 3,300 papers from three conferences, along with their corresponding review evaluations. The tool was designed to predict the review score of a new, unseen manuscript based solely on its textual content. Similarly, Boukhris and Zaâbi [25] proposed a GAN-BERT-based method to analyze the sentiment of reviewers' comments and automatically generate an objective final decision regarding the acceptance or rejection of a manuscript.

As Large Language Models (LLMs) are tested in the field of generating scientific hypotheses [26], they have also been employed in the peer review process, prompting many journals to establish specific policies to regulate their use [27, 28]. Latona et al. [29] conducted experiments on the scores and reviews of the 2024 International Conference on Learning Representations (ICLR 2024), finding that over 15% of the reviews were written with the assistance of AI (verified through experiments with GPTZero). Interestingly, these AI-generated reviews tended to assign higher scores compared to non-AI-generated reviews. The potential of LLMs for assessing the quality of scientific papers has been explored by Liang et al. [30], who developed a pipeline employing Generative Pretrained Transformer 4 (GPT-4) to generate comments on research articles. The results indicated that over half of the users rated GPT-4-generated feedback as helpful or very helpful, with many finding it more beneficial than feedback from at least one of the human reviewers. Similarly, Santu et al. [31] investigated the generation of meta-reviews by leveraging three LLMs (LLaMA2, GPT-3.5, and PaLM2) using data from ICLR from 2020 to 2023. Their qualitative analysis showed that GPT-3.5 and PaLM2 performed comparably overall, with both being rated higher by humans than LLaMA2 for manuscript-level judgments. Notably, PaLM2 demonstrated superior recall scores, while GPT-3.5 achieved better precision scores, highlighting the varying strengths of these LLMs in generating meta-reviews.

All in all, while AI and LLMs can undoubtedly enhance the efficiency of the peer review process, they also raise ethical concerns, particularly regarding the transparency of the process, disclosure agreements, and the potential replication and amplification of biases inherent in the data or systems [8]. These challenges highlight the continued necessity of human intervention in the process [32].

3. Methodology

3.1. The Readersourcing Model

Readersourcing (RSM) is a crowdsourcing approach to peer review [4, 3], which can serve as either a pre-publication alternative or a post-publication complement to the traditional peer review process. Figure 1 illustrates the general framework, and Table 1 summarizes the notation. We provide only a brief overview of the model. For additional details, see Mizzaro [4].

RSM involves three key entities: a set $\mathcal{M}$ of *manuscripts* (also referred to as *publications*, *articles*, or *papers*), a set $\mathcal{A}$ of *authors*, and a set $\mathcal{R}$ of *readers*. When a reader $l \in \mathcal{R}$ reads a manuscript $k \in \mathcal{M}$, they assign a numerical value $j_{\mathrm{r}l,\mathrm{m}k} \in [0, 100]$, referred to as *judgment* (or *rating*).

Each entity in the model is associated with a *score*:

- The *manuscript score* $s_{\mathrm{m}k}$ of a manuscript $k \in \mathcal{M}$ is calculated as an aggregation of its judgments, serving as an indicator of its quality.
- The *author score* $s_{\mathrm{a}i}$ of an author $i \in \mathcal{A}$ is derived from the aggregation of the scores of the manuscripts they have published, serving as an indicator of their reputation and skills.
- The *reader score* $s_{\mathrm{r}l}$ of a reader $l \in \mathcal{R}$ is determined by comparing their judgments on manuscripts with those of other readers, serving as an indicator of their reputation and skills.

Scores are dynamic and evolve over time based on user behavior and interactions. Each score is paired with a *steadiness value*, denoted with σ, that reflects its stability: $\sigma_{\mathrm{m}k}$ for manuscripts, $\sigma_{\mathrm{a}i}$ for

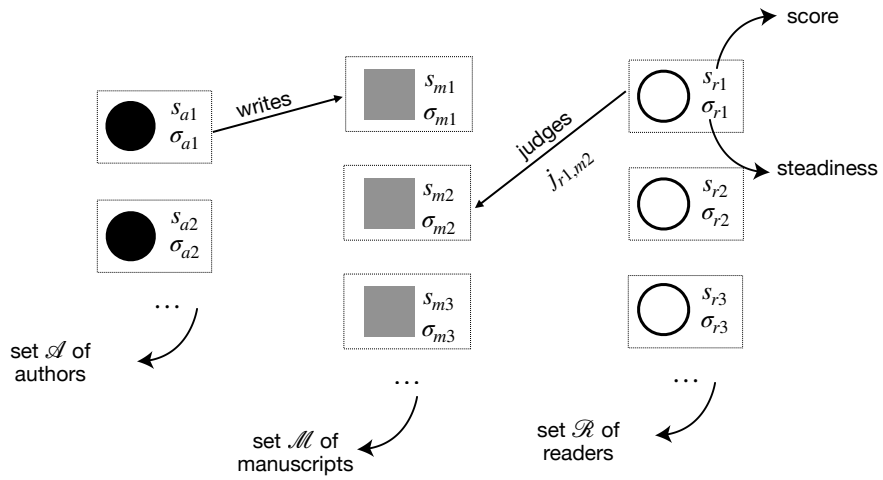

Figure 1: General framework of the Readersourcing model.

Table 1
Notation used to represent the entities and quantities involved in Readersourcing.

Symbol	Definition
$\mathcal{A}$	Set of authors.
$\mathcal{M}$	Set of manuscripts.
$\mathcal{R}$	Set of readers.
$j_{rl,\mathrm{m}k}$	Judgment of reader $l \in \mathcal{R}$ on manuscript $k \in \mathcal{M}$.
$s_{\mathrm{a}i}$	Score of author $i \in \mathcal{A}$.
$s_{\mathrm{m}k}$	Score of manuscript $k \in \mathcal{M}$.
s_{rl}	Score of reader $l \in \mathcal{R}$.
$\sigma_{\mathrm{a}i}$	Steadiness of author $i \in \mathcal{A}$.
$\sigma_{\mathrm{m}k}$	Steadiness of manuscript $k \in \mathcal{M}$.
σ_{rl}	Steadiness of reader $l \in \mathcal{R}$.

authors, and σ_{rl} for readers. For example, an older manuscript with many evaluations tends to exhibit a high steadiness value, whereas a newly registered reader will have a low steadiness value. Steadiness influences how scores are updated: lower steadiness leads to faster score adjustments in response to new inputs. As the score stabilizes, its steadiness increases.

One can consider RSM as a tripartite graph whose nodes correspond to three sets: authors, manuscripts, and readers. Authors are connected to the manuscripts they publish, and readers are connected to the manuscripts they read. More formally, an edge exists between an author $i \in \mathcal{A}$ and a manuscript $k \in \mathcal{M}$ if i publishes k (such edges are unweighted). Conversely, there is an edge between a reader $l \in \mathcal{R}$ and a manuscript $k \in \mathcal{M}$ if l reads k; the weight of this edge corresponds to the judgment $j_{rl,\mathrm{m}k}$ that l expresses on k.

Note that if a user acts as both an author and a reader, they maintain two distinct scores and steadiness values.

3.2. Simulations Flow and Assumptions

The simulation flow is illustrated in Figure 2, while the parameters used in the simulation are summarized in Table 2. Each simulation starts with a fixed number of authors and readers. In this study, these two quantities are set to be equal, i.e., $|\mathcal{A}| = |\mathcal{R}|$ (see Step 1 in Figure 2).

It is well established that scholarly publications follow Power Law distributions [33, 34]. Informally, in a Power Law distribution a small number of events occur with very high frequency, while most

Table 2
Notation used in the simulation flow.

Symbol	Definition
$P_{\mathrm{a}}(x)$	Power Law distribution modeling the exact number of manuscripts per author.
$P_{\mathrm{m}}(y)$	Power Law distribution modeling the expected number of times a manuscript is read.
$P_{\mathrm{r}}(x)$	Power Law distribution modeling the expected number of manuscripts read per reader.
$P_{\mathrm{G}}(x)$	Power Law distribution modeling the reference score of manuscripts.
$P_{\mathrm{r,m}}(x, y)$	Joint probability of reader $l \in \mathcal{R}$ reading manuscript $k \in \mathcal{M}$.
α_{a}	Exponent for the Power Law distribution modeling the number of manuscripts per author.
α_{m}	Exponent for the Power Law distribution modeling the number of times a manuscript is read.
α_{r}	Exponent for the Power Law distribution modeling the number of manuscripts read per reader.
α_{G}	Exponent for the Power Law distribution modeling the reference score of manuscripts.
$c_{\max}$	Maximum number of manuscripts that each author can publish.
$g_{\mathrm{m}k}$	Reference score of manuscript $k \in \mathcal{M}$.
$\mu_{\mathrm{m}k}$	Mean of the Gaussian distribution related to the reference score of manuscript $k \in \mathcal{M}$.
$st_{\mathrm{m}k}$	Standard deviation of the Gaussian distribution related to the reference score of manuscript $k \in \mathcal{M}$.

events occur with low frequency. Its general form is given by $P(x) \sim x^{-\alpha}$, where α is the scaling exponent. Using this distribution, we model the following quantities:

- **Number of manuscripts published by each author**: The Power Law distribution representing the number of manuscripts published by an author is denoted by $P_{\mathrm{a}}(x)$ and is parameterized by α_{a}. To avoid extreme values, we impose an upper limit $c_{\max}$ on the maximum number of manuscripts an author can publish (see Step 2 in Figure 2).
- **Number of manuscripts read by each reader**: The Power Law distribution for the number of manuscripts read by each reader is denoted by $P_{\mathrm{r}}(x)$ and is parameterized by α_{r} (see Step 3 in Figure 2).
- **Number of reads per manuscript**: The Power Law distribution for the number of times each manuscript is read is denoted by $P_{\mathrm{m}}(y)$ and is parameterized by α_{m} (see Step 4 in Figure 2).

The probability that a specific reader $l \in \mathcal{R}$ reads a specific manuscript $k \in \mathcal{M}$ is denoted as $P_{\mathrm{r,m}}(x, y)$. Assuming the two events are independent, the joint probability that reader l reads manuscript k is expressed as $P_{\mathrm{r,m}}(x, y) = P_{\mathrm{r}}(x) \cdot P_{\mathrm{m}}(y)$. (see Step 5 in Figure 2). To model judgments of readers on manuscripts, we proceed as follows. For each manuscript $k \in \mathcal{M}$, we draw a reference score (i.e., the "ideal" score of the manuscript), denoted as $g_{\mathrm{m}k}$, from a Power Law distribution $P_{\mathrm{G}}(x)$ with parameter α_{G} (see Step 6 in Figure 2).

The judgment $j_{\mathrm{r}l,\mathrm{m}k}$ assigned by reader $l \in \mathcal{R}$ to manuscript $k \in \mathcal{M}$ is drawn from a Gaussian distribution with mean $\mu_{\mathrm{m}k}$ set to the reference score $g_{\mathrm{m}k}$ and standard deviation $st_{\mathrm{m}k}$. Judgment values are bounded between 0 and 100 (see Step 7 in Figure 2). In the original work [4], these values ranged from 0 to 1.

The simulation flow is based on the following assumptions, derived from the foundational work of Mizzaro [3]:

(i) Each manuscript $k \in \mathcal{M}$ has exactly one author.

(ii) The judgments $j_{\mathrm{r}l,\mathrm{m}k}$ are independent.

(iii) The number of authors equals the number of readers ($|\mathcal{A}| = |\mathcal{R}|$).

(iv) The simulation does not enforce connections between all entities: some manuscripts may remain unread, and some readers may not read any manuscripts.

As discussed in Section 6, these assumptions will be addressed and refined in future work.

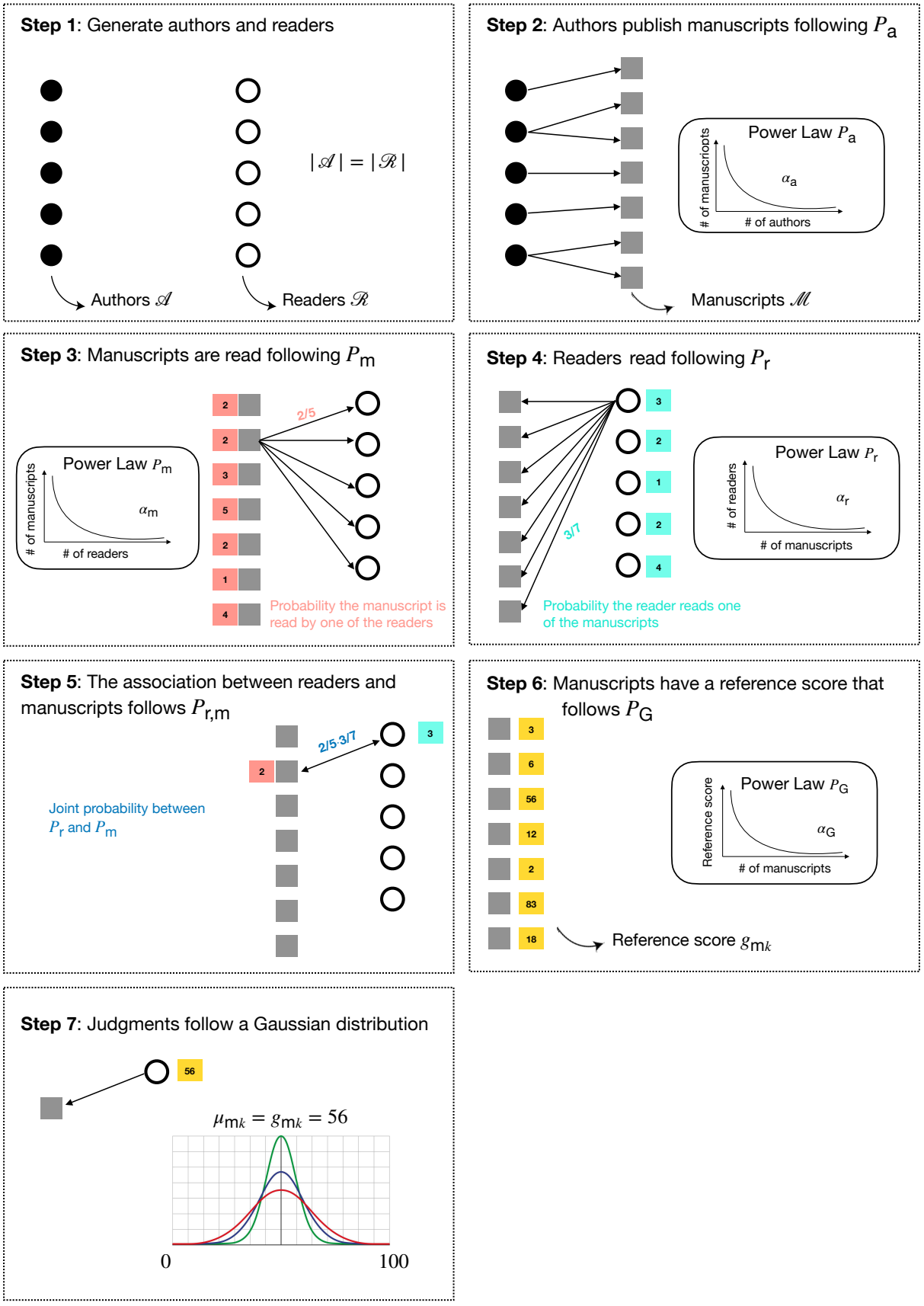

Figure 2: Overview of the simulation flow, detailing the steps involved in modeling authors, manuscripts, readers, and judgments based on Power Law and Gaussian distributions.

3.3. Experimental Setup

To conduct our simulations, we generate a set of configurations by choosing discrete values for six parameters (see Table 2): the number of authors/readers $|\mathcal{A}| = |\mathcal{R}|$, the scaling exponents for each Power Law ($\alpha_{\mathrm{a}}, \alpha_{\mathrm{m}}, \alpha_{\mathrm{r}}, \alpha_{\mathrm{G}}$), and the standard deviation of the Gaussian distribution for the reference score st_{m}. We fix the upper limit on the number of manuscripts, $c_{\max}$, to a constant value in all simulations, thereby excluding it as a parameter.

The simulations use a population of 250 authors/readers, which can be scaled up in future work. We select exponents of 1.1, 1.2, and 1.3 to explore different power law steepness levels, where higher values lead to more concentrated distributions. Standard deviations of 2.5, 5.0, 7.5, and 10.0 control data variability: smaller values keep data closer to the mean, while larger values foster the presence of outliers. These choices allow us to examine the simulation thoroughly under varied conditions. The total number of configurations is the product of the possible values for each parameter: 1 population size $\times$ 81 α values $\times$ 4 standard deviation values, for a total of 324 configurations. We repeat each configuration 10 times to account for stochasticity, leading to 3,024 executions in total.

We conducted all experiments on a machine equipped with an Intel Core i7-10700 CPU, 64 GB of DDR4 RAM, NVIDIA RTX A6000 and RTX 3090 GPUs, a 1 TB NVMe SSD, and a 4 TB HDD. We ran the simulations using Python 3.9.21 in a Conda-based environment. To ensure compatibility with the `Parquet` serialization format, we used `NumPy` 2.0.2 and `pyarrow` 18.1.0.

We use the Python-based implementation of RSM, which is publicly available on GitHub: https://github.com/EddyMaddalena/Readersourcing_OO.

3.4. Evaluation

3.4.1. Approach

Our analysis examines the relationship between simulation parameters and the quantities (score and steadiness) computed by the RSM co-determination algorithm for each entity (authors, manuscripts, and readers), based on individual judgments. We compare these computed values across all simulation runs and repetitions, with each comparison focusing on a specific entity.

To ensure reliable results, we exclude inactive entities from the simulation flow, such as unread manuscripts, readers who do not provide judgments, and authors who publish only unread or excluded manuscripts (see Section 3.2). We also omit entities with a steadiness of 0, as it indicates a lack of active participation in the model. Given their absence from the computation, we refer to them as *inert*. In contrast, we classify as *active* those entities that participate in the process, i.e., readers who provide judgments, manuscripts that receive judgments, and authors who publish manuscripts that are actively judged. From now on, we will generally refer to quantities using a single specifier. For instance, we will use s_{m} instead of $s_{\mathrm{m}k}$ to refer to the manuscript score.

3.4.2. Aggregation Functions

We evaluate the performance and distinctiveness of the co-determination algorithm in RSM by comparing its outputs with aggregations derived from simpler, widely adopted strategies for combining individual judgments.

We examine the impact of three commonly used aggregation functions: the arithmetic mean, geometric mean, and median. These functions are relevant in data aggregation tasks for different reasons: the arithmetic mean summarizes central tendencies, the geometric mean is well suited for products and rates, and the median is less affected by outliers. By comparing these results with those produced by the co-determination algorithm, we aim to assess whether RSM provides meaningful differentiation and additional insights beyond traditional aggregation methods.

Table 3
Summary statistics for active entities (authors, manuscripts, and readers) in the simulations. The target number of authors and readers is 250 (see Section 3.2).

Statistic	Active Authors	Active Manuscripts	Active Readers
mean	214	1844	162
min	175	899	115
25%	208	1557	150
50%	216	1811	163
75%	223	2110	176
max	241	3340	206

3.4.3. Random Forest Regressor (RFR)

The Random Forest Regressor (RFR) is an ensemble learning method that constructs multiple decision trees during training and outputs the average prediction from all trees. It is highly valued for its ability to capture complex, non-linear relationships between input features and the dependent variable, delivering robust performance with minimal hyperparameter tuning.

In our study, we use RFR to model the relationship between the simulation parameters and the quantities produced by the co-determination algorithm of RSM. We treat α_a, α_m, α_r, α_G, and st_m (see Table 2) as features, and s_a, s_m, s_r, σ_a, σ_m, and σ_r (see Table 1) as outputs. We apply RFR separately to the data for each of the three entities. For example, one RFR run compares α_a with s_a and σ_a.

The feature importance rankings generated by the Random Forest help pinpoint which simulation parameters most significantly affect the output quantities. We leverage these insights to evaluate the model's robustness and to confirm the importance of these parameters in subsequent analyses, such as Analysis of Variance (ANOVA). We use the `RandomForestRegressor` implementation from `scikit-learn` [35].

3.4.4. Analysis of Variance (ANOVA)

ANOVA [36] is a statistical method used to test for significant differences between the means of multiple groups. Provided its underlying assumptions are satisfied [37, 38, 39, 40, 41], it evaluates the impact of each feature on the outputs at the population level.

We use one-way ANOVA to assess the influence of simulation parameters on the quantities computed by the co-determination algorithm of RSM. We adopt the same definitions of features and outputs as for RFR (see Section 3.4.3). By analyzing how these quantities vary across different parameter settings, we aim to identify which parameters significantly affect the results, thus validating the insights from our feature importance analysis. We use the `f_oneway` implementation from the `scipy` package [42].

4. Results

4.1. RQ1: Effect of the Probabilistic Approach

To evaluate the effect of our probabilistic approach in simulations (RQ1), we focus on inert entities. By design, our approach is designed toallow their presence (see Section 3.2 and Section 3.4.1). To guide future experimental designs, we investigate inert entities using a twofold approach.

First, we examine this phenomenon quantitatively by counting the number of active and inert entities in our simulations, ensuring that inert entities do not overwhelmingly dominate. Table 3 shows the number of active authors, manuscripts, and readers. For each entity, we report the mean, minimum, maximum, and the first, second, and third quartiles. Notably, authors tend to remain active longer in the network due to higher connectivity; of the 250 simulated authors, 214 (85%) are active.

Next, we analyze correlations to understand how these values relate to the input parameters, aiming to identify potential interventions for tuning simulation parameters to reduce the occurrence of inert

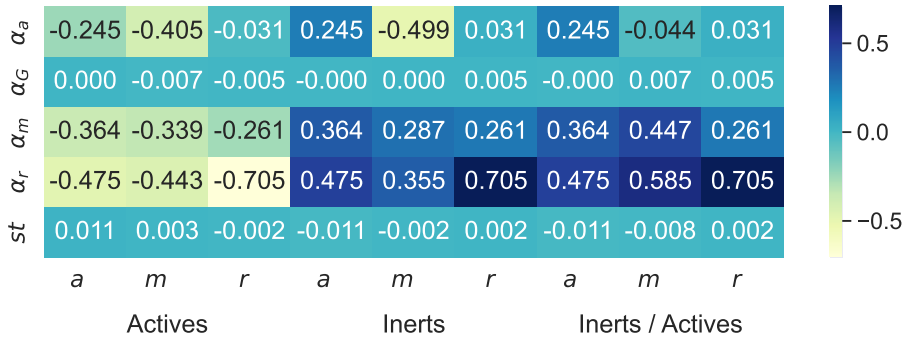

Figure 3: Heatmap showing Kendall's τ correlation values w.r.t. the number of active entities (first to third columns), the number of inert entities (fourth to sixth columns), and the ratio between inert and active entities (seventh to ninth columns). Each cell represents a correlation value.

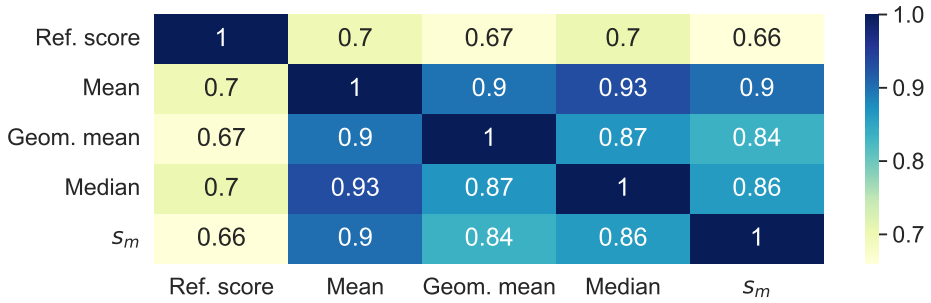

Figure 4: Heatmap showing Kendall's τ correlation values w.r.t. the reference score (first column), judgments assigned by readers in each simulation and aggregated using the arithmetic mean, geometric mean, and median (second through fourth columns), and the score assigned to the manuscript (last column). Each cell represents a correlation value.

entities. Figure 3 shows the correlation values between the simulation parameters and the number of active entities (first to third columns), the number of inert entities (fourth to sixth columns), and the ratio between active and inert entities (seventh to ninth columns). Positive correlations are shown in blue, while negative correlations appear in light green.

As the figure indicates, some simulation parameters increase the number of active entities, whereas others increase the number of inert ones. Specifically, the α_{G} and st_{m} parameters do not affect inert entities. This is expected because they govern aspects related to individual judgments and do not influence edge formation in the tripartite graph. In contrast, the three parameters α_{a}, α_{m}, and α_{r}, which govern the Power Law distributions for authors, manuscripts, and readers, do affect inert entities. The correlations show that these parameters decrease the number of active entities, increase the number of inert ones, and ultimately increment the ratio between inert and active entities.

4.2. RQ2: Comparison with Simpler Aggregation Strategies

To evaluate how effectively RSM captures distinct aspects of manuscript judgments (RQ2), we compare the model's outputs with simpler strategies for aggregating individual judgments. Figure 4 presents the resulting Kendall's τ correlation coefficients. The first column shows correlations with the reference score, while the second to fourth columns show correlations with judgments aggregated using the arithmetic mean, geometric mean, and median, respectively. The final column shows the manuscript score. Positive correlations are shown in blue, and negative correlations in light green.

The manuscript score s_{m} correlates moderately with the reference score but strongly with simpler aggregation strategies, namely the arithmetic mean, geometric mean, and median. This stronger correlation is likely due to the effect of the normal distribution, as the median aligns well with such a distribution. Furthermore, because the dataset is generated from skilled readers with highly consistent

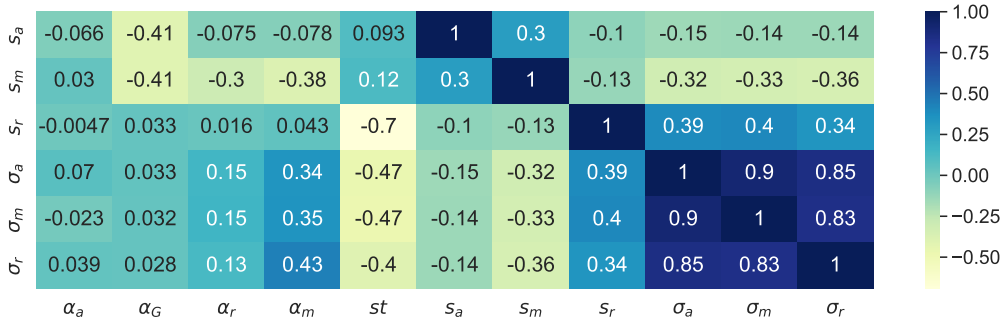

Figure 5: Heatmap showing Kendall's τ correlation values w.r.t. simulation parameters (first to fifth columns), scores (sixth to eighth columns), and steadiness values (ninth to eleventh columns). Each cell represents a correlation value.

and precise judgments, it minimizes the influence of rewarding better-performing readers. Consequently, the model emphasizes overall trends, while its weaker correlation with the reference score suggests that it captures a distinct dimension of judgment. In summary, the model places more weight on aggregated judgment patterns than on the reference score. We discuss this further in Section 6.

4.3. RQ3: Impact of Model Components on Quantities

To evaluate the impact of individual features on the outputs (RQ3), we begin by examining their correlations. Figure 5 shows Kendall's τ correlation coefficients with respect to both simulation parameters and computed quantities. In this figure, as well as in subsequent ones, positive correlations are shown in blue, while negative correlations in light green.

In the lower-right section of the heatmap, we observe a strong correlation among the steadiness values computed for each entity. High steadiness in one entity is associated with high steadiness in others, indicating that steadiness is a system-wide property observed uniformly across all entities. In the top-center blue section, the positive correlation among author scores indicates that authors with higher manuscript scores tend to receive higher individual scores. An unexpected result is the negative correlation between author and manuscript scores, on the one hand, and reader scores on the other: when authors and manuscripts have higher scores, readers tend to have lower scores, and vice versa. We discuss this further in Section 5.

Focusing on the feature importance analysis conducted using RFR, Figure 5 shows the computed feature importance scores, revealing each simulation parameter's contribution. The first three columns pertain to scores, while the fourth through sixth columns represent steadiness values.

As expected, the α_{G} parameter has the greatest influence on steadiness, since variations in the manuscript's reference score cause fluctuations in steadiness. In contrast, the α_{r} parameter affects the manuscript score by determining how many readers provide judgments. Meanwhile, the α_{a} parameter, which governs the number of manuscripts published, influences the reader score. This suggests a relationship between manuscript amount and the consistency of reader judgments.

Next, we present the results of our one-way ANOVA analysis. Figure 7 shows the F-statistic values, which quantify the strength of the relationships between simulation parameters and computed quantities. All values are statistically significant ($p < 0.0001$), except for the s_{a} column.

As expected, the manuscript score s_{m} is strongly influenced by all features, given that it arises from interactions among the parameters. The st_{m} parameter, which influences all quantities, especially the reader score s_{r}, affects judgment variability. When the standard deviation is low, errors in judgments are smaller, effectively making the reader "skilled". Examining higher variability could lead to additional insights into the reader score, as noted in Section 6.

Both the RFR and ANOVA analyses consistently indicate that simulation parameters significantly influence the manuscript score s_{m}, with α_{G} and α_{r} exerting particularly strong effects. While RFR highlights the relative importance of each feature, ANOVA confirms the statistical significance of these

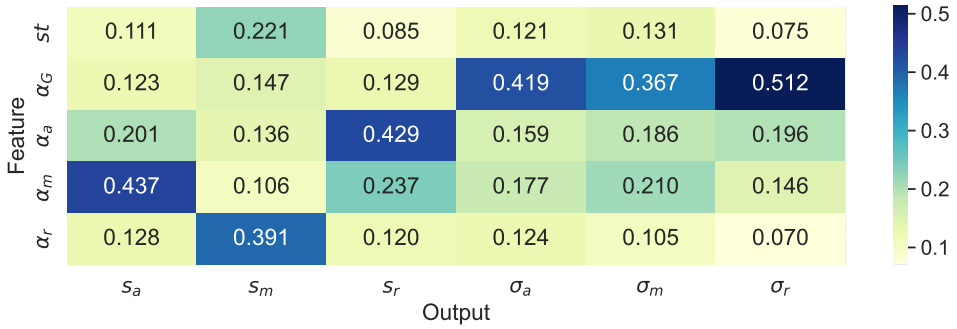

Figure 6: Heatmap showing feature importance w.r.t. scores (first to third columns) and steadiness values (fourth to sixth columns). Each cell represents a feature importance value.

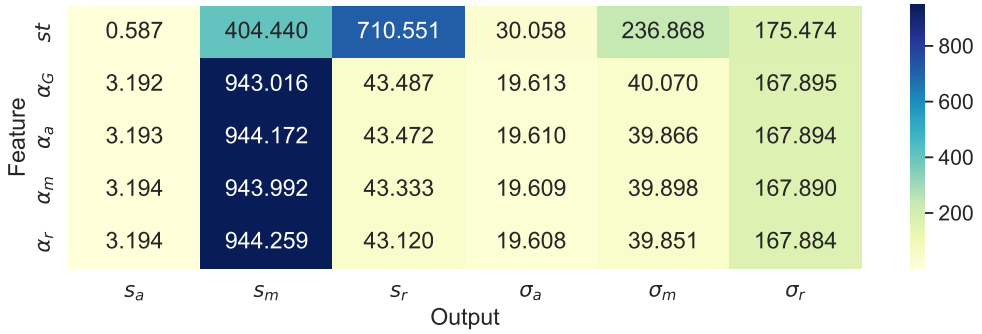

Figure 7: Heatmap showing the results of the one-way F-analysis w.r.t scores (first to third columns) and steadiness values (fourth to sixth columns). Each cell represents an F-statistic value. All values are statistically significant ($p < 0.0001$), except for the s_a column.

impacts. The high importance of α_G revealed by RFR is confirmed by ANOVA, underscoring how variations in the reference score drive changes in steadiness. Similarly, both analyses highlight the influence of the standard deviation parameter on the reader score, highlighting the role of judgment variability in consistency.

5. Discussion and Limitations

Our findings indicate that the probabilistic approach effectively accounts for inert entities within simulations. Correlation analyses reveal that certain simulation parameters affect the number of inert entities by reducing the number of active entities and increasing the number of inert ones. These insights guide parameter tuning to minimize the occurrence of inert entities and refine experimental designs in future studies (RQ1).

Additionally, our results show that RSM captures distinct aspects of manuscript judgments by focusing on aggregated judgment patterns, such as the mean and median, rather than the reference score. Its strong correlation with simpler aggregation methods reflects overall judgment trends, while its weaker correlation with the reference score suggests it captures a different dimension of judgments (RQ2).

Further analysis of RSM model components reveals how each feature contributes to the final results. The strong correlation among steadiness values highlights its role as a consistent, system-wide property. Feature importance analysis suggests that a higher manuscript count leads to more stable judgments, while ANOVA highlights the impact of standard deviation on score steadiness. Greater variability in judgments produces less stable outcomes, reflecting the complex interactions of model components (RQ3).

A limitation of our approach is that reference scores follow a Power Law distribution, often resulting in many manuscripts with near-zero scores. This issue is worsened by the bounded scale, where

clipping values outside the predefined range can distort judgments [43]. Low-quality manuscripts with near-zero scores are considered "easy" to judge, which leads to small judgment errors. This dynamic creates a countervariance effect where skilled readers provide high-quality judgments, while authors of low-quality manuscripts receive lower scores. An unbounded scale, such as Magnitude Estimation [44], could mitigate this issue.

Another aspect requiring further attention is the treatment of inert nodes. While excluding them from the analysis is reasonable (see Section 3.4.1), real-world publishing systems might indeed include unread manuscripts and inactive readers. Exploring alternative approaches to inert nodes could provide further insights, such as applying graph growth models where inert nodes are "drawn" to highly connected nodes [45].

Our simulation flow also employs four bounded power laws, which introduce four additional parameters and increase complexity. A potential solution is proposed by Antipov et al. [46], who reduce the number of parameters by randomly selecting values from a suitably scaled power law distribution at each iteration.

6. Conclusions and Future Research Directions

In conclusion, our study shows the effectiveness of the co-determination algorithm in RSM, highlighting key simulation parameters that significantly influence the model's outputs and confirming their consistent impact across various settings.

Future research should aim to enhance RSM by comparing it with other models and exploring more complex simulations with diverse scenarios and variables. Mizzaro [4] suggests improvements like considering multiple authors per manuscript for greater realism. Other possibilities include modeling manuscripts submitted to journals with varying acceptance thresholds and assigning multiple scores for authors and readers. Additionally, we propose allowing readers to revise their judgments over time, as the current model assumes fixed judgments within a set period.

A possible comparison with existing models would involve evaluating RSM reader scores against those from the co-determination algorithm in TrueReview [5]. Additionally, a network analysis using the HITS algorithm [11] could be applied to reassess the updated model.

Future simulations could benefit from real-world data, such as the 1.5 million preprints available on arXiv [47]. As proposed by Mizzaro [4], one approach is to model various reader types with different judgment patterns. For example, we could simulate readers who consistently overestimate, underestimate, or align closely with the reference score by adjusting the standard deviation of the Gaussian distribution that generates reader scores. Skilled readers would have low standard deviations, while less skilled readers would have higher ones, enabling us to evaluate how closely their scores reflect their skill levels.

Finally, incorporating LLMs into simulations is becoming increasingly relevant due to their ability to replicate human-like behavior. For instance, Park et al. [48] describe an agent-based framework that emulates realistic individual behaviors and attitudes. We believe such a framework could enrich simulations by modeling reader judgments influenced by demographic, ideological, and personality traits. However, caution is advised when using LLMs as proxies for human decision-making [32].

Acknowledgments

This research is partially supported by the PRIN 2022 Project – "MoT—The Measure of Truth: An Evaluation-Centered Machine-Human Hybrid Framework for Assessing Information Truthfulness" (Code No. 20227F2ZN3, CUP No. G53D23002800006), funded by the European Union – Next Generation EU – PNRR M4 C2 I1.1, and by the Strategic Plan of the University of Udine–Interdepartment Project on Artificial Intelligence (2020-25).

References

[1] W. Y. Arms, What are the alternatives to peer review? Quality control in scholarly publishing on the web, JEP 8 (2002). doi:10.3998/3336451.0008.103.

[2] S. Jecmen, M. Yoon, V. Conitzer, N. B. Shah, F. Fang, A Dataset on Malicious Paper Bidding in Peer Review, in: Proceedings of the ACM Web Conference 2023, WWW '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 3816–3826. doi:10.1145/3543507.3583424.

[3] S. Mizzaro, Readersourcing—A manifesto, Journal of the American Society for Information Science and Technology 63 (2012) 1666–1672. doi:10.1002/asi.22668.

[4] S. Mizzaro, Quality control in scholarly publishing: A new proposal, Journal of the American Society for Information Science and Technology 54 (2003) 989–1005. doi:10.1002/asi.10296.

[5] L. de Alfaro, M. Faella, TrueReview: A Proposal for Post-Publication Peer Review, Technical Report UCSC-SOE-16-13, University of California, Santa Cruz, 2016. URL: https://tr.soe.ucsc.edu/research/technical-reports/UCSC-SOE-16-13.

[6] R. Schulz, A. Barnett, R. Bernard, N. J. L. Brown, J. A. Byrne, P. Eckmann, M. A. Gazda, H. Kilicoglu, E. M. Prager, M. Salholz-Hillel, G. ter Riet, T. Vines, C. J. Vorland, H. Zhuang, A. Bandrowski, T. L. Weissgerber, Is the future of peer review automated?, BMC Research Notes 15 (2022) 203. doi:10.1186/s13104-022-06080-6.

[7] S. Price, P. A. Flach, Computational support for academic peer review: a perspective from artificial intelligence, Communications of the ACM 60 (2017) 70–79. doi:10.1145/2979672.

[8] F. A. Mohammed Salah, H. A. Halbusi, Debate: Peer reviews at the crossroads—'To AI or not to AI?', Public Money & Management 43 (2023) 781–782. doi:10.1080/09540962.2023.2264032.

[9] M. Soprano, S. Mizzaro, Crowdsourcing Peer Review: As We May Do, in: Manghi, Paolo and Candela, Leonardo and Silvello, Gianmaria (Ed.), Digital Libraries: Supporting Open Science, volume 988 of *Communications in Computer and Information Science*, Springer, 2019, pp. 259–273. doi:10.1007/978-3-030-11226-4_21.

[10] M. Soprano, S. Mizzaro, Crowdsourcing Peer Review in the Digital Humanities?, in: Book of Abstracts, 8th AIUCD Conference 2019 – Pedagogy, Teaching, and Research in the Age of Digital Humanities, AIUCD '19, 2019, p. 251. URL: http://aiucd2019.uniud.it/wp-content/uploads/2020/03/AIUCD2019-BoA_DEF.pdf.

[11] M. Soprano, K. Roitero, S. Mizzaro, HITS Hits Readersourcing: Validating Peer Review Alternatives Using Network Analysis, in: M. K. Chandrasekaran, P. Mayr (Eds.), Proceedings of the 4th Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries co-located with the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, volume 2414 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019, pp. 70–82. URL: http://ceur-ws.org/Vol-2414/paper7.pdf.

[12] J. A. Drozdz, M. R. Ladomery, The Peer Review Process: Past, Present, and Future, British Journal of Biomedical Science 81 (2024). doi:10.3389/bjbs.2024.12054.

[13] C. Kadaifci, E. Isikli, Y. I. Topcu, Fundamental Problems in the Peer-Review Process and Stakeholders' Perceptions of Potential Suggestions for Improvement, Learned Publishing 38 (2025) e1637. doi:10.1002/leap.1637.

[14] Publons, 2018 Global State of Peer Review, Publons Report, 2018. URL: https://publons.com/static/Publons-Global-State-Of-Peer-Review-2018.pdf, accessed: 2024-12-18.

[15] A. Team, Peer Review: How We Found 15 Million Hours of Lost Time, AJE Scholarly Publishing Blog, 2023. URL: https://www.aje.com/arc/peer-review-process-15-million-hours-lost-time/, accessed: 2024-12-18.

[16] O. M. Smith, K. L. Davis, R. B. Pizza, R. Waterman, K. C. Dobson, B. Foster, J. C. Jarvey, L. N. Jones, W. Leuenberger, N. Nourn, E. E. Conway, C. M. Fiser, Z. A. Hansen, A. Hristova, C. Mack, A. N. Saunders, O. J. Utley, M. L. Young, C. L. Davis, Peer review perpetuates barriers for historically excluded groups, Nature Ecology & Evolution 7 (2023) 512–523. doi:10.1038/s41559-023-01999-w.

[17] S. Haffar, F. Bazerbachi, M. H. Murad, Peer Review Bias: A Critical Review, Mayo Clinic Proceedings 94 (2019) 670–676. doi:10.1016/j.mayocp.2018.09.004.

[18] J. L. Blackburn, M. D. Hakel, An Examination of Sources of Peer-Review Bias, Psychological Science 17 (2006) 378–382. doi:10.1111/j.1467-9280.2006.01715.x.

[19] W. Kaltenbrunner, S. Pinfield, L. Waltman, H. B. Woods, J. Brumberg, Innovating peer review, reconfiguring scholarly communication: an analytical overview of ongoing peer review innovation activities, Journal of Documentation 78 (2022) 429–449. doi:10.1108/JD-01-2022-0022.

[20] L. Waltman, W. Kaltenbrunner, S. Pinfield, H. B. Woods, How to improve scientific peer review: Four schools of thought, Learned Publishing 36 (2023) 334–347. doi:10.1002/leap.1544.

[21] T. Ross-Hellauer, What is open peer review? A systematic review, F1000Research 6 (2017) 588. doi:10.12688/F1000RESEARCH.11369.1.

[22] J. P. Tennant, J. M. Dugan, D. Graziotin, D. C. Jacques, F. Waldner, D. Mietchen, Y. Elkhatib, L. B. Collister, C. K. Pikas, T. Crick, P. Masuzzo, A. Caravaggi, D. R. Berg, K. E. Niemeyer, T. Ross-Hellauer, S. Mannheimer, L. Rigling, D. S. Katz, B. G. Tzovaras, J. Pacheco-Mendoza, N. Fatima, M. Poblet, M. Isaakidis, D. E. Irawan, S. Renaut, C. R. Madan, L. Matthias, J. N. Kjær, D. P. O'Donnell, C. Neylon, S. Kearns, M. Selvaraju, J. Colomb, A multi-disciplinary perspective on emergent and future innovations in peer review, F1000Research 6 (2017) 1151. doi:10.12688/F1000RESEARCH.12037.1.

[23] K. Kousha, M. Thelwall, Artificial intelligence to support publishing and peer review: A summary and review, Learned Publishing 37 (2024) 4–12. doi:10.1002/leap.1570.

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

[25] I. Boukhris, C. Zaâbi, A GAN-BERT based decision making approach in peer review, Social Network Analysis and Mining 14 (2024) 107. doi:10.1007/s13278-024-01269-y.

[26] Y. J. Park, D. Kaplan, Z. Ren, C.-W. Hsu, C. Li, H. Xu, S. Li, J. Li, Can ChatGPT be used to generate scientific hypotheses?, arXiv, 2023. doi:10.48550/arXiv.2304.12208. arXiv:2304.12208.

[27] V. Mollaki, Death of a reviewer or death of peer review integrity? the challenges of using AI tools in peer reviewing and the need to go beyond publishing policies, Research Ethics 20 (2024) 239–250. doi:10.1177/17470161231224552.

[28] A. Flanagin, J. Kendall-Taylor, K. Bibbins-Domingo, Guidance for Authors, Peer Reviewers, and Editors on Use of AI, Language Models, and Chatbots, JAMA 330 (2023) 702–703. doi:10.1001/jama.2023.12500.

[29] G. R. Latona, M. H. Ribeiro, T. R. Davidson, V. Veselovsky, R. West, The AI Review Lottery: Widespread AI-Assisted Peer Reviews Boost Paper Scores and Acceptance Rates, arXiv, 2024. doi:10.48550/arXiv.2405.02150. arXiv:2405.02150.

[30] W. Liang, Y. Zhang, H. Cao, B. Wang, D. Y. Ding, X. Yang, K. Vodrahalli, S. He, D. S. Smith, Y. Yin, D. A. McFarland, J. Zou, Can Large Language Models Provide Useful Feedback on Research Papers? A Large-Scale Empirical Analysis, NEJM AI 1 (2024) AIoa2400196. doi:10.1056/AIoa2400196.

[31] S. K. K. Santu, S. K. Sinha, N. Bansal, A. Knipper, S. Sarkar, J. Salvador, Y. Mahajan, S. Guttikonda, M. Akter, M. Freestone, M. C. W. Jr, Prompting LLMs to Compose Meta-Review Drafts from Peer-Review Narratives of Scholarly Manuscripts, arXiv, 2024. doi:10.48550/arXiv.2402.15589. arXiv:2402.15589.

[32] Y. Gao, D. Lee, G. Burtch, S. Fazelpour, Take Caution in Using LLMs as Human Surrogates: Scylla Ex Machina, arXiv, 2024. doi:10.48550/arXiv.2410.19599. arXiv:2410.19599.

[33] D. J. de Solla Price, Networks of Scientific Papers, Science 149 (1965) 510–515. doi:10.1126/science.149.3683.510.

[34] S. K. Banshal, S. Gupta, H. H. Lathabai, V. K. Singh, Power Laws in altmetrics: An empirical analysis, Journal of Informetrics 16 (2022) 101309. doi:10.1016/j.joi.2022.101309.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, Édouard Duchesnay, Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830. URL: http://jmlr.org/papers/v12/pedregosa11a.html.

[36] S. F. Olejnik, J. Algina, Generalized Eta and Omega Squared Statistics: Measures of Effect Size for Some Common Research Designs, Psychological methods 8 (2004) 434–47.

[37] N. Ferro, Y. Kim, M. Sanderson, Using Collection Shards to Study Retrieval Performance Effect Sizes, ACM Trans. Inf. Syst. 37 (2019). URL: https://doi.org/10.1145/3310364. doi:10.1145/3310364.
[38] N. Ferro, G. Silvello, A General Linear Mixed Models Approach to Study System Component Effects, in: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 25–34. doi:10.1145/2911451.2911530.
[39] N. Ferro, G. Silvello, Toward an anatomy of IR system component performances, Journal of the Association for Information Science and Technology 69 (2018) 187–200. doi:doi.org/10.1002/asi.23910.
[40] K. Roitero, B. Carterette, R. Mehrotra, M. Lalmas, Leveraging Behavioral Heterogeneity Across Markets for Cross-Market Training of Recommender Systems, in: Companion Proceedings of the Web Conference 2020, WWW '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 694–702. doi:10.1145/3366424.3384362.
[41] F. Zampieri, K. Roitero, J. S. Culpepper, O. Kurland, S. Mizzaro, On Topic Difficulty in IR Evaluation: The Effect of Systems, Corpora, and System Components, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19, Association for Computing Machinery, New York, NY, USA, 2019, p. 909–912. doi:10.1145/3331184.3331279.
[42] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, et al., SciPy 1.0: fundamental algorithms for scientific computing in Python, Nature Methods 17 (2020) 261–272. doi:10.1038/s41592-019-0686-2.
[43] M. Hubert, P. J. Rousseeuw, K. Vanden Branden, Robust Statistics: Theory and Methods, 2nd ed., Wiley, 2011. doi:10.1002/0470010940.
[44] E. Maddalena, S. Mizzaro, F. Scholer, A. Turpin, On Crowdsourcing Relevance Magnitudes for Information Retrieval Evaluation, ACM Transactions on Information Systems 35 (2017). doi:10.1145/3002172.
[45] F. Menczer, S. Fortunato, C. A. Davis, A First Course in Network Science, Cambridge University Press, 2020. doi:10.1017/9781108653947.
[46] D. Antipov, M. Buzdalov, B. Doerr, Lazy Parameter Tuning and Control: Choosing All Parameters Randomly from a Power-Law Distribution, 2023. doi:10.1007/s00453-023-01098-z.
[47] C. B. Clement, M. Bierbaum, K. P. O'Keeffe, A. A. Alemi, On the Use of ArXiv as a Dataset, arXiv, 2019. doi:10.48550/arXiv.1905.00075. arXiv:1905.00075.
[48] J. S. Park, C. Q. Zou, A. Shaw, B. M. Hill, C. Cai, M. R. Morris, R. Willer, P. Liang, M. S. Bernstein, Generative Agent Simulations of 1,000 People, arXiv, 2024. doi:10.48550/arXiv.2411.10109. arXiv:2411.10109.