electronics

MDPI

Article

Enhancing Peer Fairness via Data-Driven Analysis for Outlier Detection

Zhengkun Di [1], Jinqiannan Zhang [2,*], Weixing Tan [1] and Xiaoqi Sun [1]

[1] School of Software, Shandong University, Jinan 250014, China; 202200400002@mail.sdu.edu.cn (Z.D.); 202220799@mail.sdu.edu.cn (W.T.); sunxiaoqi@mail.sdu.edu.cn (X.S.)
[2] China Science and Technology Exchange Center, Beijing 100081, China
* Correspondence: zhangjqn@nrscc.gov.cn

Abstract: Fairness in peer review is of vital importance in academic activities. Current peer review systems focus on matching suitable experts with proposals but often ignore the existence of outliers. Previous research has shown that outlier scores in reviews could decrease the fairness of these systems. Therefore, outlier detection in peer review systems is essential for maintaining fairness. In this paper, we introduce a novel method that employs data-crossing analysis to detect outlier scores, aiming to improve the reliability of peer review processes. We utilize a confidential dataset from a review organization. Due to the inability to access ground truth scores, we systematically devise data-driven deviations from an estimated ground truth through data-crossing analysis. These deviations reveal inconsistencies and abnormal scoring behaviors of different reviewers. Subsequently, the review process is strengthened by providing a structured mechanism to identify and mitigate biases. Extensive experiments demonstrate its effectiveness in improving the accuracy and fairness of academic assessments, contributing to the broader application of AI-driven methodologies to achieve more reliable and equitable outcomes.

Keywords: anomaly detection; outlier detection; cross-match; data-driven; crowdsourcing

check for updates

Citation: Di, Z.; Zhang, J.; Tan, W.; Sun, X. Enhancing Peer Fairness via Data-Driven Analysis for Outlier Detection. *Electronics* **2024**, *13*, 4735. https://doi.org/10.3390/electronics13234735

Academic Editor: George A. Papadopoulos

Received: 24 October 2024
Revised: 23 November 2024
Accepted: 27 November 2024
Published: 29 November 2024

1. Introduction

According to the "2022 Annual Report" released by the NSFC [1], 536 reports related to research integrity were received during that year, and 533 cases were addressed. Among these, actions were taken against 397 individuals and 6 affiliated institutions, including issuing 82 public criticisms, revoking 74 funded projects, and canceling 112 project applications.

Peer review, in which experts in the field critically examine and provide feedback on scientific work—including proposals, research papers, personnel evaluations, and academic manuscripts—forms the cornerstone of scientific advancement and has underpinned the scientific enterprise for over three centuries [2]. Most academic peer-reviewed journals rely heavily on high-quality peer review, which serves two primary purposes. First, peer review acts as a filter to ensure that only research of substantial validity, significance, and originality is published, especially in prestigious journals. Second, it plays a developmental role by improving the quality of the manuscript. Peer reviewers contribute by offering constructive feedback, identifying areas for improvement, and suggesting corrections for errors, thereby preparing the manuscript for publication [3].

Despite the widespread use of peer review, research indicates that the process may not always be reliable [4–6]. Ideally, reviewers should be entirely objective; however, personal factors, such as differences in scoring criteria and individual biases, can lead to deviations from the ground truth, compromising the fairness of peer review outcomes. For instance, Kaatz et al. [7] highlighted the presence of gender bias in peer review, whereby reviewers may evaluate submissions differently based on the author's gender. Likewise, Iezzoni [8] identified significant disability discrimination, with authors with disabilities facing higher rejection rates or stricter evaluation standards. Smith et al. [9] also demonstrated that the peer review system tends to disadvantage authors from historically excluded groups, perpetuating biases throughout the review process.

To address these issues, numerous researchers have sought to improve peer review mechanisms from various angles. Linton [10] integrated the Black–Scholes model into a peer review framework to evaluate proposal value, concluding that sometimes proposals with the most divergent opinions among panelists should be selected, even if they did not receive the highest average score. Gai et al. [11] proposed a framework that integrates consensus and trust thresholds to enhance communication and collaboration, aiming to foster consensus by building mutual understanding and trust; however, both approaches lack empirical research validating their practical effectiveness. Cui et al. [12] developed a knowledge tracing model based on a multi-relational Transformer, modeling interactions between students and learning content in detail. This fine-grained approach provides valuable insights into analyzing reviewer behavior in peer review, offering a comprehensive understanding of decision-making patterns and behaviors. Similarly, in social network systems, Xu et al. [13] introduced the DMPS (dynamic modeling across propagation stages) framework to dynamically model the multi-stage propagation characteristics of disinformation in online networks. This method demonstrates the value of analyzing dynamic features across stages to improve consistency and reliability in evaluation systems. Likewise, Xu et al. [14] proposed the TCA (temporal context-aware) approach to address sparse data challenges by incorporating temporal and contextual factors, demonstrating the potential of contextual modeling in complex systems. While these methods are not directly related to peer review, their dynamic and context-aware principles may inspire new directions for enhancing peer review processes. In addition, Squazzoni et al. [15] attempted to incorporate incentive programs, finding that material rewards tended to reduce the quality and efficiency of peer review, as such incentives undermined reviewers' ethical motivations. Notably, the influence of individual reviewer assessments can significantly impact overall review outcomes, particularly when a minority of reviewers deviate substantially from the consensus, leading to considerable biases.

To overcome these challenges, automated evaluation methods have emerged as a means to enhance consistency and fairness in the review process. Automated techniques have already been successfully applied in fields like data management to handle complex evaluations, reduce human bias, and improve consistency. For example, automated anomaly detection has proven effective in encrypted traffic analysis by identifying irregularities without human intervention, indicating its potential application in peer review [16]. In medical imaging, automated data classification techniques have demonstrated improved diagnostic accuracy and efficiency, showing how automation can enhance complex decision-making by reducing human error [17]. These examples suggest that integrating automation into peer review could lead to more objective and consistent evaluations.

Traditional outlier detection methods, such as the median absolute deviation (MAD) [18] and trimmed mean [19], fail to quantify the impact of outliers on the entire peer review system, hindering further analysis of its dynamics. This paper introduces a novel data cross-matching approach (DCASP), which not only detects outliers in the peer review process but also quantifies their impact on collective consensus, offering new insights into the system's dynamics.

Data cross-matching is one such automated technique employed to identify and correct inconsistencies between related records from different datasets. This method ensures data completeness and accuracy by comparing and integrating records from various data sources. In bibliometric analysis, data cross-matching is used to integrate documents from multiple bibliographic sources, ensuring accurate citation links [20]. In astronomy, it is used to match and identify the same celestial objects across different astronomical catalogs, thereby enhancing data precision [21]. In clinical trials, an AI-based cross-matching system extracts data from electronic health records and matches them with relevant trials, improving patient selection accuracy [22].

Spearman's rank correlation coefficient is a well-established statistical tool used to assess the relationship between two variables based on their ranks [23]. The data-crossing analysis based on the extended Spearman's rank correlation for peer review (DCASP) method presented in this paper extends the application of the Spearman correlation coefficient to the peer review analysis. The DCASP method combines reviewers' scoring and ranking results to estimate the true value of each item using various analytical approaches. By focusing on rankings rather than raw scores, the DCASP method aims to provide a deeper understanding of the relationships within the ranked data, thereby enhancing discussions on individual and collective reviewer consensus while improving the transparency and accuracy of peer reviews.

Through cross-matching the rating sequences of reviewers with the ground truth, the DCASP method identifies outlier reviewers and conducts an in-depth analysis of their rating data to determine the items contributing to their outlier status. Further analysis, such as comparing correlation coefficients before and after the exclusion of specific proposals, enables the assessment of the impact of individual proposals on overall scoring. This approach helps identify the main proposals contributing to discrepancies between reviewer scores and the ground truth, offering insights into the dynamics of peer review that can help mitigate sources of bias.

This paper makes two primary contributions:

1. This paper proposes a data cross-matching approach based on the extended Spearman's rank correlation for peer review (DCASP). The method not only identifies outliers but also quantifies the degree to which individual reviewers deviate from the collective consensus, laying a foundation for further analysis of fairness in peer review. Ultimately, this contributes to enhancing the transparency and accountability of scholarly assessments. The introduction of the data-crossing analysis based on the extended Spearman's rank correlation for peer review (DCASP) represents a significant advancement in evaluating outliers among reviewers. The method quantifies the extent of deviations from consensus among reviewers and establishes a foundation for further analysis of fairness in peer review, contributing to greater transparency and accountability in scholarly assessments.
2. The DCASP method further quantifies the impact of individual proposal ratings on reviewers' deviation from the collective consensus. This approach provides a method for analyzing reviewer behavior patterns and offers valuable insights for improving the peer review system. The DCASP method also assesses the impact of individual proposals on the correlation between reviewer rankings and ground truth rankings. By generating a weighted sequence where each proposal's contribution to reviewer deviation is quantified, this approach helps pinpoint proposals that significantly influence outlier behavior, thereby enhancing our understanding of the factors contributing to reviewer inconsistencies.

The remainder of this paper is organized as follows: Section 2 reviews existing literature on enhancing the fairness of peer review systems, focusing on strategies for reducing biases and improving evaluation accuracy. Section 3 describes the DCASP methodology in detail, including its conceptual foundation, three approaches for deriving estimated ground truths, and methods for calculating reviewer outlier indices. Section 4 introduces the experimental setup used to validate the DCASP method, along with an analysis and summary of the experimental results. Section 5 concludes by summarizing the findings and discussing broader implications.

2. Related Work

To date, significant efforts have been undertaken to enhance the fairness and accuracy of peer review, which can be broadly categorized into two primary approaches:

1. Detecting Bias in Existing Peer Review Processes:
   Zardi et al. [24] proposed an innovative method for anomaly detection in network communities, integrating structural deviation analysis with attribute consistency

checks. This dual approach significantly improves the accuracy and reliability of anomaly detection. Kaatz et al. [7] conducted surveys and statistical analyses to examine gender bias in peer review, finding that female authors often receive more critical or unfair evaluations, indicating a correlation between reviewer gender and review outcomes. Iezzoni [8] conducted an experiment using double-blind review, submitting papers with and without disability-related information. By comparing the evaluations, they identified explicit disability bias and recommended measures to mitigate such biases in peer review. Stelmakh et al. [25] conducted an experiment at ICML 2020 and EC 2021 by assigning submissions to reviewers who had cited them. Their analysis revealed statistically significant differences in the behavior of cited versus uncited reviewers, indicating the presence of citation bias. Smith et al. [9] utilized a dataset comprising 312,740 biological sciences manuscripts across 31 studies to analyze peer review outcomes based on author demographics, revealing that peer review may systematically disadvantage authors from historically excluded groups by perpetuating biases throughout the review process.

2. Introducing Strategies and Mechanisms to Improve the Peer Review System:
Linton [10] applied the Black–Scholes model to the peer review process, using research characteristics such as innovativeness, expected impact, and implementation difficulty as input parameters. This approach quantifies the potential value of high-risk, high-reward research, offering a novel perspective for assessment. Shayegan et al. [26] introduced a collective anomaly detection method to identify fraudulent activities within the Bitcoin network by detecting anomalies at the user level with a trimmed K-means algorithm, thereby improving fraud detection. Fernández et al. [27] developed a group multi-objective optimization framework using multi-criteria ordinal classification to handle imperfect information and maximize overall group satisfaction. Applied to project portfolio optimization, this method demonstrated its effectiveness in reconciling diverse objectives in group decision-making. Papadopoulos et al. [28] investigated a peer review setting in which students had the freedom to select multiple reviews, finding that providing choice positively influenced students' attitudes, suggesting that peer review deficiencies could be mitigated through self-review processes. Hosseini and Horbach [29] examined the potential of large language models (LLMs) in assisting with peer review writing and decision letter drafting, potentially increasing productivity while addressing reviewer shortages. However, concerns were raised regarding biases, data privacy, and review integrity due to the opaque nature of LLM training data and algorithms. Ji and Ma [30] developed an enhanced consensus model incorporating risk aversion strategies to optimize decision-making in uncertain environments by considering experts' risk preferences. Gai et al. [11] proposed an advanced consensus-enhancing method for large groups, employing a bidirectional feedback mechanism based on consensus and trust thresholds. This hybrid feedback strategy aims to improve group cohesion while minimizing opinion adjustment efforts.

Building on extensive prior research, this study introduces an innovative data-crossing analysis methodology designed to address challenges in the peer review system and enhance both the accuracy and fairness of evaluations. The subsequent sections will elaborate on the proposed methodology, the experimental setup, and the evidence supporting the effectiveness and validity of our approach.

3. DCASP Method

3.1. Preliminary

3.1.1. Data Cross-Matching

In the context of anomaly detection within peer review systems, we employ data cross-matching techniques, which are extensively utilized in fields such as astronomy. Data cross-matching involves comparing and aligning records from different datasets to ensure consistency and accuracy. This approach systematically identifies and resolves discrepancies by cross-referencing data points across multiple sources. By detecting outliers and inconsistencies, cross-matching maintains data integrity, ensuring that the final dataset is comprehensive and reliable [31].

3.1.2. Spearman's Correlation Coefficient

Spearman's rank correlation coefficient measures the monotonic relationship between two variables. Unlike Pearson's correlation, Spearman's coefficient is based on ranks rather than actual values, making it suitable for non-normally distributed datasets or those containing outliers. The formula for Spearman's coefficient is as follows:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{1}$$

where d_i represents the difference between the ranks of the two variables for the i-th observation, and n is the total number of observations. The coefficient r_s ranges from -1 to 1, where $r_s = 1$ indicates a perfect positive correlation, $r_s = -1$ indicates a perfect negative correlation, and $r_s = 0$ indicates no correlation.

3.2. DCASP Method

3.2.1. Overview

In this section, we present the DCASP method for detecting outlier reviewers in peer review systems. Traditional anomaly detection methods, such as the median absolute deviation (MAD) [18], the trimmed mean [19], and the 1.5 IQR rule [32], primarily focus on the values themselves and do not account for the significance of rankings. In the context of peer review, relying solely on numerical scores to identify outliers is insufficient, or even unreasonable, due to the inherent personal biases among reviewers. It is natural for individual reviewers to assign different scores to the same proposal. Instead, our approach differs from traditional outlier detection methods by emphasizing rankings rather than raw scores. The core idea is that while individual scores may vary, the overall ranking of proposals should remain relatively consistent. This approach also helps avoid situations where some reviewers consistently give high or low scores to all proposals. If the scores lack differentiation, our method can still effectively identify these outliers. We begin by introducing three different ground truth computation methods to approximate consensus rankings. We then discuss the rank computation process, including how we handle ties. Subsequently, we describe our cross-matching-based outlier detection approach using Spearman's rank correlation coefficient. Finally, we introduce the computation of outlier weights to identify specific proposals contributing to a reviewer's deviation from the consensus.

3.2.2. Ground Truth Computation

Reviewers exhibit diverse preferences, cognitive processes, and evaluative criteria, making the establishment of an absolute ground truth challenging. To address this, we experimented with various methodologies to derive a suitable estimated ground truth. We employ three representative scoring methods to comprehensively analyze deviations among reviewer outliers from different perspectives. These methodologies enhance our understanding of consensus within peer review settings. Throughout this paper, the ground truths derived from these strategies are referred to as Ground Truth 3, Ground Truth 2, and Ground Truth 3.

Given the score matrix D, with m proposals and n reviewers, we compute the ground truth score $\bar{S}$ in the following three ways:

1. Ground Truth 1:
   Ground Truth 1 approximates the ground truth by averaging the ratings for each proposal after excluding the highest and lowest scores to reduce bias and mitigate the influence of extreme values. For each proposal j, all reviewers P_i provide ratings S_{ij},

where S_{ij} is the rating given to proposal j by reviewer P_i, and n is the total number of reviewers. The ratings are sorted in ascending order:

$$S_{(1)j} \leq S_{(2)j} \leq \cdots \leq S_{(n)j},$$

where $S_{(1)j}$ and $S_{(n)j}$ are the lowest and highest ratings, respectively. The lowest and highest ratings are excluded, and the modified average score $\bar{S}_j$ is computed as follows:

$$\bar{S}_j = \frac{1}{n-2} \sum_{k=2}^{n-1} S_{(k)j}.$$

2. Ground Truth 2:
   Ground Truth 2 uses a ranking point system in which each proposal is assigned points based on its rank by each reviewer. Specifically, for each proposal j, the rank r_{ij} assigned by reviewer P_i is determined, with the lowest-rated proposal receiving a rank of 1, the next lowest a rank of 2, and so on. The total points for proposal j are calculated as follows:

   $$T_j = \sum_{i=1}^{n} (n - r_{ij} + 1).$$

   The normalized average score $\bar{S}_j$ is then defined as follows:

   $$\bar{S}_j = \frac{T_j}{n}.$$

3. Ground Truth 3:
   Ground Truth 3 employs a recommendation voting mechanism that differentiates proposals with equal votes based on their average scores. Specifically, for each proposal j, the recommendation vote V_{ij} is defined as follows:

   $$V_{ij} = \begin{cases} 1 & \text{if } S_{ij} > 75, \\ 0 & \text{otherwise.} \end{cases}$$

   The total recommendation votes for proposal j are then computed as follows:

   $$V_j = \sum_{i=1}^{n} V_{ij}.$$

   The average score for each proposal is calculated as follows:

   $$\bar{S}_j = \frac{1}{n} \sum_{i=1}^{n} S_{ij}.$$

3.2.3. Rank Computation

Given the ground truth scores $\bar{S}$ obtained from the three methods of ground truth calculation, we determine the rank of each proposal by sorting these scores, resulting in the ground truth ranks, denoted as R^t. Similarly, by sorting the ratings in the raw peer review data, we obtain the rankings assigned by each reviewer, denoted as R^r.

Handling ties during ranking is inevitable. For instance, if multiple proposals are tied for a particular rank, assigning them the same rank can lead to inconsistencies in subsequent rankings. To address this, we use an average ranking approach. Specifically, if a subset $\mathcal{T}$ consisting of $m_{\mathcal{T}}$ proposals is tied at rank r, each proposal in $\mathcal{T}$ is assigned a rank calculated as follows:

$$R = \frac{2r + m_{\mathcal{T}} - 1}{2} \tag{2}$$

As illustrated in Figure 1, if two proposals are tied for first place, each is assigned a rank of 1.5. This method ensures a balanced representation of each proposal's standing relative to others in cases of equivalence.

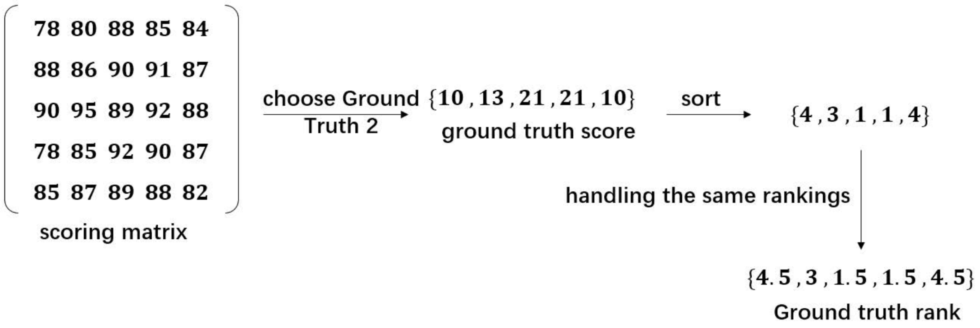


Figure 1. Example of ground truth rank computation.

3.2.4. Cross-Matching-Based Outlier Detection

In this section, we present the computation of the outlier index.

Outlier index computation

Given that peer review data consist of discrete ratings and rankings, we integrate Spearman's rank correlation coefficient into our cross-analysis approach. This allows us to evaluate the correlation between reviewers' ranking sequences and the estimated sequence of underlying truths. Spearman's rank correlation coefficient is particularly well-suited for ordinal data, providing a robust means of assessing the alignment between reviewers' assessments and established benchmarks.

The formula for Spearman's correlation coefficient is given as follows:

$$\rho_i = 1 - \frac{6\sum_{j=1}^{m} d_j'^2}{m(m^2-1)} \tag{3}$$

where

$$d_j' = R_j^t - R_{ij}^r \tag{4}$$

In this equation, ρ_i denotes the rank correlation coefficient for reviewer P_i, d_j' represents the difference between the ground truth rank R_j^t and the rank R_{ij}^r assigned by reviewer P_i to proposal j, and m is the total number of proposals.

Once the estimated ground truth is established, reviewers are ranked based on their scores in descending order. We compute the Spearman rank correlation coefficient between each reviewer's ranking and the three ground truth rankings using the cross-matching method described earlier. The correlation coefficient ranges from −1 to 1, measuring the alignment between each reviewer's ranking and the ground truth. Specifically, the Spearman correlation coefficient is calculated by comparing the ground truth ranking sequence with each reviewer's ranking sequence, which is derived directly from sorting their scores in the original data.

Ranking reviewers solely by this coefficient may not adequately highlight differences in deviations from the ground truth. To address this limitation, we introduce the *outlier index*, which amplifies the discrepancies between reviewers and the ground truth based on correlation coefficients. Given the diverse contexts of peer review, defining a universal threshold for excessive deviation is challenging. Therefore, we implement a dynamic thresholding approach that can be tailored to different scenarios. In this experiment, we define the threshold interval to range from −0.5 to 0.9 with incremental steps of 0.1. Starting at a threshold of −0.5, for each subsequent threshold value (i.e., −0.4, −0.3, . . ., 0.9), reviewers with correlation coefficients at or below the current threshold are identified and recorded. For each of these reviewers, the outlier index is calculated using the following formula:

$$\text{Outlier Index}_i = \sum_{s=1}^{\|Thres\|} \frac{10}{(s+1)} \cdot \mathbb{I}(\rho_i \geq \text{Thres}_s) \tag{5}$$

where s represents the threshold step, ranging from 1 to the total number of steps (in this case, 16). ρ_i is the Spearman correlation coefficient for reviewer i, and $\mathbb{I}(\cdot)$ is the indicator function, which is 1 if ρ_i is greater than or equal to the threshold Thres_s, and 0 otherwise. The outlier index is calculated by summing the weighted indicators across all threshold steps for each reviewer. This dynamic calculation assigns greater weight to reviewers with smaller correlation coefficients as the threshold increases. The stepwise approach captures varying degrees of deviation from the ground truth, allowing for a more detailed identification of outliers. By adjusting the threshold range and step size, the method is adaptable to different peer review contexts, ensuring flexibility and robustness in outlier detection.

Outlier Weight Computation

While our methodology successfully identifies reviewers with substantial deviations from the ground truth, pinpointing the specific proposals causing these deviations remains unresolved. To address this, we introduce the concept of *proposal weights* w_{ij}, which quantifies the influence of each proposal j on reviewer P_i's alignment with the ground truth. The weight w_{ij} is based on the relative change in the correlation coefficient ρ_i before and after the removal of proposal j.

After removing proposal j, we recalculate the ground truth rank using the same method described earlier and compute the revised correlation coefficient ρ_i^{-j}. The weight w_{ij} is then computed as follows:

$$w_{ij} = \frac{\rho_i^{-j}}{\rho_i} \tag{6}$$

where we have the following:

- ρ_i: Original correlation coefficient for reviewer P_i.
- ρ_i^{-j}: Correlation coefficient for reviewer P_i after recalculating the ground truth and removing proposal j.
- w_{ij}: Weight of proposal j for reviewer P_i.

The interpretation of w_{ij} depends on the sign of the original correlation coefficient ρ_i:

- When $\rho_i > 0$:
  - $w_{ij} > 1 \Rightarrow$ Removal of proposal j increases the correlation, suggesting that the proposal's original ranking was disproportionately influential and possibly erroneous.
  - $w_{ij} < 1 \Rightarrow$ Removal of proposal j decreases the correlation, indicating the proposal's original ranking was reasonable and stabilizing.
- When $\rho_i < 0$:
  - $w_{ij} > 1 \Rightarrow$ Removal of proposal j decreases the correlation, suggesting that the proposal's original ranking was overly detrimental.
  - $w_{ij} < 1 \Rightarrow$ Removal of proposal j increases the correlation, indicating the proposal's original ranking was beneficial and contributed positively to the overall assessment.

To improve consistency in weight interpretation and enhance clarity, particularly for visualization purposes, we refine the calculation method by normalizing the weights around zero as follows:

$$w'_{ij} = w_{ij} - 1 = \frac{\rho_i^{-j} - \rho_i}{|\rho_i|} \tag{7}$$

The refined weight w'_{ij} can be interpreted as follows:

- $w'_{ij} > 0$: Removing proposal j increased the correlation, indicating that proposal j's original ranking negatively influenced the overall correlation.
- $w'_{ij} < 0$: Removing proposal j had a stabilizing effect, implying that its original ranking was more reasonable.
- $|w'_{ij}|$: The magnitude of w'_{ij} reflects the strength of the proposal's impact on the correlation.

This refined method of calculating proposal weights ensures that the influence of each proposal's removal is quantified in a way that is both intuitive and visually interpretable, offering clearer insights into how individual proposals affect the overall evaluation process. The pseudocode for the proposal weighting algorithm is shown in Algorithm 1.

Algorithm 1 DCASP algorithm.

```
Input:
  $D$: Score matrix from data
  $m$: The number of proposals
  $n$: The number of reviewers
  $str$: Strategy to estimate the ground truth
  $k$: Output the first $k$ outlier reviewer numbers
Output:
  $X$: Sequence of reviewer numbers in the top $k$ of the outlier index
  $Y$: Ranking of weights for each proposal by reviewers in $X$
$M, M^{'} \leftarrow \emptyset$    ▷ Initialize the set used to store the sorted results
$\bar{S} \leftarrow Groundtruth(D, str, m, n)$    ▷ Ground truth scores are obtained according to the ground truth strategy
$R^t, R^r \leftarrow Rank(D, \bar{S}, m, n)$    ▷ Get the ground truth rank $R^t$ and reviewer rank $R^r$
$\rho \leftarrow Crossmatching(R^t, R^r)$ ▷ Obtain correlation coefficient $\rho$ using the cross-matching method
$P \leftarrow$ Compute Outlier Index using $\rho$ by Equation (5)
$O \leftarrow$ Sort $P$ in descending order
$X \leftarrow Select(k, O)$    ▷ Get the top $k$ reviewer numbers as needed
for each reviewer $P_i$ in $X$ do
    for each proposal $j$ in $1 \ldots m$ do
        $D^{'} \leftarrow Delete(D, j)$    ▷ Getting the scoring data after deleting the proposal $j$
        $\bar{S}^{'} \leftarrow Groundtruth(D^{'}, str, m-1, n)$    ▷ Get the average rating after deleting proposal $j$
        $R^{t'}, R^{r'} \leftarrow Rank(D^{'}, \bar{S}^{'}, m-1, n)$ ▷ Get ground truth rank $R^{t'}$ and reviewer rank $R^{r'}$ after deleting proposal $j$
        $\rho^{-} \leftarrow Crossmatching(R^{t'}, R^{r'})$    ▷ Calculate new correlation coefficient $\rho^{-}$ after deletion
        $W_{ij} \leftarrow$ Compute Outlier Weight using $\rho, \rho^{-}$ by Equation (7)
    end for
end for
for each reviewer $P_i$ in $X$ do
    $Y_i \leftarrow$ Sort $W_{i*}$ in descending order
end for
return $X, Y$
```

4. Experiment

In this section, we present a series of confirmatory and exploratory experiments conducted to validate the methodology outlined previously. This comprehensive experimental approach not only confirms the robustness of our methodology but also enhances our understanding of its applicability in various evaluative contexts. In this experiment, the threshold increment was set at 0.1, which often results in cases where multiple reviewers share the same outlier index. In such instances, we prioritize reviewers with lower ordinal numbers for ranking.

4.1. Dataset

In our experimental data, we obtained confidential peer review ratings from a review organization, collected via a standardized review process, to validate the accuracy and applicability of our method. Table 1 presents an example of the reviewers' ratings, where R_i represents the i-th reviewer, and O_j represents the j-th proposal. All reviewers provided ratings on a percentage scale for each proposal. Each group of data is independent, representing a set of proposals evaluated by a corresponding set of reviewers, with both the number of proposals and the number of reviewers varying across different groups. In the experiments, we used two groups of data: Group one involved ratings from non-trusted reviewers obtained from a review organization, used for validation to compare the effectiveness of our method with MAD and trimmed mean. Group two utilized original data to demonstrate the accuracy of our method and support further analysis.

Table 1. Example data.

	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8
R_1	90	88	87	73	92	77	89	78
R_2	70	85	79	80	72	84	67	75
R_3	89	90	93	85	79	85	72	92
R_4	86	80	73	75	90	88	78	70
R_5	85	88	95	80	82	94	90	84

4.2. Baselines

We applied two conditional methods—MAD and trimmed mean—to identify anomalous reviewer scores for each proposal.

In the MAD method, we first calculate the median M_j of reviewer scores S_{ij} for each proposal. The absolute deviation D_{ij}^M from the median is given by the following:

$$D_{ij}^M = |S_{ij} - M_j| \tag{8}$$

The MAD for each proposal, denoted as MAD_j, is then calculated as the median of these deviations:

$$\text{MAD}_j = \text{median}(D_{ij}^M) \tag{9}$$

Scores are considered outliers if their deviation from the median exceeds a threshold of three times the MAD:

$$D_{ij}^M > 3 \times \text{MAD}_j \tag{10}$$

In the trimmed mean approach, we first sort the reviewer scores S_{ij} for each proposal in ascending order and then remove the highest and lowest scores. The trimmed mean of the remaining $n - 2$ scores is denoted as $\tilde{S}_j$. The deviations from the trimmed mean are calculated as follows:

$$D_{ij}^T = |S_{ij} - \tilde{S}_j| \tag{11}$$

The average of these deviations, denoted as $\tilde{D}_j$, is then calculated as follows:

$$\tilde{D}_j = \frac{1}{n-2}\sum_{i=2}^{n-1} D_{ij}^T \tag{12}$$

A score is considered an outlier if its deviation exceeds three times the trimmed mean of the deviations:

$$D_{ij}^T > 3 \times \tilde{D}_j \tag{13}$$

In both methods, the threshold factor was consistently set to three to reduce the influence of extreme scores, ensuring a reliable and robust measure of variability and central tendency. We then aggregated the outliers identified across all proposals and ranked them by frequency of occurrence, selecting the top three reviewers with the highest frequencies as the overall outliers. For the DCASP method, we used Ground Truth 3, deemed more reasonable, to identify outliers.

4.3. Quantitative Analysis of Group One

We adopt the proposed method and two baselines i.e., MAD and trimmed mean on collected review data from a private peer review platform. To give a quantitative performance of the proposed method, we also collect the outlier rank of reviews with the help of the review organization (e.g., $Rank(R_1) < Rank(R_2)$ means that reviewer R_1 is more like an outlier than R_2 and might be firstly outputted in the algorithm). The review scores are shown in Table 2.

Table 2. Scoring by group one of proposal reviewers.

	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9	O_{10}	O_{11}
R_1	85	84	66	74	70	94	94	76	94	87	85
R_2	92	65	66	94	74	72	90	76	73	74	78
R_3	63	80	81	72	79	89	86	74	92	73	67
R_4	70	88	73	89	82	94	94	92	80	66	74
R_5	94	96	71	94	88	80	80	86	83	82	74
R_6	92	95	84	80	95	81	78	66	87	74	73
R_7	79	82	80	81	90	76	78	81	71	72	73
R_8	81	89	62	76	86	73	91	69	78	73	73
R_9	91	90	74	94	93	74	74	63	86	66	66
R_{10}	74	74	84	90	90	86	80	72	86	82	82
R_{11}	80	79	76	86	69	94	74	73	87	74	83
R_{12}	97	96	90	94	88	80	94	86	94	72	72
R_{13}	89	84	87	80	89	78	80	74	55	74	74
R_{14}	74	72	83	78	88	76	74	74	74	74	73
R_{15}	67	89	79	86	84	73	94	73	73	74	74
R_{16}	92	80	74	90	70	74	65	86	73	74	53
R_{17}	86	92	80	86	84	70	71	66	62	53	74

We use the rank-based score $DCG@k$ as the evaluation metric, which is calculated via the following equation:

$$DCG@k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2(i+1)} \tag{14}$$

In practice, we take into consideration the fact that the number of outlier reviews is small among the reviewers. Therefore, we use $DCG@1$, $DCG@2$, $DCG@3$ as the metrics. The quantitative analysis results of the proposed methods along with the two baselines are shown in Table 3. DCASP achieves the highest values in $DCG@1$, $DCG@2$, and $DCG@3$, indicating its superior effectiveness in identifying highly relevant outliers.

Table 3. Comparison of MAD, trimmed mean, and DCASP methods.

Method	MAD	Trimmed Mean	DCASP
DCG@1	0	1	**3**
DCG@2	0.631	1	**3.631**
DCG@3	2.131	1	**3.631**

4.4. Results and Analysis of the Group Two

Group two is presented in Table 4.

Table 4. Scores from group two of proposal reviewers.

	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9	O_{10}	O_{11}	O_{12}	O_{13}
R_1	57	73	84	63	88	82	56	83	85	75	86	63	62
R_2	88	84	87	87	81	82	90	89	85	86	87	91	82
R_3	80	82	72	83	71	80	74	90	72	84	70	92	74
R_4	79	92	66	62	68	55	64	60	71	70	61	68	71
R_5	76	51	82	67	71	59	78	48	54	71	46	49	52
R_6	62	60	75	60	55	53	70	55	62	70	62	77	77
R_7	73	72	88	73	65	69	85	74	72	68	73	84	69
R_8	87	74	77	72	73	74	79	72	73	73	74	72	74
R_9	77	78	79	80	71	72	73	72	74	80	78	77	73
R_{10}	88	78	90	80	84	92	90	88	82	86	86	93	92
R_{11}	70	69	64	60	77	70	72	71	75	78	79	73	70
R_{12}	80	70	88	72	76	82	72	70	88	90	70	85	86
R_{13}	86	73	83	71	73	87	83	72	80	79	70	72	82
R_{14}	85	65	85	75	85	70	75	71	72	80	61	90	82
R_{15}	84	85	79	72	70	69	88	86	71	69	72	82	76

- Ground Truth 1: [5.0, 7.0, 6.0, 10.0, 3.0, 1.0, 2.0, 8.0, 4.0, 9.0, 11.0, 12.0, 13.0]
- Ground Truth 2: [2.5, 9.0, 1.0, 12.5, 12.5, 11.0, 4.0, 8.0, 7.0, 5.0, 10.0, 2.5, 6.0]
- Ground Truth 3: [2.0, 10.0, 1.0, 12.0, 8.0, 7.0, 5.0, 11.0, 9.0, 3.0, 13.0, 4.0, 6.0]

In the subsequent sections, we will analyze one or two reviewers who exhibit the highest outlier index across the three ground truths. This analysis aims to validate the accuracy of DCASP.

As illustrated in Figure 2, using Ground Truth 1, the reviewer rankings are as follows: [6, 9, 3, 2, 4, 14, 7, 10, 11, 12, 15, 1, 8, 5, 13]. Reviewer nos. 6 and 9 are identified as the most significant outliers. Reviewer no. 6's rankings of proposals exhibit notable deviations; for instance, Proposal no. 6 is ranked first by Ground Truth 1 but thirteenth by Reviewer no. 6. Similarly, Proposal nos. 12 and 13, ranked twelfth and thirteenth by Ground Truth 1, are both ranked first by Reviewer no. 6. Reviewer no. 9 also shows a substantial divergence in rankings, particularly for Proposal nos. 4, 5, 6, 7, 10, and 11, with his rankings differing significantly from those of Ground Truth 1. These discrepancies between the rankings by Reviewers nos. 6 and 9 and Ground Truth 1 suggest that variations in personal preferences alone cannot account for the observed gaps. Consequently, it is reasonable to classify them as the most significant outliers in the context of Ground Truth 1.

The data reveal that the outlier index values range broadly from 2.775 to 33.807. Notably, certain reviewers, such as Reviewer no. 5, exhibit significantly higher outlier indices, suggesting their evaluations deviate markedly from the collective scoring norm. Conversely, multiple reviewers sharing the same outlier index, such as Reviewer nos. 9, 10, and 11, each with a value of 10.974, might indicate a similar scoring criterion or behavioral pattern among them. Reviewer no. 14, possessing the lowest outlier index, appears to align most closely with the consensus, potentially indicating the highest consistency with the collective evaluations. This variation in indices underscores the effectiveness of the outlier recommendation algorithm in distinguishing among reviewers, enhancing our understanding of individual and collective scoring behaviors within this dataset.

As illustrated in Figure 3, the weighting of proposals for Reviewer no. 6, derived from Ground Truth 1, is sequenced as [6, 12, 13, 5, 11, 8, 9, 1, 2, 3, 4, 7, 10]. Focusing on Proposal nos. 6 and 12, we examine the impact of their removal on the correlation coefficients. Initially, the raw correlation coefficient stands at −0.506. Upon removal of Proposal no. 6, the reordered ground truths are [4.0, 6.0, 5.0, 9.0, 2.0, 1.0, 7.0, 3.0, 8.0, 10.0, 11.0, 12.0], and the reviewer's reordered ratings are [7.0, 9.5, 3.0, 9.5, 11.5, 4.5, 11.5, 7.0, 4.5, 7.0, 1.5, 1.5]. This results in a new correlation coefficient of −0.369 and a calculated weight of 0.270. Similarly, after removing Proposal no. 12, the ground truths are reordered to [5.0, 7.0, 6.0, 10.0, 3.0, 1.0, 2.0, 8.0, 4.0, 9.0, 11.0, 12.0], and the reviewer rankings to [6.0, 8.5, 2.0, 8.5, 10.5, 12.0, 3.5, 10.5, 6.0, 3.5, 6.0, 1.0], with a correlation coefficient of −0.372 and a weight of 0.265. These calculated weights align with the outcomes estimated by the algorithm, confirming the accuracy of our method in assessing the influence of individual proposals on the correlation metrics.

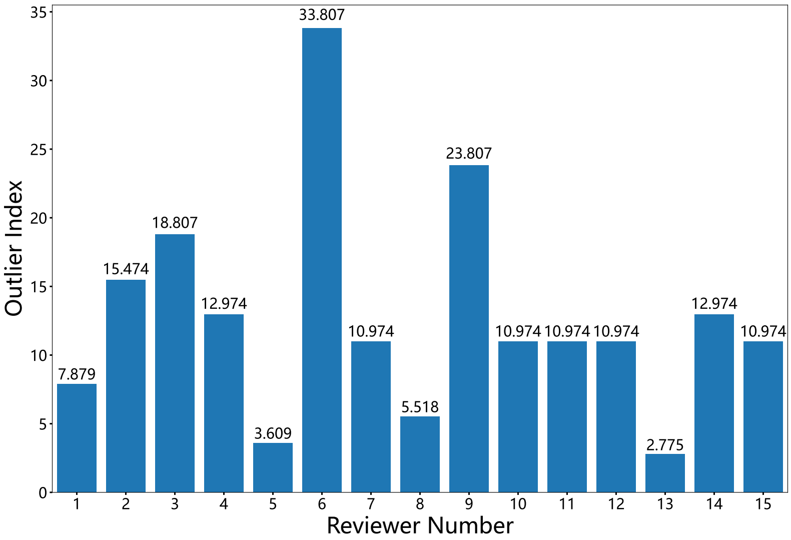


Figure 2. Reviewer outlier index graph generated based on Ground Truth 1.

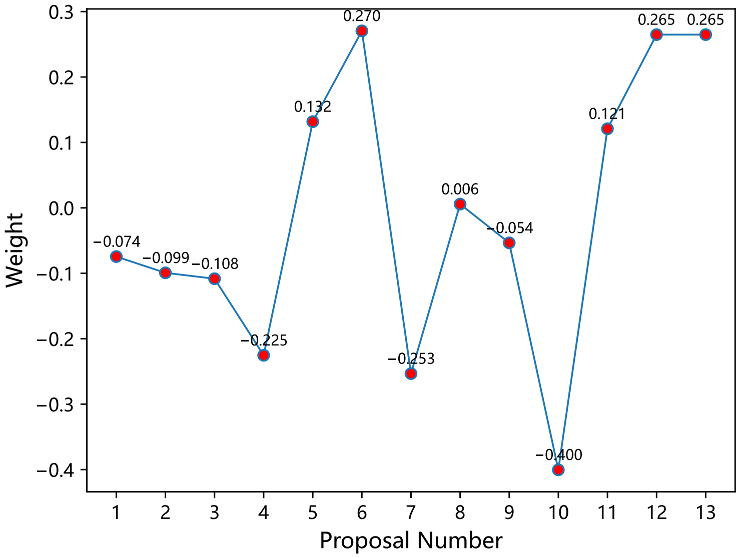
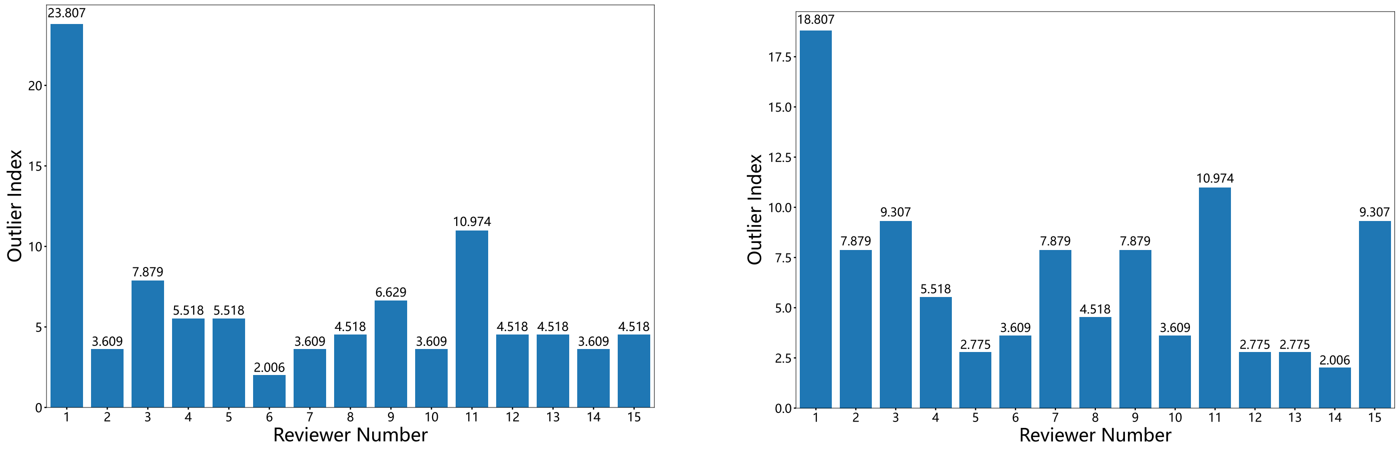


Figure 3. Weights calculated by Reviewer no. 6 based on Ground Truth 1.

The graphical representation illustrates a pronounced fluctuation in the weights assigned to the proposals, ranging from −0.400 for Proposal no. 9 to +0.265 for Proposal nos. 6 and 12. This variability suggests a notable inconsistency in the reviewer's evaluations of different proposals relative to the estimated ground truths. Such fluctuations underscore the likelihood of personal biases influencing the reviewer's assessments. This observation corroborates the previously determined outlier status of this reviewer, affirming that the evaluations deviate significantly from the collective norm. To enhance fairness in peer review, Proposal nos. 6, 12, and 13 may share certain common characteristics, while Reviewer no. 6's understanding of this domain significantly deviates from that of other reviewers. Targeted training or reassigning this reviewer away from similar proposals could mitigate such inconsistencies.

As depicted in Figure 4a,b, focusing on Ground Truth 2, Reviewer nos. 1 and 11 are identified as the most significant outliers. Reviewer no. 1's rankings of proposals are [12.0, 8.0, 4.0, 9.5, 1.0, 6.0, 13.0, 5.0, 3.0, 7.0, 2.0, 9.5, 11.0], and Reviewer no. 11's are [11.0, 4.5, 9.0, 10.0, 4.5, 4.5, 4.5, 7.0, 1.0, 8.0, 2.0, 9.0, 11.0, 12.0, 13.0, 3.0, 9.0, 6.0, 7.0, 4.0, 2.0, 1.0, 5.0, 9.0]. For Reviewer no. 1, the rankings for Proposal nos. 1, 5, 7, and 11 show significant deviations from their respective ground truths, highlighting a stark divergence in evaluation criteria. Similarly, for Reviewer no. 11, Proposal nos. 1, 3, 5, and 11 also display considerable misalignments from the ground truths. These findings illustrate the pronounced outlier behavior of these reviewers and validate the use of these ground truths in identifying deviations in reviewer assessments.



(**a**)

(**b**)

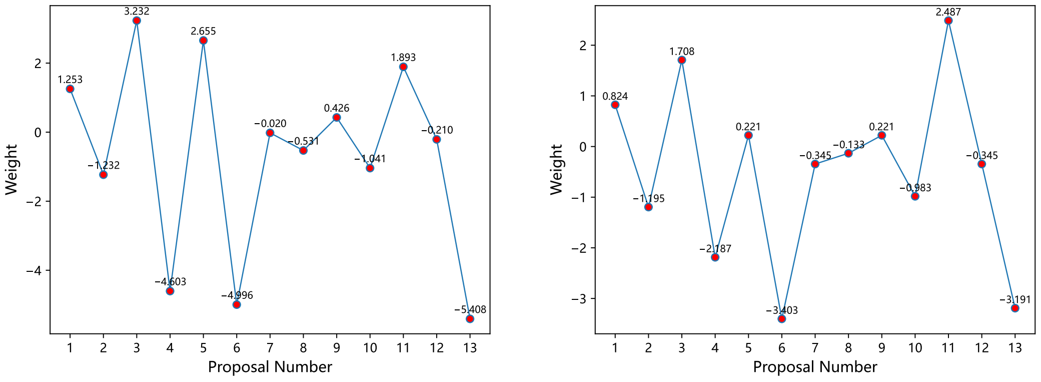
Figure 4. Reviewer outlier index graphs based on different ground truths. (**a**) Reviewer outlier index graph based on Ground Truth 2. (**b**) Reviewer outlier index graph based on Ground Truth 3.

Similar to the observations from Ground Truth 2, Ground Truth 3 identifies Reviewer nos. 1 and 11 as exhibiting the highest degree of outlier behavior. Specifically, for Reviewer no. 1, the proposals markedly divergent from the ground truths remain consistent with those previously noted, underscoring a persistent deviation in his assessments. In contrast, for Reviewer no. 14, notable deviations are observed in the evaluations of Proposal nos. 1, 3, and 11, indicating significant discrepancies from the ground truths. These findings reaffirm the appropriateness of the algorithm's classification of Reviewer nos. 1 and 11 as the most significant outliers. The consistency of these results across different ground truths supports the robustness of the outlier detection methodology and underscores its effectiveness in identifying reviewers whose evaluations consistently deviate from the established consensus.

Compared to Ground Truth 1, we observe a significant reduction in the maximum outlier index when utilizing Ground Truth 2, indicative of a tendency towards tighter clustering among reviewers. With the implementation of Ground Truth 3, the outlier index not only decreases further but also demonstrates a more uniform distribution across reviewers. Notably, the mean outlier index calculated from Ground Truth 1 is higher than those derived from Ground Truths 2 and 3. This suggests that reviewer evaluations under Ground Truths 2 and 3 are more consistent with the collective consensus compared to those under Ground Truth 1. However, a higher mean outlier index in Ground Truth 1, while reflecting greater variability among reviewers, may facilitate a clearer differentiation of outlier degrees among them. This condition suggests that while Ground Truths 2 and 3 provide a more consolidated view of reviewer consensus, they may obscure the identification of extreme deviations that are critical to understanding reviewer biases and the overall reliability of the peer review process.

An examination of the outlier reviewer recommendation plots derived from each ground truth reveals significant similarities between the rankings provided by Ground Truths 2 and 3, while both differ markedly from those determined by Ground Truth 1. Ground truths serve as quantitative representations of the collective consensus, and the observed discrepancies suggest that Ground Truth 1 encapsulates a distinct quantification of this consensus compared to the other two ground truths. Furthermore, across all three ground truths, Reviewer no. 5 consistently exhibits a lower outlier index, indicating that his evaluations align closely with the collective consensus, irrespective of the measurement perspective. Conversely, Reviewer no. 11 maintains a consistent outlier index across all ground truth conditions, positioned within the medium range. This consistency highlights that while Reviewer no. 11's evaluations do not represent extreme deviations, they consistently diverge to some extent from the collective consensus. These insights underscore the nuanced differences in how each ground truth captures reviewer biases and the overall dynamics within the peer review process.

Figure 5a,b present the weighting analysis for Reviewer no. 11, detailing the sequence in which proposals influence the overall correlation coefficients based on Ground Truths 2 and 3. The sequence derived from Ground Truth 2 is [3, 5, 11, 1, 9, 7, 12, 8, 10, 2, 4, 6, 13], with an initial correlation coefficient of −0.055. In contrast, the sequence from Ground Truth 3 is [11, 3, 1, 5, 9, 8, 7, 12, 10, 2, 4, 13, 6], with an initial correlation coefficient of −0.099. Taking Proposal no. 3 as an example, upon its removal, the reordered ratings by Reviewer no. 11 are [9.0, 11.0, 12.0, 3.0, 9.0, 6.0, 7.0, 4.0, 2.0, 1.0, 5.0, 9.0]. Correspondingly, the Ground Truth 2 sequence adjusts to [1.5, 8.0, 11.5, 11.5, 10.0, 3.0, 7.0, 6.0, 4.0, 9.0, 1.5, 5.0], resulting in a revised relevance coefficient of 0.124 and a calculated weight of 3.232. Similarly, with the removal of Proposal no. 3 under Ground Truth 3, the new sequence becomes [1.0, 9.0, 11.0, 7.0, 6.0, 4.0, 10.0, 8.0, 2.0, 12.0, 3.0, 5.0], leading to a revised correlation coefficient of 0.070 and a derived weight of 1.708. These outcomes confirm the accuracy of the weighting algorithm, as the derived weights align with the anticipated changes in correlation coefficients following the removal of influential proposals. This analysis demonstrates the effectiveness of the algorithm in quantifying the impact of individual proposals on the overall evaluation consistency of reviewers.



(**a**) (**b**)

Figure 5. Comparison of weights calculated by Reviewer no. 11. (**a**) Weights calculated by Reviewer no. 11 based on Ground Truth 2. (**b**) Weights calculated by Reviewer no. 11 based on Ground Truth 3.

The weight analysis graphs for Reviewer no. 11, based on both Ground Truth 2 and Ground Truth 3, effectively illustrate the volatility of the weights associated with this reviewer′s evaluations. These visual representations reveal significant fluctuations in the weights assigned to different proposals, indicating that Reviewer no. 11′s evaluations deviate considerably from the collective consensus. Such variability suggests that his assessments are influenced by factors distinct from those guiding the majority of reviewers, emphasizing his role as a significant outlier in the peer review process.

Figure 6 presents the weight analysis plot for Reviewer no. 11 based on Ground Truth 1, enabling direct comparison with those derived from Ground Truths 2 and 3. The analysis reveals that the plots for Ground Truths 2 and 3 are markedly similar, whereas the plot based on Ground Truth 1 diverges significantly from the other two. This variance can be attributed to the distinct methodologies underlying each ground truth, which inherently produce different outcomes. Ground Truth 1 emphasizes the score itself, incorporating a method that excludes the highest and lowest scores to minimize error. However, given the variability in scoring habits and standards among reviewers, the rankings derived from Ground Truth 1 tend to reflect an amalgamation of individual reviewer biases; it represents an average that integrates all personal factors of the reviewers. In contrast, Ground Truth 2 focuses more on the ordinal ranking of proposals. While large discrepancies in scores may exist, these do not necessarily translate into substantial differences in rankings, potentially obscuring true variance among scores. Ground Truth 3 builds on the framework of Ground Truth 2 by introducing a recommendation voting mechanism, which not only emphasizes the impact of scores but also aims to establish a more uniform standard among reviewers, mitigating the influence of personal biases. Thus, Ground Truth 3 is essentially an enhancement of Ground Truth 2, making the similarity between the results from these two reasonable. The fundamental difference in the approach of Ground Truth 1 explains the disparity in outcomes when compared to Ground Truths 2 and 3. This analysis underscores the importance of considering the specific methodologies of ground truths to understand their impact on the evaluation of reviewers in peer review systems.

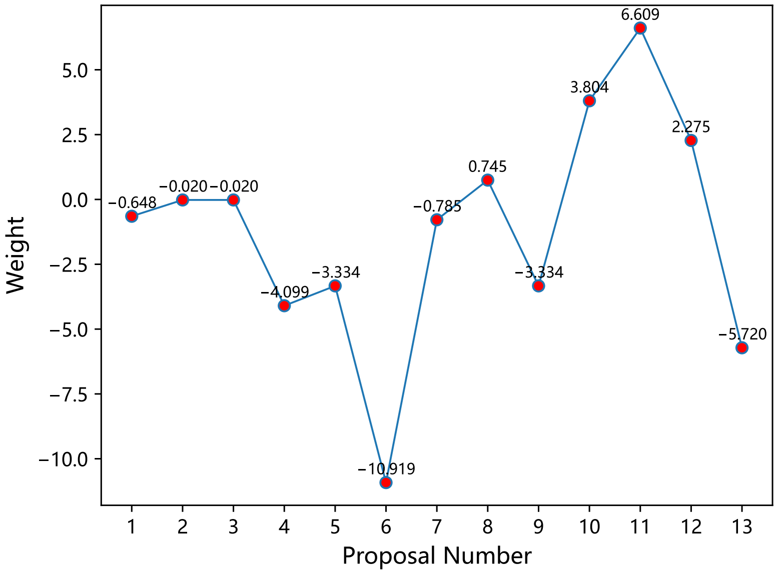


Figure 6. Weights calculated by Reviewer no. 11 based on Ground Truth 1.

Proposal no. 11 consistently receives high weights across all three ground truths, indicating that Reviewer no. 11′s perspective on this proposal significantly diverges from others. To enhance fairness in peer review, it would be beneficial to investigate the reasons behind this divergence and consider follow-up measures, such as targeted training.

Despite the distinct philosophical underpinnings and approaches to quantifying collective consensus inherent in the three ground truths, a consistent observation emerges regarding Proposal nos. 7 and 8. Across all ground truths, these proposals consistently receive weights close to zero, indicating a minimal impact on the outlier index of Reviewer no. 11. This uniformity suggests that regardless of the methodological differences, these proposals do not significantly alter the reviewer's deviation from the collective consensus. This finding highlights the potential stability of certain evaluations across varied evaluative frameworks and underscores the value of analyzing multiple ground truths to gain a comprehensive understanding of the factors influencing reviewer assessments.

5. Conclusions

In conclusion, the DCASP method has been validated through empirical testing on two distinct datasets, demonstrating its efficacy in enhancing the integrity and fairness of peer review systems. The proposed data-crossing analysis method effectively identifies and mitigates the influence of outlier scores, contributing to a more rigorous and equitable evaluation process. By revealing deviations from the estimated ground truth, the methodology successfully detects abnormal scoring behaviors and irrational patterns, thereby ensuring a higher degree of accuracy and fairness in assessments. These findings highlight the robustness and reliability of the DCASP approach in addressing the inherent subjectivity and biases of individual reviewers, ultimately leading to the improvement of academic evaluation processes. This advancement holds promise for fostering a more trustworthy and balanced peer review environment, benefiting both reviewers and authors within the academic community.

Author Contributions: Conceptualization, Z.D. and J.Z.; methodology, Z.D.; software, Z.D. and W.T.; validation, Z.D., J.Z. and X.S.; formal analysis, Z.D.; writing—original draft preparation, Z.D.; writing—review and editing, J.Z. and X.S.; visualization, Z.D.; supervision, J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Due to the fact that our data is derived from a real peer review process and involves sensitive reviewer information, it cannot be shared publicly in order to protect privacy.

Conflicts of Interest: J.Z. is employed by the company China Science and Technology Exchange Center, and remaining authors declare that no conflicts of interest.

References

1. National Natural Science Foundation of China. *Annual Report of the National Natural Science Foundation of China 2022*; Zhejiang University Press: Hangzhou, China, 2023.
2. Aly, M.; Colunga, E.; Crockett, M.; Goldrick, M.; Gomez, P.; Kung, F.Y.; McKee, P.C.; Pérez, M.; Stilwell, S.M.; Diekman, A.B. Changing the culture of peer review for a more inclusive and equitable psychological science. *J. Exp. Psychol. Gen.* **2023**, *152*, 3546–3565. [CrossRef] [PubMed]
3. Kelly, J.; Sadeghieh, T.; Adeli, K. Peer review in scientific publications: Benefits, critiques, & a survival guide. *eJIFCC* **2014**, *25*, 227. [PubMed]
4. Gregory, A.T.; Denniss, A.R. Everything you need to know about peer review—The good, the bad and the ugly. *Heart Lung Circ.* **2019**, *28*, 1148–1153. [CrossRef] [PubMed]
5. Tennant, J.P.; Ross-Hellauer, T. The limitations to our understanding of peer review. *Res. Integr. Peer Rev.* **2020**, *5*, 6. [CrossRef]
6. Shoham, N.; Pitman, A. Open versus blind peer review: Is anonymity better than transparency? *BJPsych Adv.* **2021**, *27*, 247–254. [CrossRef]
7. Kaatz, A.; Gutierrez, B.; Carnes, M. Threats to objectivity in peer review: The case of gender. *Trends Pharmacol. Sci.* **2014**, *35*, 371–373. [CrossRef]
8. Iezzoni, L.I. Explicit disability bias in peer review. *Med. Care* **2018**, *56*, 277–278. [CrossRef]
9. Smith, O.M.; Davis, K.L.; Pizza, R.B.; Waterman, R.; Dobson, K.C.; Foster, B.; Jarvey, J.C.; Jones, L.N.; Leuenberger, W.; Nourn, N.; et al. Peer review perpetuates barriers for historically excluded groups. *Nat. Ecol. Evol.* **2023**, *7*, 512–523. [CrossRef]
10. Linton, J.D. Improving the Peer review process: Capturing more information and enabling high-risk/high-return research. *Res. Policy* **2016**, *45*, 1936–1938. [CrossRef]
11. Gai, T.; Cao, M.; Chiclana, F.; Zhang, Z.; Dong, Y.; Herrera-Viedma, E.; Wu, J. Consensus-trust driven bidirectional feedback mechanism for improving consensus in social network large-group decision making. *Group Decis. Negot.* **2023**, *32*, 45–74. [CrossRef]

12. Cui, J.; Chen, Z.; Zhou, A.; Wang, J.; Zhang, W. Fine-grained interaction modeling with multi-relational transformer for knowledge tracing. *ACM Trans. Inf. Syst.* **2023**, *41*, 1–26. [CrossRef]
13. Xu, S.; Xu, J.; Yu, S.; Li, B. Identifying Disinformation from Online Social Media via Dynamic Modeling across Propagation Stages. In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, Boise, ID, USA, 21–25 October 2024; pp. 2712–2721.
14. Xu, S.; Fu, X.; Pi, D.; Ma, Z. Inferring Individual Human Mobility From Sparse Check-in Data: A Temporal-Context-Aware Approach. *IEEE Trans. Comput. Soc. Syst.* **2024**, *11*, 600–611. [CrossRef]
15. Squazzoni, F.; Bravo, G.; Takács, K. Does incentive provision increase the quality of peer review? An experimental study. *Res. Policy* **2013**, *42*, 287–294. [CrossRef]
16. Ji, I.H.; Lee, J.H.; Kang, M.J.; Park, W.J.; Jeon, S.H.; Seo, J.T. Artificial intelligence-based anomaly detection technology over encrypted traffic: A systematic literature review. *Sensors* **2024**, *24*, 898. [CrossRef] [PubMed]
17. Meng, X.; Ma, J.; Liu, F.; Chen, Z.; Zhang, T. An Interpretable Breast Ultrasound Image Classification Algorithm Based on Convolutional Neural Network and Transformer. *Mathematics* **2024**, *12*, 2354. [CrossRef]
18. Leys, C.; Ley, C.; Klein, O.; Bernard, P.; Licata, L. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *J. Exp. Soc. Psychol.* **2013**, *49*, 764–766. [CrossRef]
19. Gervini, D. Outlier detection and trimmed estimation for general functional data. *Stat. Sin.* **2012**, *22*, 1639–1660. [CrossRef]
20. Visser, M.; Van Eck, N.J.; Waltman, L. Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quant. Sci. Stud.* **2021**, *2*, 20–41. [CrossRef]
21. Etsebeth, V.; Lochner, M.; Walmsley, M.; Grespan, M. Astronomaly at scale: Searching for anomalies amongst 4 million galaxies. *Mon. Not. R. Astron. Soc.* **2024**, *529*, 732–747. [CrossRef]
22. Zhang, W.; Lai, X.; Wang, J. Social link inference via multiview matching network from spatiotemporal trajectories. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *34*, 1720–1731. [CrossRef]
23. Spearman, C. The proof and measurement of association between two things. *Am. J. Psychol.* **1987**, *100*, 441–471. [CrossRef] [PubMed]
24. Zardi, H.; Karamti, H.; Karamti, W.; Alghamdi, N.S. Detecting anomalies in network communities based on structural and attribute deviation. *Appl. Sci.* **2022**, *12*, 11791. [CrossRef]
25. Stelmakh, I.; Rastogi, C.; Liu, R.; Chawla, S.; Echenique, F.; Shah, N.B. Cite-seeing and reviewing: A study on citation bias in peer review. *PLoS ONE* **2023**, *18*, e0283980. [CrossRef] [PubMed]
26. Shayegan, M.J.; Sabor, H.R.; Uddin, M.; Chen, C.L. A collective anomaly detection technique to detect crypto wallet frauds on bitcoin network. *Symmetry* **2022**, *14*, 328. [CrossRef]
27. Fernández, E.; Rangel-Valdez, N.; Cruz-Reyes, L.; Gomez-Santillan, C. A new approach to group multi-objective optimization under imperfect information and its application to project portfolio optimization. *Appl. Sci.* **2021**, *11*, 4575. [CrossRef]
28. Papadopoulos, P.M.; Lagkas, T.D.; Demetriadis, S.N. Technology-enhanced peer review: Benefits and implications of providing multiple reviews. *J. Educ. Technol. Soc.* **2017**, *20*, 69–81.
29. Hosseini, M.; Horbach, S.P. Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. *Res. Integr. Peer Rev.* **2023**, *8*, 4. [CrossRef]
30. Ji, Y.; Ma, Y. The robust maximum expert consensus model with risk aversion. *Inf. Fusion* **2023**, *99*, 101866. [CrossRef]
31. Hsu, F.C.; Elvidge, C.D.; Baugh, K.; Zhizhin, M.; Ghosh, T.; Kroodsma, D.; Susanto, A.; Budy, W.; Riyanto, M.; Nurzeha, R.; et al. Cross-matching VIIRS boat detections with vessel monitoring system tracks in Indonesia. *Remote Sens.* **2019**, *11*, 995. [CrossRef]
32. Vinutha, H.; Poornima, B.; Sagar, B. Detection of outliers using interquartile range technique from intrusion dataset. In *Information and Decision Sciences, Proceedings of the 6th International Conference on Ficta*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 511–518.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.