Rebutting Rebuttals

Nachum Dershowitz
School of Computer Science
Tel Aviv University
nachumd@tau.ac.il

Rakesh M. Verma
Department of Computer Science
University of Houston
rmverma2@central.uh.edu

Draft of September 29, 2021

Introduction

Many prominent computer science conferences, including AAAI, ACL, ASPLOS, CCS, CODASPY, EACL, EMNLP, EUROCRYPT, ICML, IJCAI, LICS, MOBICOM, MobiHoc, NDSS, NeurIPS, OSDI, PLDI, POPL, SIGCOMM, SOSP, STOC, and USENIX, have in recent years instituted rebuttal/feedback periods during which authors get to see reviews and are offered the opportunity to answer specific queries or to otherwise respond to issues raised in the referee reviews.

To understand the extent to which such rebuttals affect the ultimate decisions, we analyzed the recently-released review data for the 2018 *Annual Meeting of the Association for Computational Linguistics (ACL 2018)*. Considering the significant effort that goes into composing and writing rebuttals by authors and into reading and reconsidering by referees, and the concomitant extra weeks of delay in decisions, this is an important question.

It turns out that it is far from clear that the net advantages justify the significant effort involved of authors in writing rebuttals and of reviewers in considering them for score changes.

There have been numerous studies of the quality and possible biases in conference reviews. In a controlled study of reviewing of the *Conference on Neural Information Processing Systems* (*NIPS*; now *NeurIPS*), the outcome for all but the extreme cases was more or less random.[1] Many of the problems, including the disadvantage of novelty, are exacerbated by the ever-growing scale of major conferences.[2] Collusion among reviewers is another growing problem.[3] Some other relevant works are cited in Wang et al. [4]

We address the issue of rebuttal impact only. Members of the *Conference on Human Factors in Computing Systems (CHI)* 2016 and 2020 committees reported that although rebuttals led to score changes, they had minimal impact on final outcomes, concluding, “Perhaps there is a conversation to be had in the community about whether those 1.7m words are worth the effort.”[5,6] Our analysis is a followup on the work of Gao et al.,[7] which analyzed the reviews themselves for impact of style and other matters, concluding inter alia that impoliteness hurts.

The Dataset

The dataset consists of review scores for all submissions to *ACL 2018*.[8] For each paper, besides an ID number, there are the following data items:

1. submission type (long or short);
2. final status (withdrawn before rebuttal, withdrawn after, rejected without review, rejected, accepted for oral paper, accepted for poster, accepted with shepherding);
3. had rebuttal or not;
4. for each reviewer, there are the following scores:
   (a) initial overall score (range 1..6);
   (b) initial reviewer confidence level (1..5);
   (c) final (post-rebuttal) overall score;

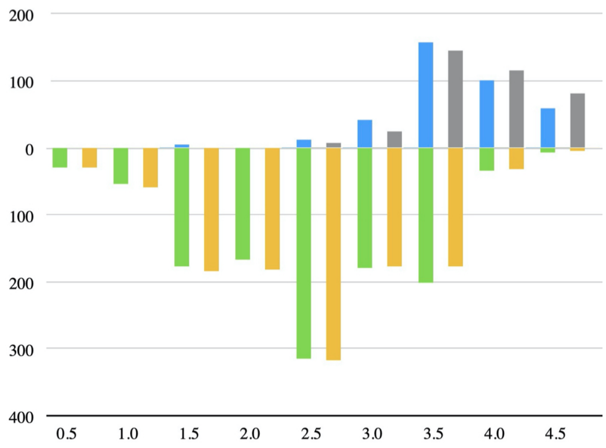


Figure 1: Acceptance by mean pre- and post-rebuttal overall score. The green/blue bars show the number of rejections (below "sea-level") and acceptances (above) for each initial, pre-rebuttal score; orange/grey for the final score. The orange/grey bars reflect post-rebuttal rejections/acceptances.

(d) final confidence level.

For late reviews, there are only final scores.

The figures are as follows:

- 1545 papers submitted;
- 3875 reviews initially, averaging 2.5 reviews per submission; 4059 all told, including late arrivals (2.6 reviews/submission);
- 1197 rebuttals (77% of the papers);
- 39 confidence level changes, more up (23) than down (16);
- 493 score changes (13% of the reviews): 245 positive (50%) and 248 negative (50%);
- only 72 papers had more than one change: 26 were all downwards; 17 were upwards; 28 were evenly balanced; 1 was imbalanced;
- 480 papers had a change in mean overall score on account of revised scores and/or new reviews: 237 positive (49%) and 243 negative (51%);
- 393 changes in overall score when averaged over original reviewers only 198 positive (50%) and 195 negative (50%);
- 381 acceptances (as poster or paper), which amount to a 25% acceptance rate;

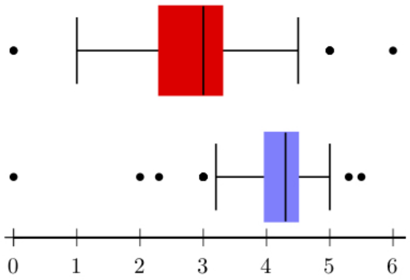


Figure 2: Box-and-whisker plots for acceptance (blue) and rejection (red), showing medians (bars in middle of colored rectangles), first and third quartiles (rectangle edges), upper and lower fences (whiskers extend to include all data points within 1.5 times the interquartile range), and scattered outliers (dots). Among the outliers were 3 papers with no reviews at all (score 0), one of which was accepted as a poster post rebuttal.

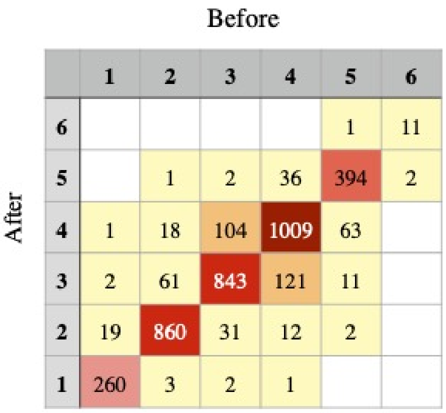
- 352 (92%) accepted submissions supplied a rebuttal; 845 (73%) of rejected did.

We concentrate on the degree of impact of rebuttals on the reviewers' overall scores. Figure 1 shows acceptance rates for the different final mean overall scores. The box plot in Figure 2 shows the distribution of scores for accepted and rejected submissions.

In analyzing the dataset, Gao et al. [7] found, unsurprisingly, that peer pressure to homogenize scores is a major factor in score change. These changes presumably had no impact on final decisions but are merely cosmetic. They go on to assert:

> The 227 papers that receive at least one INC [increase after rebuttal] review, their acceptance rate is 49.8%, much higher than those 221 papers with at least one DEC [decrease] (7.2%) and those 1119 papers with no score update (22.8%). Hence, the score update has a large impact on the final accept/reject decision.

This conclusion is unwarranted. Score changes are often in line with the intended decision, as indicated in their analysis. Increases are likely indicative of a more positive average, so a higher acceptance rate is



After \ Before	1	2	3	4	5	6
6					1	11
5		1	2	36	394	2
4	1	18	104	1009	63	
3	2	61	843	121	11	
2	19	860	31	12	2	
1	260	3	2	1		

Table 1: The number of reviews for each level of overall scores, before and after rebuttal. (Shading is darker the more there are.) The overwhelming majority of reviewers leave their scores intact during the second round. Higher scores are more likely to move down; lower scores, to move up. (From Gao et al.[7])

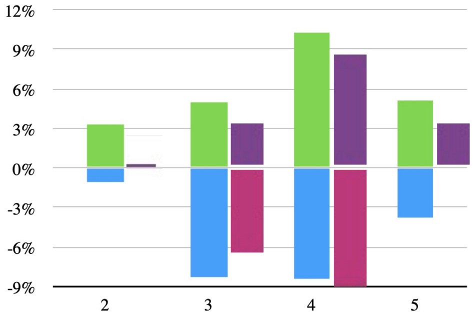


Figure 3: Percentages of positive and negative score changes for papers with (green/blue) and without (purple/magenta) rebuttal. Columns are for initial average score ranges (1–2, 2–3, etc.).

to be expected. To be sure, we concur that updates are "largely determined by the scores of peer reviewers."

Rebuttal Effectivity

Table 1 shows changes in individual review scores after the rebuttal period. Table 2 lists post-rebuttal changes in overall score averaged over all reviewers. Virtually all score changes are minor as can be seen from the sparseness of the matrices other than on or near the antidiagonals.

Significantly, the average mean overall score for all papers was the same before and after rebuttals, with or without late reviews (3.15–3.16). Scores were as likely to go down after rebuttal as up. This all suggests that rebuttals—intended to leave a positive impression on reviewers—are not the main impetus for changes, but rather the consideration of other referees' evaluations and opinions is.

Figure 3 shows the distribution of score changes up and down for both when there was or wasn't a rebuttal. Most reviewers, of course, don't modify their scores regardless.

Examining the tables as well as the raw data, we make the following additional observations:

- There were 52 dramatic changes of ±2 or more to individual scores, close to half (24) positive and slightly over half negative (28). Presumably many decreases are because a serious flaw was identified.
- There were hardly any big changes in mean scores: 5 increased more than 1 point and 6 decreased that much. Most are ascribable to additional reviews.
- There were only 4 contributions whose original referee(s) changed their average score more than 1 point: 2 increased and were accepted; 2 decreased and were rejected.
- Another 8 accepted papers with rebuttals increased 1 point, likely due to the positive influence of the author response; some would have been accepted regardless.
- Ignoring late reviews, mean scores *for 348 papers with rebuttals* changed. For 180, post-rebuttal scores were higher; for 168 they were lower. Virtually all were under 1 point.
- Of the 180 with increased scores, 98 were accepted. Of the 168 with lowered scores, almost all (158) were ultimately rejected.
- Of the 158 rejections, 31 had had an even chance of acceptance before their scores decreased.

Table 3 highlights the 80 (in blue and green) accepted papers with increased post-rebuttal scores. Of

	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0
6.0	0	0	0	0	0	0	0	0	0	1
5.5	0	0	0	0	0	0	4	8	25	1
5.0	0	0	0	0	1	3	20	82	8	0
4.5	0	0	1	0	11	51	224	8	0	0
4.0	0	0	2	1	38	118	28	4	1	0
3.5	0	0	6	11	246	42	16	0	1	0
3.0	0	1	11	136	43	7	0	1	0	0
2.5	1	4	221	26	9	0	0	0	0	0
2.0	0	56	3	6	1	0	0	0	0	0
1.5	53	2	1	0	0	0	0	0	0	0

Table 2: Mean overall scores before (columns) and after (rows) rebuttal. Heading x refers to range $[x - 1/2, x)$. The green area counts papers whose likelihood of acceptance changed from relatively low to relatively high; orange are those that moved in the opposite direction.

	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0
6.0	0	0	0	0	0	0	0	0	0	1
5.5	0	0	0	0	0	0	5	8	24	0
5.0	0	0	0	0	1	1	16	68	3	0
4.5	0	0	1	0	8	29	133	1	0	0
4.0	0	0	0	0	9	24	1	0	0	0
3.5	0	0	2	0	12	1	0	0	0	0
3.0	0	0	0	0	0	0	0	0	0	0
2.5	0	0	2	1	0	0	0	0	0	0
2.0	0	0	0	0	0	0	0	0	0	0
1.5	0	0	0	0	0	0	0	0	0	0

Table 3: Accepted papers with rebuttals (351 in number) and their post-rebuttal scores, *ignoring late reviews*. The 51 blue-highlighted ones were unlikely to have been accepted prior to post-rebuttal increase. The green ones likely would have been.

them, 29 with pre-rebuttal scores over 4.5 were likely ($> 87\%$) to be accepted in any event. Another 30 had a better than even chance. The remaining 21 had only a small prior chance ($< 23\%$).

Rebuttals are more likely for middling initial scores; see Figure 4. Among the 348 papers sans rebuttal, with their significantly lower scores (mean 2.4), there were only 29 acceptances (mean 4.2). Besides new reviews for 5 of the 29, there were a number of score changes, 13 up and 4 down, affecting 15. All but 3 changes moved scores toward consensus. Rejected papers without rebuttals also had a fair number of changes (39 papers; 49 changes), primarily downward (34). Clearly, reviewers often modify their scores even in the absence of rebuttals (15% of papers; 8% of reviews), suggesting that about half the changes have other motivations.

There are several reasons for reviewers to modify overall scores between initial and final evaluations:

1. Re-reading and re-evaluating the submission.
2. Taking the other reviews and scores into account, which were unseen by the reviewer before submitting her original review.
3. Taking into account new reviews, which arrived after the more timely reviews were sent to authors for feedback.
4. Considering clarifications provided by authors in their responses to issues raised in reviews.

Gao et al. [7] already determined that “peer pressure” and “conformity bias” motivate many changes. Only 121 review-score changes out of 498, up or down, were further from the mean after the changes than before.

Only 31 papers had a pre-rebuttal overall score below 4, yet saw an increase by 0.5 or more after rebuttal (not necessarily on account of the rebuttal) and were accepted (in any category). That amounts to 2.0% of submitted papers and 2.6% of papers with rebuttals. In 10 instances, more positive reviews arrived. Only in 9 was there an increase of an already above-average score.

For the sake of argument, let’s deem a rebuttal “effective” if:

(a) there was a rebuttal;
(b) the paper was accepted;

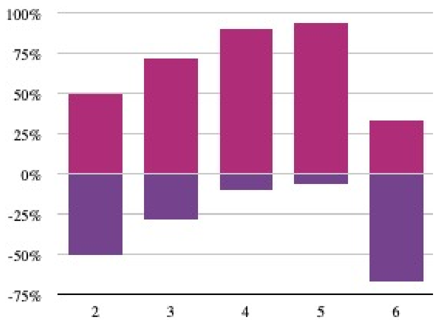


Figure 4: Percentage of papers with (up) and without (down) rebuttal. Columns are for initial average score ranges (1–2, 2–3, etc.).

(c) at least one referee increased his/her score to be further from the pre-rebuttal mean than it had been;
(d) no referee counterproductively decreased the score; and
(e) the mean overall grade increased non-negligibly (> 0.1) ignoring added reviews.

Only 23 rebuttals were effective in this sense, a mere 2%. Eight or nine of these would have been accepted anyway; for another five, late strong reviews came in to tilt the scales.

All 98 papers satisfying conditions (a,b)—with any increase at all—increased at least 0.3, obviating (e). None of these had negative changes, making (d) superfluous, too. Of those 98, some 39 would have been expected to be accepted based on their unrevised scores. Condition (c) suppresses the many changes that appear to be consensus building. Figure 5 accentuates the scanty decisions in both directions that may be attributed to the rebuttal.

A χ^2 test indicates that the distributions of changes (up, down, none) are significantly different (at $p = 0.01$), but that is only because of the confounding factor that there are more rebuttals for papers in the range 3–5, which are also those most likely to have their scores modified. Compare Figure 3 with 4. Clearly, authors are more motivated to rebut and reviewers to revise when the score is midrange. Indeed, for each initial score range, there is no significance (χ^2 gives p values of 0.25, 0.76, and 0.51 for ranges 2–3, 3–4, and 4–5, respectively).

Thus, it would seem that rebuttals do not lead significantly to improved referee scores. Rather, scores are changed up and down as they would have been regardless.

The Upshot

We estimate that only 1% of the rebuttals achieved their presumed goal of leading to acceptance by clearing up misconceptions or clarifying matters. A qualitatively similar conclusion regarding the large CHI conference was reached based on observations of mean score movement.[6] It is also likely that some disappointing rebuttals led to rejection of papers that might have otherwise been given the benefit of doubt. All in all, it is far from clear that the very minimal favorable impact justifies the enormous investment of effort entailed by across-the-board rebuttals.

Another moral to consider: Don't homogenize scores. Why mask the frequent diversity of evaluations, presenting a false façade of relatively consistent judgments? Perhaps conferences should insist that reviewers choose from a list of reasons to justify any change of score.

References

(1) Price, E. The NIPS Experiment, Moody Rd. Blog, December 15, 2014.

(2) Church, K. W. *Natural Language Engineering* **2020**, *26*, 245–257, DOI: 10.1017/S1351324920000030.

(3) Littman, M. L. *Communications of the ACM* **2021**, *64*, 43–44, DOI: 10.1145/3429776.

(4) Wang, G.; Peng, Q.; Zhang, Y.; Zhang, M. *What Have We Learned from OpenReview?*; preprint; arXiv, 2021, arXiv: 2103.05885 [cs.DL].

(5) Kaye, J. Do Rebuttals Change Reviewer Scores?, A Tumblr for SIGCHI, December 8, 2015.

(6) McGrenere, J.; Cockburn, A.; Gould, S. CHI 2020 – The Effect of Rebuttals, CHI 2020 blog, December 15, 2019.

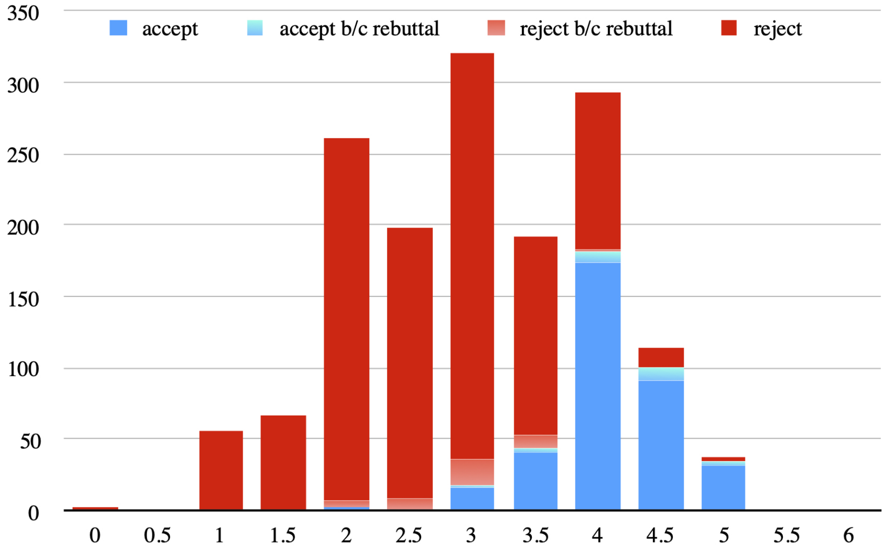


Figure 5: Distribution of final mean overall scores and outcomes. Shaded areas are attributable to rebuttal ("effective" if accepted, or its analog if rejected).

(7) Gao, Y.; Eger, S.; Kuznetsov, I.; Gurevych, I.; Miyao, Y. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, Association for Computational Linguistics: Minneapolis, MN, 2019, pp 1274–1290, DOI: 10.18653/v1/N19-1129.

(8) Gao, Y.; Eger, S.; Kuznetsov, I.; Gurevych, I.; Miyao, Y. ACL-18 Numerical Peer Review Dataset, https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2639, 2021.