Analysis of the ICML 2023 Ranking Data: Can Authors' Opinions of Their Own Papers Assist Peer Review in Machine Learning?

Buxin Su* Jiayao Zhang* Natalie Collina* Yuling Yan† Didong Li‡
Kyunghyun Cho§ Jianqing Fan¶ Aaron Roth* Weijie J. Su*

August 23, 2024

Abstract

We conducted an experiment during the review process of the 2023 International Conference on Machine Learning (ICML) that requested authors with multiple submissions to rank their own papers based on perceived quality. We received 1,342 rankings, each from a distinct author, pertaining to 2,592 submissions. In this paper, we present an empirical analysis of how author-provided rankings could be leveraged to improve peer review processes at machine learning conferences. We focus on the Isotonic Mechanism, which calibrates raw review scores using author-provided rankings. Our analysis demonstrates that the ranking-calibrated scores outperform raw scores in estimating the ground truth "expected review scores" in both squared and absolute error metrics. Moreover, we propose several cautious, low-risk approaches to using the Isotonic Mechanism and author-provided rankings in peer review processes, including assisting senior area chairs' oversight of area chairs' recommendations, supporting the selection of paper awards, and guiding the recruitment of emergency reviewers. We conclude the paper by addressing the study's limitations and proposing future research directions.

1 Introduction

The peer review process plays a key role in advancing research by identifying high-quality, high-impact research. Nevertheless, a widespread decline in the quality of peer review has been observed across various disciplines (Brezis and Birukou, 2020; Cheah and Piasecki, 2022), with notable concerns in machine learning and artificial intelligence centered around the noisiness or arbitrariness of the process (Langford and Guzdial, 2015; Yuan et al., 2022). For instance, a randomized experiment at NeurIPS 2021 indicated that about half of the accepted papers would be rejected upon a second round of reviews (Cortes and Lawrence, 2021; Beygelzimer et al., 2023). One factor contributing to this decline is the stunning increase in the number of submissions to machine learning conferences. For example, NeurIPS 2023—one of three annual top machine learning conferences—received 12,343 submissions; it seems nearly impossible to recruit enough experienced and capable reviewers

*University of Pennsylvania. Email: suw@wharton.upenn.edu.
†Massachusetts Institute of Technology.
‡University of North Carolina at Chapel Hill.
§New York University.
¶Princeton University.

to provide low variance assessments for this volume of submissions on a regular basis (Sculley et al., 2019; Stelmakh et al., 2021).

This reality has spurred a continued research effort aimed at improving the peer review process for machine learning conferences, often focusing on reviewer assignments and reviewer biases (Kobren et al., 2019; Wang and Shah, 2019; Jecmen et al., 2020; Leyton-Brown et al., 2022; Stelmakh et al., 2023). In contrast to existing research on the peer review process, which has focused primarily on the reviewers, the recently proposed Isotonic Mechanism seeks to enhance peer review by leveraging authors' opinions of the quality of their *own* submissions to yield more robust review scores (Su, 2021). This mechanism requires authors with multiple submissions to rank their papers according to their perception of the relative quality of the papers and generates calibrated scores that are modified from the original review scores to align with the author-provided rankings. This ranking-based calibration can be seen as a "de-noising" of the original review scores from the perspective of the authors (Su, 2021, 2022). Accurate review scores are crucial to improving peer review in very large-scale publication venues because they are the most important factor in determining accept/reject decisions.[1]

The Isotonic Mechanism is well-suited for contemporary machine learning conferences, where it is commonplace for an author to submit multiple papers at the same time (Sun, 2020; Rastogi et al., 2022). An advantage of this methodology is that it requires minimal effort from authors and imposes no additional burden on reviewers. Rather than exacerbating reviewer workload, this approach uses the authors themselves as a resource for adding information to the peer review process.

In this paper, we aim to evaluate the empirical effectiveness of the Isotonic Mechanism for peer review. To this end, we conducted a survey experiment at the 2023 International Conference on Machine Learning (ICML), which is one of the top-tier conferences in machine learning and artificial intelligence. In 2023, ICML received 6,538 submissions from 18,535 authors. On January 26, 2023, right after the ICML submission deadline, we sent a survey to all submitting authors who have OpenReview profiles to ask them to provide rankings of their submissions if they submitted at least two papers.

Specifically, we address the following questions by analyzing the ICML 2023 ranking data together with review scores and accept/reject decisions:[2]

(a) *How do the outputs of the Isotonic Mechanism, which we refer to as isotonic scores, compare to original/raw review scores in terms of accurately reflecting submission quality?*

(b) *What near-term applications might there be for leveraging isotonic scores to enhance the peer review process?*

(c) *What are the limitations of the study, and which aspects of the mechanism should future experiments investigate?*

Addressing the first question requires understanding the relationship between review scores and paper quality, which is challenging due to the absence of ground truth quality of the ICML 2023 submissions. To address this, we leverage the fact that submissions typically receive multiple review scores. We use the average of the remaining scores as a proxy for the ground truth "expected

[1]For example, the NeurIPS 2023 guidelines for area chairs state that any decision "should be properly explained" if any paper with an average score above (below) the threshold is rejected (accepted).

[2]Our code is publicly available on GitHub.

review score" of a submission when evaluating the performance of an estimator applied to a single randomly selected review score. In Section 3.1, we show that although the squared error of the review scores computed with respect to this proxy is an upwards-biased estimator for the unobserved "ground truth" squared error with respect to the true expected review score, because the bias terms exactly cancel, it can still be used to obtain an unbiased estimator for the *difference* in squared error between two estimators, with respect to the ground truth. Figure 1 shows that the Isotonic Mechanism substantially reduces proxy estimation errors—specifically, mean squared error (MSE) and mean absolute error (MAE)—for the 2,592 submissions ranked by authors. Moreover, it suggests that the improvement becomes more substantial as the number of submissions of an author increases. This dependence implies that more author-provided rankings could lead to even more significant error reduction through the Isotonic Mechanism.[3] Because the change in proxy estimation error is an unbiased estimator for the change in ground truth estimation error, this allows us to produce confidence intervals for the decrease in ground truth squared error, and we find substantial decreases at the 99% confidence level. See more details in Section 3.

In response to the second question, Section 4 proposes three cautious applications of isotonic scores and author-provided rankings during peer review processes. First, we suggest that senior area chairs (SACs) can use isotonic scores to help direct additional scrutiny of accept/reject recommendations made by area chairs (ACs). Here the scores are used to focus additional human scrutiny, not directly used for accept or reject decisions. Second, we suggest using author-provided rankings to help direct attention during the selection of paper awards. This application occurs after the accept/reject decisions are made, ensuring its impact is circumscribed, and again used to direct human attention rather than to make decisions directly. Third, we propose using the discrepancy between isotonic scores and raw review scores as an indicator of review quality. A significant discrepancy, particularly in comparison to other submissions, could signal the need for an emergency reviewer to provide additional evaluation. We provide empirical evidence from the ICML 2023 data supporting the effectiveness of the latter two applications. These three applications share the crucial feature that the sensitive information of rankings and isotonic scores is visible only to certain high-level roles, such as SACs and program chairs (PCs), and their use does not directly impact the accept/reject decisions.

In Section 5, we conclude our paper by discussing the third question, the limitations of our experimental results, and suggesting avenues for future research. While this method has shown empirical effectiveness at increasing the accuracy of noisy reviews, and is supported by theoretical underpinnings—including "truthfulness" guarantees—in a stylized setting (Su, 2022), using the isotonic mechanism to make important decisions may give rise to unforeseen consequences because of the possibility of strategic manipulation by authors in ways that are not captured by the stylized analysis. Thus, caution is warranted which is why our policy recommendations are limited, and we recommend a more comprehensive investigation of the Isotonic Mechanism in future experiments.

1.1 Related Work

Continued efforts have been made to evaluate and quantify the randomness and quality of peer review from the standpoint of authors. It has been demonstrated that there is a consensus among researchers about the declining quality of peer review (Lipton and Steinhardt, 2019; Wang et al.,

[3]If all ICML 2023 authors were to provide their rankings, this would increase the average length of rankings. See Figure 11 in the Supplementary Material.

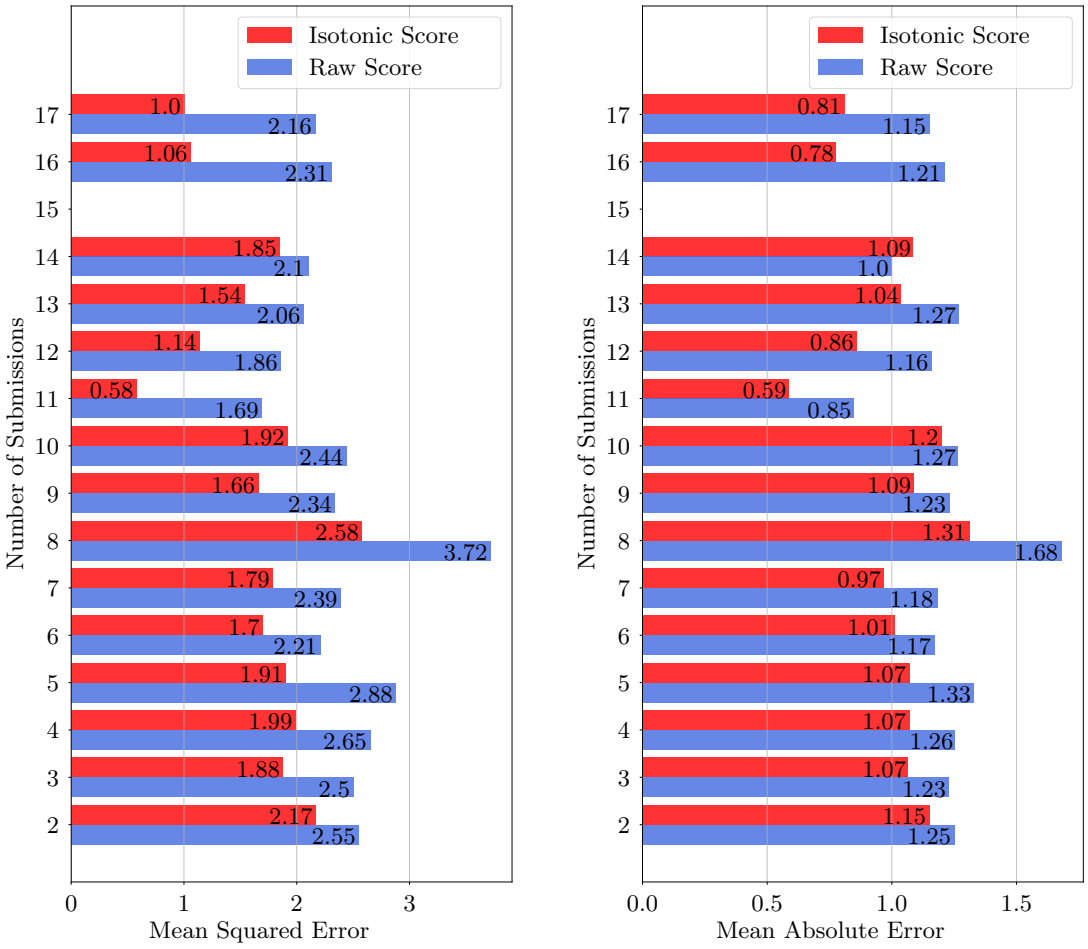

Figure 1: (Proxy) MSE and MAE averaged over ICML 2023 authors who submitted rankings of the same length in our survey experiment, ranging from 2 to 17 submissions. The original/raw review scores are in red, while the isotonic scores are in blue. Overall, the relative improvement using the isotonic scores grows with the number of submissions by an author. For MSE, the average percentage improvement for numbers of submissions between 2 and 10 is 25% and it is 41% for submissions more than 11. See more details in Figure 10. The experimental setup is described in Section 3.2.

2020; Russo, 2021; Liang et al., 2023a). By analyzing 1,313 reviews, Pranić et al. (2021) found that authors were most satisfied with reviews recommending acceptance, but reviews suggesting revisions were of the highest quality. Two studies particularly relevant to our paper are those conducted by Gardner et al. (2012) and Rastogi et al. (2022). Gardner et al. (2012) surveyed authors to rate their submissions to the Australasian Association for Engineering Education Annual Conference, and Rastogi et al. (2022) asked NeurIPS 2021 authors to estimate the acceptance probabilities of their submissions and to compare them pairwise. Both studies revealed that authors tend to overestimate their submissions' chances of acceptance.

From a methodological standpoint, various approaches have been introduced to refine peer review, primarily from the perspectives of reviewers (Van Rooyen et al., 1999; Arous et al., 2021; Jecmen et al., 2020; Liang et al., 2023b). For instance, Ugarov (2023) proposed a mechanism that incentivizes reviewers via peer prediction. More recently, Liang et al. (2023b) explored the potential of employing large language models such as GPT-4 to generate initial reviews for research papers. For completeness, an emerging body of work considers enhancing peer review from the authors' perspectives (Aziz et al., 2019; Mattei et al., 2020; Srinivasan and Morgenstern, 2021). For example, Srinivasan and Morgenstern (2021) proposed a mechanism that requires authors to submit a bid for a review slot for each submitted paper.

2 Experimental Design and Summary Statistics

We first provide an overview of some basic statistics from ICML 2023.

(a) Number of submissions: 6,538.

(b) Number of authors: 18,535.[4]

(c) Number of submissions with at least one author having more than one submission: 5,035 (77.0%).

(d) Number of authors with two or more submissions: 4,505 (24.3%).

(e) Number of authors with at least 5 submissions: 508.

(f) Number of authors with at least 10 submissions: 74.

(g) Number of authors with at least 15 submissions: 26.

(h) Number of authors with at least 20 submissions: 7.

An ICML 2023 submission was typically reviewed by three or four reviewers. Reviewers rated submissions on a scale from 1 (very strong reject) to 10 (award quality). For the 6,538 submissions, the decisions were as follows:

(a) Number of "Withdrawn or Desk Rejected" submissions: 991 (15.2%).[5]

(b) Number of "Rejected" submissions: 3,719 (56.9%).

[4] Among the 18,535 authors, 20 did not have OpenReview profiles. Our survey was sent via the OpenReview API to those 18,515 authors who had OpenReview profiles.

[5] This category includes three papers withdrawn after having been accepted.

(c) Number of "Accepted as Poster" submissions: 1,674 (25.6%).

(d) Number of "Accepted as Oral"[6] submissions: 154 (2.36%).

(e) Number of submissions awarded the "Outstanding Paper Award": 6 (0.09%).

The mean of the average review scores[7] following the rebuttal period for the categories "Rejected", "Accepted as Poster", "Accepted as Oral", and "Outstanding Paper Award" is 4.32, 5.93, 6.82, and 7.72, respectively.

The survey-based experiment was conducted in OpenReview, which hosted the peer review process for ICML 2023, in conjunction with OpenRank.cc, which we developed to implement the experiment. On January 26, 2023, immediately following the submission deadline, an official email was sent through OpenReview to all ICML authors requesting information about their submissions. Importantly, participants were informed that the survey data would not be used in the decision-making process for ICML 2023. Figure 8 in the Supplementary Material shows the survey interface. Specifically, we solicited the following information:

- Ranking: Authors with multiple submissions were asked to rank their papers based on their perceived quality, with allowances for ties in the rankings. Authors could order their papers by dragging them up or down at the OpenRank.cc interface.
- Additional questions: All authors, including those with only one submission, were asked to respond to some questions, such as their confidence in the provided rankings and their perceived probability of inconsistencies between their expectations and the review outcomes. All these questions are shown in Figure 8 in the Supplementary Material.

Additionally, review scores and final decisions were retrieved from OpenReview.

This study was conducted with the approval of the Institutional Review Boards (IRB) at the University of Pennsylvania. The experiment and subsequent analyses adhered to strict privacy and confidentiality standards. Specifically, the data were anonymized by excluding all personal identifying information, and analysis began only after the accept/reject decisions were announced. For further information regarding our privacy policy, please refer to `https://openrank.cc/legal/privacy`. Furthermore, it is our policy that all data collected from this experiment be completely deleted by December 31, 2024.

Summary statistics from the experiment. We first provide statistics about the rankings obtained from the survey:

(a) Number of authors who completed the survey: 5,634 (30.4%).

(b) Number of authors who had multiple submissions and provided rankings of their submissions: 1,342.

(c) Number of submissions that were ranked by at least one author: 2,592 (39.6%).

[6] Officially termed "Accepted as Oral & Poster".

[7] In this paper, the term "average review score" of a paper refers to the simple average of all ratings provided by reviewers for that paper.

(d) Number of reviews received by these 2,592 submissions: 7,974 (3.08 reviews on average per submission).[8]

(e) The longest ranking list provided by an author: 17 submissions.

The dependence between the number of submissions by an author and their likelihood of completing the survey is shown in Figure 11 in the Supplementary Material. It appears that authors with more submissions were less likely to provide rankings.

Regarding the additional questions in the survey, 59.8% of the authors reported high confidence in their rankings, and 59.4% would likely provide the same rankings if they were to be used in the decision-making process. More details are given in Figure 12 in the Supplementary Material.

In response to the question, "What is your estimated probability that your lowest-ranked paper will have a higher or equal average rating than your highest-ranked paper?", over half of the authors estimated this probability to be at least 40%. The average of these estimated probabilities is 36.6%. The distribution of these probabilities is illustrated in Figure 13 in the Supplementary Material. In contrast, the actual proportion of authors whose lowest-ranked papers received higher or equal scores prior to the rebuttal period, compared to their highest-ranked papers, is 42.2%.

Preliminary analysis of rankings. Examining the relevance of author-provided rankings in predicting the quality of submissions is a key focus of our study. If these rankings were not predictive of review outcomes at all, incorporating them into the Isotonic Mechanism would be unlikely to enhance the efficacy of the review process. Yet, our preliminary analysis suggests that these rankings are indeed predictive, indicating their potential value to improve peer review.

To investigate this aspect, we grouped the highest-ranked paper by an author into one category and the lowest-ranked paper into another. The mean of average review scores received by the highest-ranked papers is 4.80 before the rebuttal period, while it is 4.50 for the lowest-ranked papers. This difference is statistically significant, with a p-value of 6.17×10^{-16} using a one-sided t-test. A more detailed comparison between the two groups depending on the categories of decision is given in Table 1, which shows that highest-ranked submissions were more likely to obtain better review outcomes.

Moreover, this positive correlation extends beyond the rebuttal period, where highest-ranked papers were more likely to receive an increase in score, with an average increase of 0.23, compared to 0.20 for the lowest-ranked. However, our analysis found no statistically significant correlation between the rankings and the length of the rebuttals, as measured by word count. This is illustrated in Table 6 in the Supplementary Material.

3 Statistical Analysis of Isotonic Scores

In this section, we analyze the ICML 2023 ranking data, which comprises 1,342 rankings associated with 2,592 submissions. Our main finding is that the Isotonic Mechanism can reduce the MSE of review scores, and this improvement in estimation accuracy is not only statistically significant but also becomes more pronounced as the number of an author's submissions increases. This empirical finding aligns with the theoretical analysis of the method (Su, 2021, 2022).

[8] This is the number of reviews received before the rebuttal, as of March 12, 2023. The average number of reviews per submission increased to 3.29 after the rebuttal period, as of April 22, 2023.

	"Withdrawn or Desk Rejected"	"Rejected"	"Accepted as Poster"	"Accepted as Oral"
Highest-ranked	9.02%	53.20%	33.31%	4.47%
Lowest-ranked	16.24%	57.68%	23.85%	2.24%
p-value	2.43×10^{-7}	2.20×10^{-3}	7.31×10^{-8}	1.87×10^{-3}

Table 1: Comparison of outcomes ("Withdrawn/Desk Rejected", "Rejected", "Accepted as Poster", and "Accepted as Oral") for highest-ranked versus lowest-ranked submissions. Highest-ranked submissions received significantly better review outcomes. We use a one-sided t-test to obtain p-values.

3.1 Method and Evaluation

The Isotonic Mechanism operates as follows (Su, 2021). Consider an author who submits n papers to a conference. The mechanism requires the author to rank these submissions in descending order of perceived quality. The ranking is denoted by π, which allows for ties. Given the (average) raw review scores $\boldsymbol{y} := (y_1, y_2, \ldots, y_n)$ for the n submissions, the Isotonic Mechanism outputs the ranking-calibrated scores as the solution to the following optimization program:

$$\min_{\boldsymbol{r}:=(r_1,\ldots,r_n)} \quad \|\boldsymbol{y} - \boldsymbol{r}\|^2 \qquad \text{s.t.} \quad r_{\pi(1)} \geq \cdots \geq r_{\pi(n)},$$

where $\|\cdot\|$ denotes ℓ_2 norm or, equivalently, Euclidean distance. Formally, this convex optimization program is equivalent to isotonic regression (Barlow and Brunk, 1972). For example, letting $\boldsymbol{y} = (8, 7, 4, 3)$ and $(\pi(1), \pi(2), \pi(3), \pi(4)) = (1, 3, 2, 4)$,[9] the isotonic scores are $(8, 5.5, 5.5, 3)$.

An essential aspect of this method lies in the assumption that the author is knowledgeable about the quality of their submissions. The Isotonic Mechanism is perhaps the simplest blend of authors' and reviewers' perspectives. Furthermore, under certain conditions, authors are incentivized to truthfully report their rankings when these modified scores are used for decision-making (Su, 2021). With truthful author-provided rankings, the isotonic scores are more accurate than the raw scores in estimating the ground truth quality of submissions. Extensions of this mechanism for broader practical settings are discussed in Yan et al. (2023) and Wu et al. (2023).

To adapt the Isotonic Mechanism to the practical setting where most papers each have multiple authors, we consider three strategies.

Simple-averaging strategy. The first strategy runs the mechanism for each author who provides a ranking. For a given submission, different authors often yield different modified scores. The isotonic score for the submission is calculated as the simple average of these different modified scores.

Greedy strategy. This strategy starts by running the Isotonic Mechanism for the author who provides the longest ranking, and then removes this author and all submissions of the author from further consideration. This process is repeated for the remaining authors and their submissions until each submission has exactly one isotonic score.[10]

[9] From this ranking, the author has the opinion that the last paper has the lowest quality and the first paper has the highest quality.

[10] If the length of the ranking is 1, the isotonic score will be identical to the raw review score.

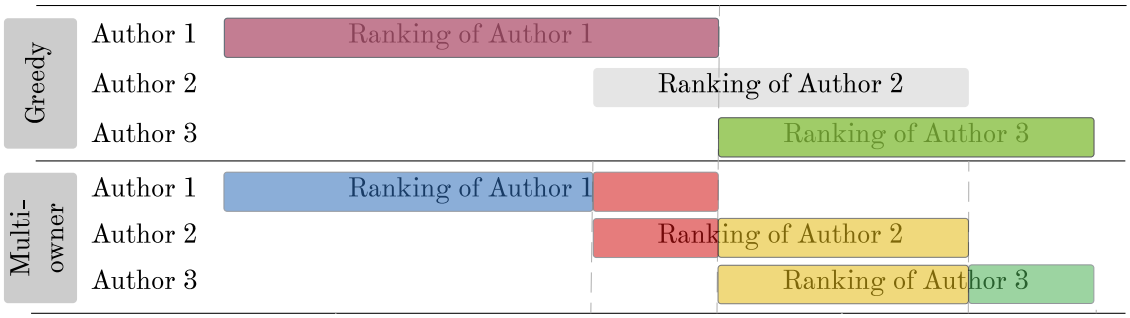

Figure 2: Illustration of the greedy and multi-owner strategies for the Isotonic Mechanism in the setting of multiple authors. Author-submission pairs highlighted in gray (Author 2's submissions in the greedy strategy) are excluded from consideration in the mechanism. In the multi-owner strategy, any paper in the red block has its score averaged over the two isotonic scores from Author 1 and Author 2.

Multi-owner strategy. The first step of this strategy is to partition all submissions into disjoint blocks such that in each block every submission shares a common set of authors. In each block, run the Isotonic Mechanism taking as input a ranking within the block from each author to yield modified scores. The last step is to average the modified scores separately for each block. See more details about this approach in Wu et al. (2023).

In both the greedy and multi-owner strategies, each run of the Isotonic Mechanism operates on a sub-ranking that involves a subset of the submissions from an author. However, it is important to note that these sub-rankings can be derived from complete rankings. Therefore, authors can provide complete rankings, regardless of the strategy being implemented. Notably, under certain conditions, the greedy and multi-owner strategies ensure the truthful reporting of rankings by authors (Wu et al., 2023). In contrast, the simple-averaging approach does not generally guarantee this.

Evaluation metrics. Evaluating the performance of isotonic scores compared to raw scores presents a challenge due to the unknown ground truth quality of submissions. To address this challenge, we leverage the fact that a submission typically receives multiple reviews, resulting in multiple review scores. For simplicity, consider y and y' as two independent scores of the same submission, both assumed to be unbiased estimators of the ground truth.[11] Let $\widehat{y}$ denote any estimator of the ground truth using only the data y. The performance of $\widehat{y}$ is measured using either $(\widehat{y} - y')^2$ or $|\widehat{y} - y'|$, which we refer to as the "proxy" MSE and MAE, respectively. Note that the conventional MSE and MAE of $\widehat{y}$ are defined as $\mathbb{E}(\widehat{y} - \text{ground truth})^2$ and $\mathbb{E}|\widehat{y} - \text{ground truth}|$, respectively, which are not observable. In contrast, $(\widehat{y} - y')^2$ and $|\widehat{y} - y'|$ can be precisely calculated from the raw scores.

Both proxy MSE and MAE are upward-biased estimators of their conventional counterparts.

[11] Scores may exhibit biases conditional on certain factors (Wang et al., 2020). In our context, "unbiasedness" is understood in a marginal sense, achieved through random selection of scores without conditioning on variables such as confidence level or reviewer seniority. Moreover, it is more appropriate to interpret "ground truth" as the "ground truth score" rather than the intrinsic merit of a paper. Practically, the ground truth score could be considered as the average score given by a very large number (say, 1,000) of reviewers.

This can be seen, as the expectation of the proxy MSE is expressed as

$$\mathbb{E}(\widehat{y} - y')^2 = \mathbb{E}\left(\widehat{y} - \mathbb{E}y'\right)^2 + \mathrm{Var}(y') = \mathrm{MSE}(\widehat{y}) + \mathrm{Var}(y').$$

In essence, the bias of the proxy MSE is equal to the variance of the "noisy target" y'. For the proxy MAE, note that it satisfies $\mathbb{E}|\widehat{y} - y'| \geq \mathbb{E}|\widehat{y} - \mathbb{E}y'| = \mathrm{MAE}(\widehat{y})$. Here, the inequality follows from Jensen's inequality when applied to the convex function $|c - x|$ for any constant $\widehat{y} = c$.

Despite this bias, the proxy MSE retains the ability to compare two estimators in expectation: for any two estimators $\widehat{y}$ and $\widetilde{y}$, their proxy MSE's difference satisfies

$$\begin{aligned} \mathbb{E}(\widehat{y} - y')^2 - \mathbb{E}(\widetilde{y} - y')^2 &= \mathrm{MSE}(\widehat{y}) + \mathrm{Var}(y') - \mathrm{MSE}(\widetilde{y}) - \mathrm{Var}(y') \\ &= \mathrm{MSE}(\widehat{y}) - \mathrm{MSE}(\widetilde{y}). \end{aligned} \tag{3.1}$$

Therefore, if $\widehat{y}$ outperforms $\widetilde{y}$ in terms of MSE, then $\widehat{y}$ will also have a smaller proxy MSE than $\widetilde{y}$ in expectation, and vice versa.

When analyzing the ICML 2023 ranking data, we randomly selected one review score per submission as the data y for estimating the submission's quality using either the Isotonic Mechanism or the raw-score-estimator. The average of the remaining review scores serves as the noisy target y'. For this purpose, a submission must have at least two review scores. This is applicable to 2,530 out of the 2,592 ranked submissions.

3.2 Results

To compare the isotonic and raw scores, Figure 3 presents scatter plots for each of the proxy MSE and MAE across the 2,530 submissions. A least-squares fit without an intercept between the proxy errors using isotonic scores and raw scores indicates that, on average, the proxy MSE of isotonic scores is smaller than that of raw scores, a difference that is statistically significant at the 10^{-8} level. Similarly, the proxy MAE of isotonic scores is found to be statistically smaller than that of raw scores, with the gap being narrower yet still significant. This is consistent across all three strategies of the Isotonic Mechanism, suggesting an overall improvement in estimating submission quality. Furthermore, Figures 4 and 5 corroborate the findings in Figure 3.

Table 2 demonstrates that the Isotonic Mechanism using any of the three strategies reduces both the overall proxy MSE and MAE compared to raw scores. Specifically, the greedy strategy achieves a 21.3% reduction in the proxy MSE and 11.7% in the proxy MAE. Furthermore, evidence suggests that the reduction in conventional MSE is likely to exceed 21.3% when employing isotonic scores, as elaborated in Section C in the Supplementary Material.

	Proxy MSE			Proxy MAE		
	Error	Improvement	p-value	Error	Improvement	p-value
Raw Score	2.57	NA	NA	1.26	NA	NA
Simple-averaging Strategy	1.97	23.48%	9.99×10^{-11}	1.10	12.75%	6.69×10^{-10}
Greedy Strategy	2.02	21.30%	5.53×10^{-9}	1.11	11.71%	1.66×10^{-8}
Multi-owner Strategy	2.07	19.38%	1.22×10^{-7}	1.12	10.62%	3.10×10^{-7}

Table 2: Reduction of proxy MSE and MAE using the Isotonic Mechanism with various strategies. A two-sample t-test shows that the reduction in proxy errors is statistically highly significant.

Denote by $\widehat{y}^{\text{Iso}}$ and $\widehat{y}^{\text{Ave}}$ the isotonic score and raw score, respectively. As shown in (3.1), $(\widehat{y}^{\text{Ave}} - y')^2 - (\widehat{y}^{\text{Iso}} - y')^2$ is an unbiased estimator of $\text{MSE}(\widehat{y}^{\text{Ave}}) - \text{MSE}(\widehat{y}^{\text{Iso}})$. This observation allows us to construct confidence intervals for the average reduction in ground truth MSE—that is, $\text{MSE}(\widehat{y}^{\text{Ave}}) - \text{MSE}(\widehat{y}^{\text{Iso}})$ averaged over all ranked submissions—and we present the results in Table 3. At the 95% confidence level, $\text{MSE}(\widehat{y}_i^{\text{Iso}})$ on average is smaller than $\text{MSE}(\widehat{y}_i^{\text{Ave}})$ by 0.4 or more.

Confidence Level	Simple-averaging Strategy	Greedy Strategy	Multi-owner Strategy
95%	$[0.52, 0.69]$	$[0.46, 0.63]$	$[0.42, 0.58]$
99%	$[0.49, 0.71]$	$[0.44, 0.66]$	$[0.39, 0.60]$

Table 3: 95% and 99% confidence intervals for the average reduction of MSE by using isotonic scores compared to raw scores. This makes use of the fact that the reduction in proxy MSE is an unbiased estimator of ground truth MSE reduction, as shown in (3.1).

In investigating the impact of an author's number of submissions on the Isotonic Mechanism's performance, Figure 6 averages the proxy MSE and MAE for isotonic and raw scores across authors with the *same* number of submissions. The results indicate a tangible and statistically significant improvement in estimation accuracy using the Isotonic Mechanism, irrespective of the number of submissions. Overall, this improvement becomes more pronounced with an increase in submission quantity, as shown in Figure 10 in the Supplementary Material.

These findings align with the theoretical analysis in Su (2022), which proves that the Isotonic Mechanism achieves greater performance with larger numbers of submissions. However, shorter rankings are typically more common in our dataset, as authors with more submissions in ICML 2023 tended to have lower response rates. Taken together, our results imply that the advantages of the Isotonic Mechanism would be more pronounced if all authors in ICML 2023 provided their complete rankings.

4 Applications

Section 3 provides strong evidence that the isotonic scores are more accurate estimates of ground truth compared to the raw review scores, as measured by both MSE and MAE. Despite these results, however, we do not advocate for using them to make a major change in how paper acceptance decisions are made at major machine learning conferences at the moment. Instead, we suggest a more modest and cautious approach for using isotonic scores in the review process, while at the same time conducting additional empirical evaluations of the Isotonic Mechanism to attempt to better understand the consequences of using it in the review process.

Towards this end we have identified several specific applications of the Isotonic Mechanism and author-provided rankings that appear to be beneficial without significant negative consequences. These applications primarily target scenarios where authors are aware of important nuances regarding the scientific value of their papers, which might be overlooked by reviewers or ACs. These applications share the common feature that the isotonic scores or author-provided rankings are accessible only to certain high-level roles within the peer review hierarchy, such as SACs and above, and thus give a separation between the isotonic scores and the majority of accept/reject decisions.

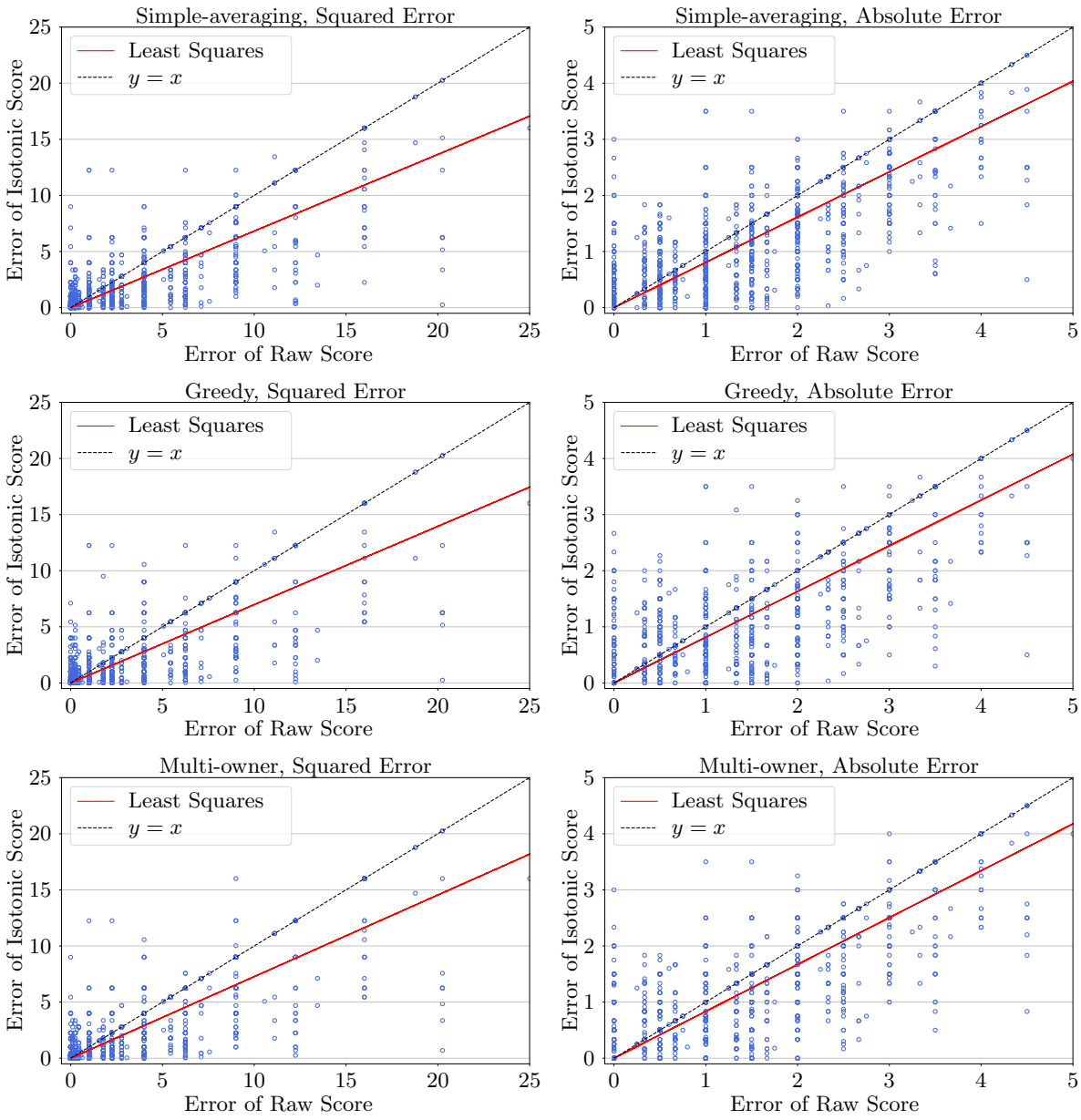

Figure 3: Scatter plots comparing isotonic and raw scores in terms of proxy MSE (left) and proxy MAE (right). Each plot represents a submission, with the x and y axes indicating the proxy errors of the raw score and isotonic score of the submission, respectively. The least-squares fit line is consistently below the 45° line passing through the origin, demonstrating that the proxy error of isotonic scores is, on average, smaller than that of the raw scores. The analyses in this section are based on review scores collected prior to the rebuttal phase, as of March 12, 2023. For analyses based on post-rebuttal review scores, see Section B in the Supplementary Material.

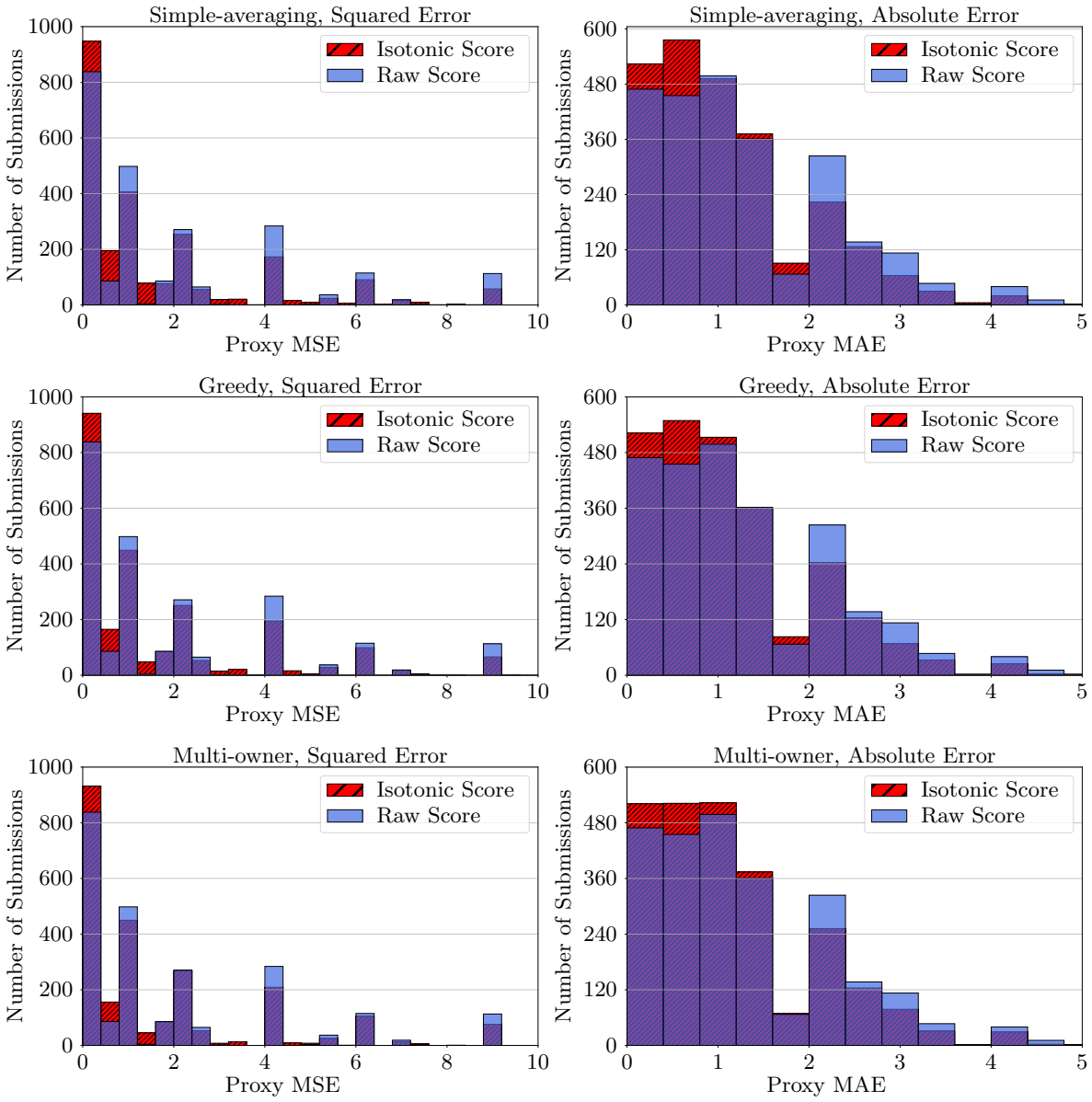

Figure 4: Histograms comparing the distributions of isotonic and raw scores in terms of proxy MSE (left) and proxy MAE (right). The distribution of isotonic scores is more heavily weighted towards smaller errors.

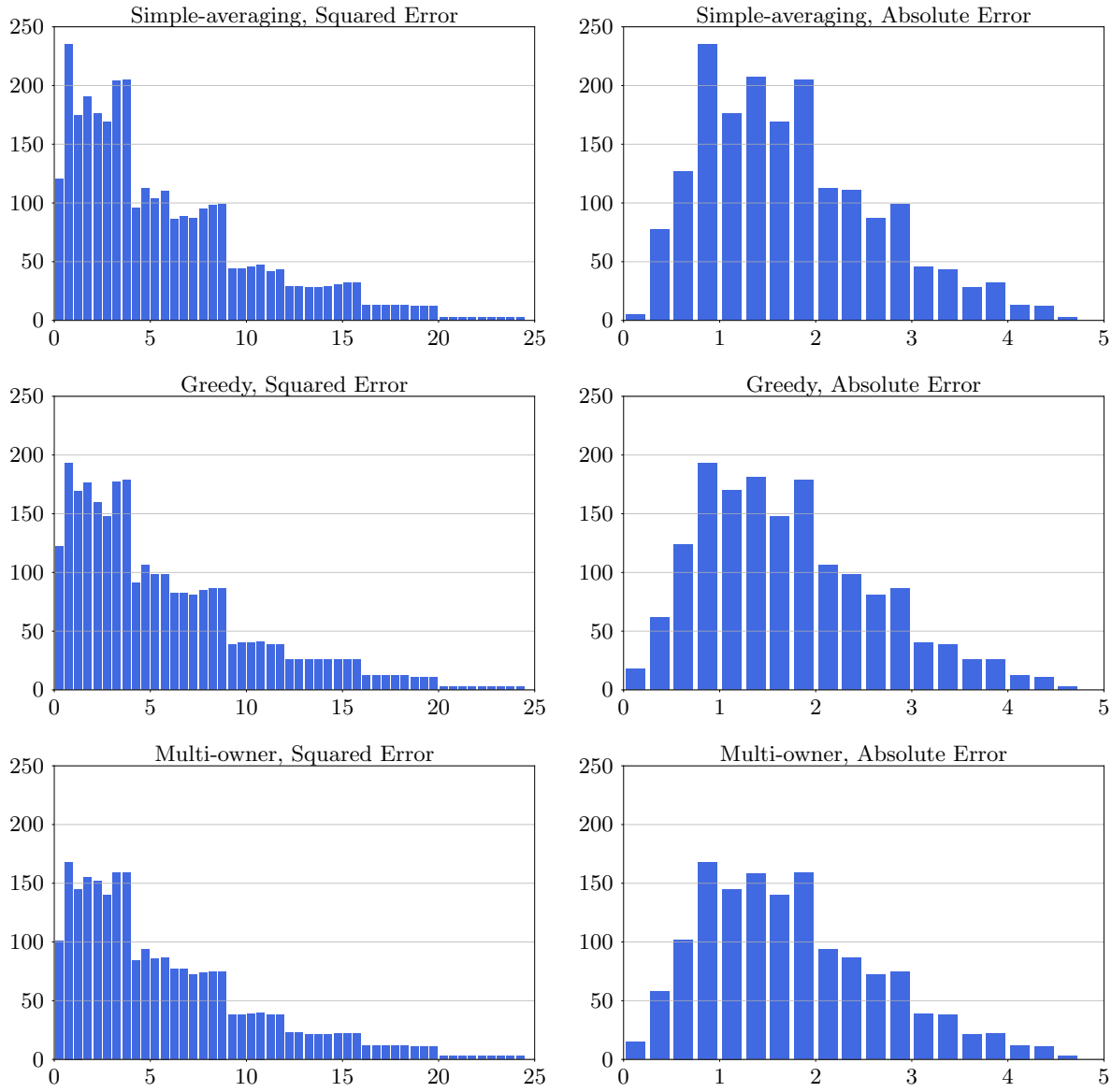

Figure 5: Difference between cumulative distributions of proxy errors for isotonic and raw scores. Left panel: At x-axis value x, the y-axis value represents $|\{i : (\widehat{y}_i^{\text{Iso}} - y'_i)^2 \leq x\}| - |\{i : (\widehat{y}_i^{\text{Ave}} - y'_i)^2 \leq x\}|$, where $|\{i : (\widehat{y}_i^{\text{Iso}} - y'_i)^2 \leq x\}|$ denotes the number of submissions with isotonic scores having proxy MSE less than or equal to x. Right panel: At x-axis value x, the y-axis value represents $|\{i : |\widehat{y}_i^{\text{Iso}} - y'_i| \leq x\}| - |\{i : |\widehat{y}_i^{\text{Ave}} - y'_i| \leq x\}|$. The consistently positive difference demonstrates that isotonic scores generally yield smaller proxy errors than raw scores in distribution.

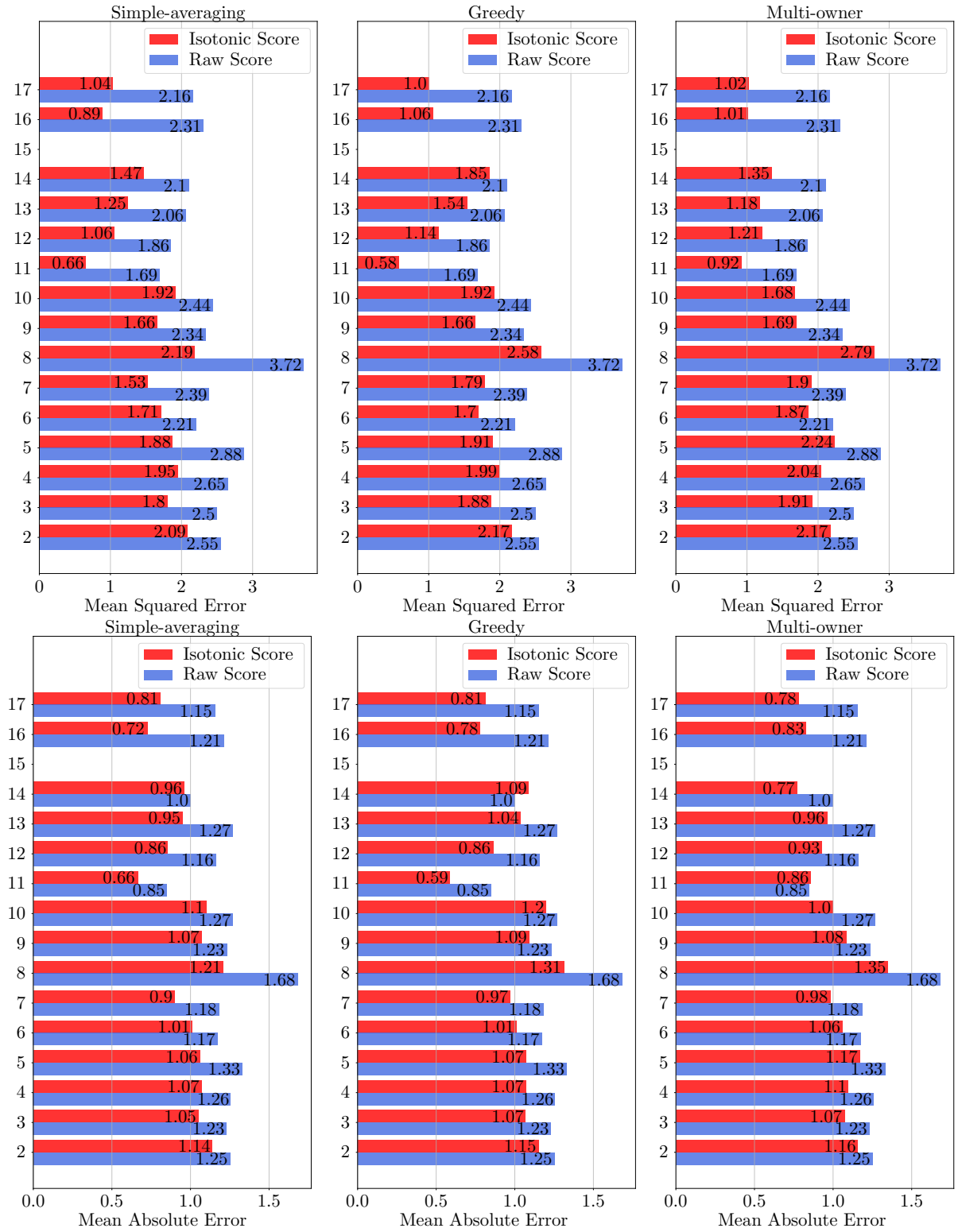

Figure 6: Comparison between isotonic and raw scores in terms of proxy MSE and MAE averaged over ICML 2023 authors who submitted the same-length rankings. Figure 1 corresponds to the middle column.

4.1 Oversight of ACs' Recommendations

The isotonic scores can be used to flag submissions in need of more scrutiny by SACs. In this application, the isotonic scores are made visible to SACs and those in higher roles, who can then use these scores to more effectively oversee the recommendations made by ACs. For instance, significant discrepancies between isotonic scores and ACs' recommendations could serve as red flags, prompting SACs to scrutinize and discuss these cases with the ACs.

The use of isotonic scores in this application presents low risk because ACs, who make the initial accept/reject recommendations (the majority of which are also the final decisions), do not have access to these scores. To further mitigate risk, when an SAC identifies a red flag for a submission's review, the SAC could request that the AC conduct a further review of the submission without specifying that the request is due to a large discrepancy between the accept/reject recommendation and the authors' own opinions.

4.2 Selection of Paper Awards

Machine Learning conferences select certain papers to receive awards. The process typically begins with the formation of a shortlist, including papers with high average scores or those nominated by ACs. A committee then carefully reviews these shortlisted papers to identify the award recipients. The committee is intended to carefully weigh each paper on its merits, and not simply choose based on reviewer scores. However, the process does not always go smoothly; for insight into the difficulties and controversies that some recent award decisions have generated, we refer the reader to Carlini et al. (2022) and Orabona (2023), which critically examined the ICML Outstanding Paper Awards in 2022 and 2023.

Author-provided rankings could be given as an additional useful piece of information for the committee involved in the selection of paper awards. As evidence that this information might be useful, three out of the six papers awarded as Outstanding Papers at ICML 2023 were ranked by one of their authors and, notably, were all ranked first by their authors. Furthermore, of the 84 submissions that received oral presentations (a distinction that is given to the top few percent of papers) and had rankings from their authors, 69.1% were ranked first by at least one of their authors. These statistics highlight a strong correlation between the authors' rankings and the recognition the papers received.

In the selection of papers for awards, the rankings could be made visible to some PCs who are not on the selection committee.[12] The committee relies on their expertise in the selection of the paper awards without knowledge of the author-provided rankings. Once the recommendation is made by the selection committee, the PCs could then scrutinize and raise flags if a recommended paper receives low rankings from its authors, in which case the committee may need to gather additional evidence before considering it for an award.

The award selection takes place following the accept/reject decisions. This phase does not impact most authors, thereby minimizing the potential for unforeseen outcomes when using author-provided rankings.

[12] Here, the rankings instead of the isotonic scores are used for a reason that will be elaborated in Section 4.3.

4.3 Recruitment of Emergency Reviewers

In machine learning conferences, it is common practice to recruit emergency reviewers in response to indicators of low review quality, often triggered by low-confidence reviews or significant disagreement among reviewers for a submission. For instance, NeurIPS 2023 recommended recruiting an additional emergency reviewer for each low-confidence review in addition to the four regular reviewers. An effective mechanism for assigning emergency reviewers is an economical way of utilizing the limited pool of qualified reviewers (Peng, 2018; Stelmakh et al., 2021) by adaptively assigning them to papers based on the quality of the initial round of reviews.

Determining review quality is an inherently noisy process. Incorporating authors' elicited rankings into this determination when assigning emergency reviewers could both improve its accuracy and enhance the community's trust in the credibility of the peer review processes. It is crucial to convey to authors that their concerns are taken seriously, especially when they disagree with the reviews. This can be achieved by leveraging isotonic scores, based on the premise that discrepancies between raw review scores and isotonic scores, which we refer to as isotonic residuals,[13] might signal concerns about review quality from the authors' viewpoint.

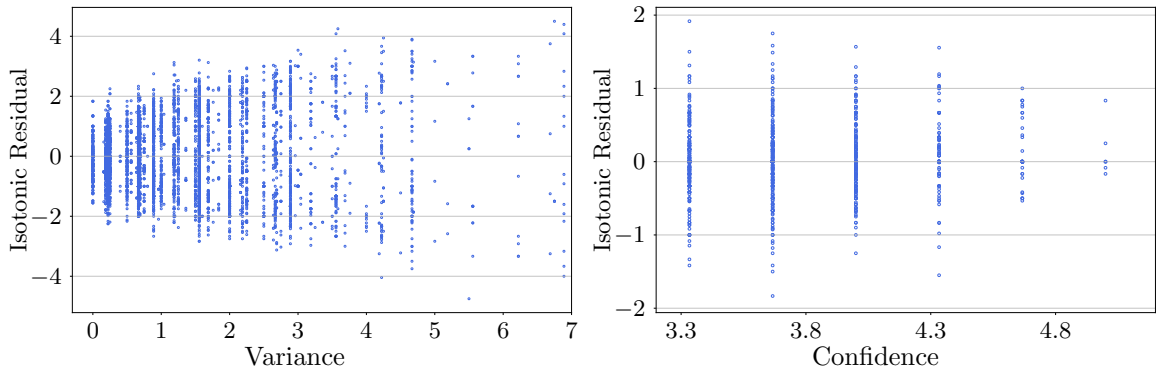

Figure 7: Scatter plots showing isotonic residuals using the simple-averaging strategy, plotted against review confidence levels and against raw score variance, for submissions with a confidence level of 3 or above.

To provide empirical support, we examine the relationship between isotonic residuals and both the variance of review scores and review confidence levels, using data from ICML 2023. High score variance or low confidence levels are often used as indicators for recruiting emergency reviewers (Shah et al., 2018). Figure 7 illustrates that isotonic residuals in absolute value have a strong negative correlation with confidence levels and a positive correlation with score variance. Furthermore, little dependence is found between score variance and confidence levels, with a correlation of only 2.05×10^{-2}. This suggests that isotonic residuals might offer a more comprehensive measure of review quality.

We also evaluate the effectiveness of isotonic residuals using data from a second survey, which asked authors to identify submissions where the review outcomes differed most from their expectations. Table 4 shows that isotonic residuals are the most predictive of submissions with the most "unexpected" review outcomes. This finding is expected as isotonic scores, by definition, reflect

[13] In contrast to the setup in Section 3.2, here we run the Isotonic Mechanism on the average of *all* review scores for each submission.

authors' expectations.

	Isotonic Residual	Score Variance	Score Confidence
Prediction Accuracy	254/322 = 78.9%	162/322 = 50.3%	136/322 = 42.2%

Table 4: Prediction accuracy of the most "unexpected" review outcomes, determined using the largest mean isotonic residual in absolute value, the greatest variance of review scores, and the lowest average confidence levels

Given this empirical evidence, we propose using large isotonic residuals as an indicator of the need for emergency reviewers to provide additional expert opinions. In implementing this mechanism, it is crucial to ensure that submissions receive roughly the same number of reviewers on average, regardless of whether a submission is included in this mechanism or not.[14] One approach to achieving this balance is to assign three initial reviewers to papers participating in the mechanism, while four reviewers to those not included. For papers in the participating group, we assign two emergency reviewers for isotonic residual magnitudes in the top 30%, and one for magnitudes between the 30th and 70th percentiles.[15] Consequently, on average, a paper has four reviewers, regardless of its group. To ensure a cautious approach, the quantile of the isotonic residual in absolute value will be made available to ACs and above, but not the raw isotonic score itself. Without knowing whether the isotonic residual is positive or negative, ACs cannot ascertain whether authors hold a high or low opinion of their submissions, thereby minimizing the likelihood of bias influencing the ACs' decisions.

It is important to note that the isotonic score differs from the raw review score only when at least one author has multiple submissions to the conference. Consequently, the Isotonic Mechanism cannot uniformly increase review accuracy across all submissions, which is also the rationale for suggesting the use of rankings rather than isotonic scores in the selection of paper awards, as discussed in Section 4.2. As demonstrated, isotonic scores are better estimates of the ground truth "expected review scores" than the raw review scores, and the accuracy improves with the number of papers an author has submitted and ranked (Su, 2021). In particular, authors with fewer submissions are more likely to have higher variance in their isotonic scores. Thus, caution is warranted when comparing isotonic residuals of submissions across authors. Nevertheless, despite the fact that the Isotonic Mechanism does not improve accuracy uniformly for all submissions, as it is accuracy improving, it seems prudent to (cautiously) use the information to improve the review process—using the isotonic residuals to recruit emergency reviewers is one such example of cautious use.

5 Limitations and Future Work

This paper has presented analyses of the ICML 2023 ranking data collected from 1,342 submitting authors to empirically evaluate the Isotonic Mechanism. Our findings indicate that this mechanism

[14] A submission might not be included because no author has multiple submissions. However, this applies to a minority of submissions in large machine learning conferences. For example, 77.0% of the ICML 2023 submissions have at least one author with more than one submission.

[15] The numbers 30% and 70% are arbitrary as long as they ensure that the expected number of emergency reviewers for a paper is one.

can effectively mitigate noise in review scores. Additionally, we introduce three cautious applications of the Isotonic Mechanism and author-provided rankings to improve peer review processes.

In interpreting these results, it is crucial to note that the rankings were provided under the condition that they would not influence decision-making processes. Authors might behave strategically if their rankings were used in decision-making. We note here that 59.4% of authors in our survey stated that they would submit the same rankings even if they were to be used for decision making. Although the Isotonic Mechanism has been shown to be dominant strategy truthful in stylized scenarios (Su, 2022), we cannot rule out the possibility that authors could benefit from strategic behaviour in a real world deployment. Examples of strategic behaviors not covered in the game theoretic analysis might include authors showing a preference for papers where they are the first author over those where they are secondary, or a professor ranking a student's paper higher than warranted to boost the student's job market prospects, or an author would like to get a weak paper published first and defer strong ones for the next conference series. The possibility of strategic behavior is perhaps the most concerning obstacle for deployment of the Isotonic Mechanism for consequential decision making, and should be the focus of future investigation. Another concern is that the variance reduction effect of the Isotonic Mechanism increases with the number of papers an author submits (and does nothing at all for authors who submit only a single paper). This complicates comparisons of scores across authors with different numbers of submissions, and is another aspect that is in need of further study.

The analysis of the results could be potentially improved by properly addressing the potential for non-response bias, which exists due to the 30.4% response rate in our experiment. For example, the mean of average review scores of all submissions to ICML 2023 is 4.53 before the rebuttal period, while it is 4.66 for the 2,592 ranked submissions. The response rate is largely influenced by the number of submissions per author, with more prolific authors being less inclined to provide rankings. However, it is these authors for whom the Isotonic Mechanism could be most effective. Consequently, this bias might lead to a conservative estimate of the mechanism's effectiveness. In future experiments, incentivizing authors to participate could increase participation. One possible incentive mechanism would be to provide an earlier notification of decisions to authors who provided rankings.

Percentile	5%	10%	15%	20%	25%	30%
Overlapping Fraction	53.17%	64.03%	67.81%	66.40%	73.42%	81.55%

Table 5: Fractions of overlapping top-scored submissions between isotonic scores (using greedy strategy) and raw scores, for different percentiles among the 2,530 ranked submissions.

Another potentially useful approach to improving the quality of reviews is to aggregate rankings provided by reviewers. Imagine a conference where every reviewer is asked to rank the submissions they have reviewed. Consider each submission as a node and draw an edge between two nodes if they share a common reviewer. This creates a submission-reviewer network, which divides the submissions into several connected graphs. For each connected graph, the Plackett–Luce model can be employed to estimate the preference score of each paper using the spectral method described in Fan et al. (2024b). This approach yields ranking-based scores, which refine raw review scores through the aggregation of preferences from all reviewers. These scores can also be integrated with those generated by the Isotonic Mechanism. Notably, preference scores can be consistently estimated even when reviewers only identify their top choices or provide partial rankings (Fan et al.,

2024a).

To conclude, we discuss several avenues for future research. From a methodological viewpoint, the Isotonic Mechanism in its current form does not account for reviewers' confidence levels; developing weighted review scores that feed into the mechanism could potentially enhance its performance. Additionally, investigating how the mechanism could leverage rankings provided by reviewers (Fan et al., 2024b) presents another valuable research direction. The practical challenge of coauthors holding differing opinions on their submissions, as evidenced by the 28.7% of coauthor pairs providing inconsistent comparisons (dropping to 25.9% for submissions with substantial review score differences), calls for developing a variant of the mechanism that adapts to these ranking inconsistencies (Wu et al., 2023). From a practical perspective, whether a paper would be accepted depends largely on whether its score is above the top, say, 20% of all submissions. Table 5 shows how much the top-scored papers would change from using raw scores to using isotonic scores. Notably, the overlapping fraction increases as the cutoff percentage increases. An interesting question for future study is to compare these two different lists of top-scored papers in terms of the citations they have accumulated over the years.

Acknowledgments

We would like to thank Melisa Bok, Emma Brunskill, Barbara Engelhardt, Xiao-Li Meng, Nihar Shah, Sherry Xue, and James Zou for helpful discussions in early stages of this project. We are grateful to the ICML Board for providing the opportunity for this survey experiment. This research was supported in part by NSF grant CCF-1934876 and Wharton AI for Business.

References

I. Arous, J. Yang, M. Khayati, and P. Cudré-Mauroux. Peer grading the peer reviews: A dual-role approach for lightening the scholarly paper review process. In *Proceedings of the Web Conference 2021*, pages 1916–1927, 2021.

H. Aziz, O. Lev, N. Mattei, J. S. Rosenschein, and T. Walsh. Strategyproof peer selection using randomization, partitioning, and apportionment. *Artificial Intelligence*, 275:295–309, 2019.

R. E. Barlow and H. D. Brunk. The isotonic regression problem and its dual. *Journal of the American Statistical Association*, 67(337):140–147, 1972.

A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan. Has the machine learning review process become more arbitrary as the field has grown? The NeurIPS 2021 consistency experiment. *arXiv preprint arXiv:2306.03262*, 2023.

E. S. Brezis and A. Birukou. Arbitrariness in the peer review process. *Scientometrics*, 123(1):393–411, 2020.

N. Carlini, V. Feldman, and M. Nasr. No free lunch in "privacy for free: How does dataset condensation help privacy". *arXiv preprint arXiv:2209.14987*, 2022.

P. Y. Cheah and J. Piasecki. Should peer reviewers be paid to review academic papers? *The Lancet*, 399 (10335):1601, 2022.

C. Cortes and N. D. Lawrence. Inconsistency in conference peer review: Revisiting the 2014 NeurIPS experiment. *arXiv preprint arXiv:2109.09774*, 2021.

J. Fan, Z. Lou, W. Wang, and M. Yu. Ranking inferences based on the top choice of multiway comparisons. *Journal of American Statistical Association*, page to appear, 2024a.