DOI:10.1145/3584664

Nachum Dershowitz and Rakesh M. Verma

Opinion
Rebutting Rebuttals

Evaluating the impact of the conference review rebuttal process.

Peer review of research output stands as a cornerstone of quality control in scientific and scholarly publishing. In addition to articles appearing in journals—be they nonprofit or commercial—conference proceedings have become a prime venue for the dissemination of advances in many areas of the computer science discipline, and are often highly cited.[3] Publishing books, journal articles, and/or conference papers is critical for advancing one's career in academia or in research laboratories.

The extent of the conference reviewing process differs significantly from event to event: Workshop submissions normally undergo only light review, for instance, whereas large, focused conferences typically ask multiple experts to review and evaluate each potential contribution.

Many prominent computer science conferences, large and small, theoretical or applied, and spanning the whole gamut of research areas, have in recent years instituted rebuttal or feedback periods during which authors see preliminary reviews and are offered the opportunity to answer specific queries or to otherwise respond to issues raised in the referee reviews. And many conference management frameworks, such as ConfTool (Pro), EasyChair (paid Executive version), HotCRP, and OpenConf, support rebuttals.

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 issue.

Our analysis of five conferences shows the net impact in the acceptance/rejection decision does not seem to justify the significant effort required of authors in writing rebuttals and of reviewers in considering them before making a final decision.

There have been numerous studies of the quality and possible biases in conference reviews. In a controlled study of reviewing for the Conference on Neural Information Processing Systems (NIPS; now NeurIPS), the outcome for all but the extreme cases was more or less random.[12] 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.[10] It has been found that the order in which reviewers express their views in discussions does not seem to impact their post-discussion scores.[14] Some other relevant works are cited in Wang et al.[15] and in Shah.[13]

We address here the issue of rebuttal impact only. We follow up on the work of Gao et al.,[5] who analyzed the ACL 2018 reviews for impact of style and other matters, concluding inter alia that impoliteness hurts.

Members of the Conference on Human Factors in Computing Systems (CHI) 2016 and 2020 committees report-

ed 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."[9,11] Accordingly, we consider the effect of rebuttals for CHI 2020 and CHI 2021. We also briefly examine the impact of rebuttals on acceptance at two smaller 2022 conferences, namely, the International Conference on Formal Structures for Computation and Deduction (FSCD) and the International Conference on Theory and Applications of Satisfiability Testing (SAT). Several years' worth of International Conference on Learning Representations (ICLR) review data is available[7] (see Wang et al.[15]). Unfortunately, pre-rebuttal scores are not included. Another dataset of reviews, PeerRead,[8] lacks actual numerical evaluations and does not include rebuttals.

The Reviewing Process

The *Proceedings of the ACL 2018* conference begin with a report from the program committee co-chairs explaining the reviewing procedure.[6]

The process began with the ACL steering committee inviting two senior academics to serve as program committee co-chairs. The co-chairs first decided on the different areas to be covered, and then proceeded—in September 2017—to post a call for nominations of area chairs (AC) and (primary) reviewers, the latter constituting the "program committee." In the end, 61 area chairs were selected, out of about 300 nominations. There were 936 valid nominees for reviewing; more than half, self-nominations. Additional reviewers were invited for a total of 1,473 people, ranging in seniority from doctoral students to full professors.

Each committee member reviewed three papers on average. The assignment of papers to areas and reviewers was done in stages: first, an initial assignment by area of interest; then, area chairs were chosen to serve as "meta-reviewers" for each paper, whose task was to summarize and evaluate all the reviews; next, the committee took possible conflicts of interest into account and also balanced loads; finally, area chairs made the final assignments to individual reviewers.

The reviewers' initial reviews were shared with the authors, who were asked to respond and point out factual inaccuracies. These rebuttals were then supposed to be read by the reviewers, who could revise their scores and reviews based on them. Area chairs attempted to iron out disagreements between the different reviews by engaging in online discussions. They were then tasked with writing a meta-review, taking everything into consideration. All these were used by the co-chairs to make final accept/reject decisions.

One innovation was a structured, argument-based review form, in which reviewers were asked to provide arguments pro and con acceptance. These were found to be quite helpful in making final recommendations. Authors were asked specifically to respond in their rebuttals to the con arguments in the reviews. ACL 2018 did not have an option for revising and resubmitting the paper, as do some other conferences.[a]

Quoting from the report,[1] the main criteria of acceptance were as follows:

- strengths/weaknesses raised by reviewers and their significance;
- the result of discussions and author responses;
- contribution to [computational linguistics] as the science of language: whether the paper advances (or contributes to) our understanding of language in any way; and
- diversity: we do not want to fill ACL with similar papers like achieving 1% improvement on a well-known task.

Out of the 1,610 reviewers, 1,473 primary and 137 secondary (that is, appointed by primary reviewers), 192 "were recognized by the area chairs as outstanding reviewers who have turned in exceptionally well-written and constructive reviews and who have actively engaged themselves in the post-rebuttal discussions."[1] A report on the reviewing process and its quality is available.[1]

a This option was introduced with ACL 2022.

It would seem rebuttals did not lead significantly to improved referee scores.

The Dataset

The released dataset consists of review scores for all submissions to ACL 2018.[4] For each paper, besides an ID number, there are the following data items:

1. submission type (long or short) and track;
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, low to high);

(b) final (post-rebuttal) overall score;

(c) initial and final reviewer confidence levels (1..5);

(d) initial and final subscores (for originality, soundness/correctness, substance, replicability, meaningful-comparison, and readability);

(e) initial and final assessment of contributions in various categories;

(f) some checks for appropriateness and adherence to guidelines.

For late reviews, there are only final scores.

The figures are as follows:

- 1,545 papers submitted (excluding papers that were withdrawn prior to refereeing or summarily rejected);
- 3,875 reviews initially, averaging 2.5 reviews per submission; 4,059 all told, including late arrivals (2.6 reviews/submission);
- 1,197 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 downward; 17 were upward; 28 were evenly balanced; one 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;
- 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.

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, given as ranges (0.5–1.0, 1.0–1.5, and so forth) on the bottom. The orange/grey bars reflect post-rebuttal rejections/acceptances per final score.

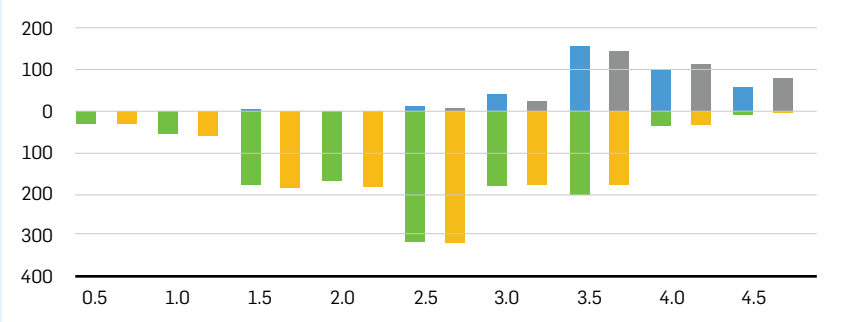


Figure 2. Box-and-whisker plots for acceptance (purple) and rejection (red), showing medians (bars in middle of colored rectangles), first and third quartiles (rectangle edges), upper and lower fences (whiskers extend to include all data points within 1.5 times the interquartile range), and scattered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

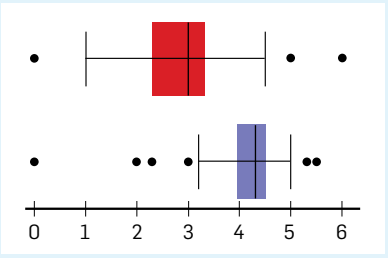


Table 1. The number of reviews, before (columns) and after (rows) rebuttal, at each level of overall scores, (0, 1], (1, 2], and so forth. The overwhelming majority of reviewers leave their scores intact during the second round. Higher scores are more likely to move down; lower scores, to move up. (After Gao et al.[5])

	1	2	3	4	5	6
6	0	0	0	0	1	11
5	0	1	2	36	395	2
4	1	18	104	1011	63	0
3	2	61	844	121	11	0
2	19	861	31	12	2	0
1	260	3	2	1	0	0

Prior Analysis

Most of the authors of Gao et al.[5] were deeply involved in the refereeing process of ACL 2018, two as program committee chairs, and two as their assistants. The original intention was to release the full reviews (85% of the reviewers consented to this), anonymized, but concerns regarding pressure to consent precluded fulfilling that promise.[b] Thus, we were only given access to submission type, overall scores, confidence levels, various subscores, and the decisions, whereas the analysis in Gao et al.[5] is based also on the texts of the referee reviews and author responses, plus additional information about reviewers.

Using quantitative and qualitative analysis of the reviews before and after rebuttals, Gao et al.[5] found (1) the overall scores correlated best with subscores for soundness and substance, (2) reviewers who submit early have slightly lower scores on average, and those papers are more likely to have post-rebuttal score increases, (3) the more successful rebuttals were more likely to refer back explicitly to locations in the original paper, (4) impoliteness can harm the final score, (5) final scores are largely determined by a reviewer's initial score in comparison to that of other reviewers, (6) author responses have a significant, but marginal impact on the final scores, and (7) unsurprisingly, reviewers seem to pay more attention to responses when it can make the difference between acceptance or rejection. Furthermore, manual assessment of weaknesses identified by reviewers showed that 28% of the criticisms were with regard to evaluation and analysis and 18% with regard to the quality of writing. The authors proposed a task of machine prediction of final scores based on initial reviews and author responses.

Table 1, from Gao et al.,[5] shows changes in individual review scores after the rebuttal period. In their detailed analysis, the authors found 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 1,119 papers with no score update (22.8%). Hence, the score update has a large impact on the final accept/reject decision."

This conclusion, however, is unwarranted in our opinion. 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 to be expected. To be sure, we concur updates are "largely determined by the scores of peer reviewers."

Our Analysis

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 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

b Personal communication, Ilia Kuznetsov, December 2020.

Table 2. Mean overall scores before (columns) and after (rows) rebuttal. The heading refers to range [x – 0.5, 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	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	2	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	7	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 3. Accepted papers with rebuttals (351 in number) and their post-rebuttal scores, ignoring late reviews. Headings are as in Table 2. The 51 blue highlighted ones were unlikely to have been accepted prior to post-rebuttal increase. The green ones likely would have been.

	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	1	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	6	0	0	0	0	0	0
1.5	0	2	0	0	0	0	0	0	0	0

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 was not a rebuttal. Most reviewers, of course, do not 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: five increased more than one point and six decreased that much. Most are ascribable to additional reviews.
- There were only four contributions whose original referee(s) changed their overall score more than one point: 2 increased and were accepted; two decreased and were rejected.
- Another eight accepted papers with rebuttals increased one point, likely due to the positive influence of the author response; some would have been accepted regardless.
- Ignoring late reviews, mean scores for 413 papers with rebuttals changed. For half (206), post rebuttal scores were higher; for half they were lower. Virtually all changes were less than 1 point.
- Of the 206 with increased scores, 104 were accepted. Of the 207 with lowered scores, almost all (186) were ultimately rejected.
- Of the 186 rejections, 51 had had an even chance or better of acceptance before their scores decreased and they were rejected.

Table 3 highlights the 80 (in blue and green) accepted papers with increased post-rebuttal scores—at a granularity of 0.5. (So 26 papers moved up too little to show in the table.) Of them, 29 with prerebuttal 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 three 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 approximately 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.

5. Procedural issues raised by chairs or other reviewers.

As mentioned, Gao et al.[5] 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.

Interestingly, whenever there was a rebuttal and the mean overall scores of the original reviewers increased afterward, there were no reviewers who lowered their score. In the opposite direction, this uniformity was also nearly always the case; there were only four contributions in which mixed score revisions resulted in an overall decrease. This suggests that uniform motion is indicative of re-evaluation by the cadre of referees involved in the review of the paper. Still, this may be because of consideration of points raised in other reviews, rather than in the rebuttal. Certainly, that is likely the case when scores decrease.

It may also be interesting to note that in 32 cases a reviewer giving the highest score increased it, while in a comparable 35 cases the lowest scorer lowered the score even further.

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 ten instances, more positive reviews arrived. Only in nine was there an increase of an already above-average score.

For the sake of argument, let us deem a rebuttal "effective" if:

(a) there was a rebuttal;

(b) the paper was accepted;

(c) the mean overall grade of the original reviewers increased non-negligibly (≥ 0.1);

(d) no reviewer counterproductively decreased the score;

(e) there were no above-average late reviews; and

(f) at least one reviewer increased his/her score to be further from the pre-rebuttal mean than it had been.

Only 24 rebuttals were effective in this sense, a mere 1.6%. Eight or nine of these would have been accepted anyway, so maybe 1% were accepted on account of an effective rebuttal.

All 100 papers satisfying conditions (a,b)—with any increase at all—increased at least 0.17, obviating (c). Only three of these had negative changes (d), but nine had late reviews that upped the average overall score (e). Condition (f) suppresses 64 additional papers whose changes appear to be mainly 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 rebuttals did not lead significantly to improved referee scores. Rather, scores are changed for the most part up and down as they would have been regardless.

Additional Analyses

We performed similar analyses for four additional meetings: CHI 2020; CHI 2021; FSCD 2022; and SAT 2022. Note that different conferences use varying score scales and increments.

CHI 2020 and 2021. These are large conferences with close to 3,000 submissions. In 2020, out of 3,125 (com-

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), and so forth.

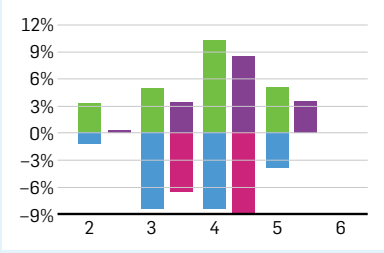


Figure 4. Percentage of papers with (up) and without (down) rebuttal. Columns are as in Figure 3.

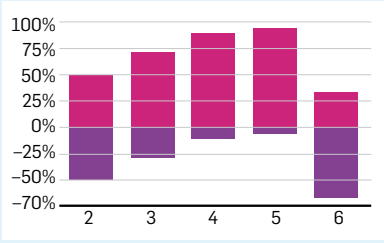
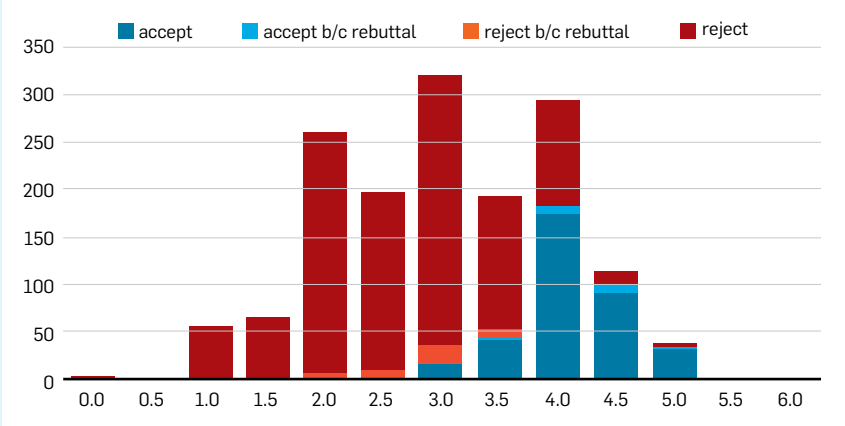


Figure 5. Distribution of final mean overall scores and outcomes. Shaded areas are attributable to rebuttal ("effective" if accepted, or its analogue if rejected).



pleted) submissions, 647 were accepted outright and 113 were accepted contingent on shepherding. Of the 760 rebutted acceptances, 449 saw their score (on a scale 1..5) increase (by at least 0.1) post-rebuttal (0.89 points on average) from an initial marginal value (below the highest reject score of 3.63) and were ultimately accepted, about half with shepherding.

The numbers, following our definition of effectiveness, are as follows:

(a) 2,274 papers (73%) of the 3,125 completed submissions had rebuttals.

(b) 760 acceptances. All had rebuttals. Of these, 175 were to begin with clear accepts.

(c) 574 accepted papers had a post-rebuttal increase of at least 0.1.

(d) 545 of them did not also have an above-average late review.

(e) 518 of the latter also had no post-rebuttal decreased score.

(f) 344 submissions had at least one review score increase to move further away from the prerebuttal average. The other 174 only had movements toward consensus, so were more likely changed on account of consideration of other reviews.

Of these 344 effective rebuttals, 229 (69%) would likely have been accepted anyway based on the acceptance rates for their pre-rebuttal scores. A score of 3, for example, had a 50-50 a priori chance. This leaves 115 papers, or 3.7% of submissions, that likely would have been rejected were it not for the rebuttal, new reviews, and subsequent discussions. This is the percentage that meets the definition of effective rebuttal given above and would otherwise not have been expected to be accepted.

The statistics were quite similar in 2021. Out of 2,844 submissions, 749 were accepted (outright or shepherded). There were 2,120 rebuttals. Reviewer scores showed a non-negligible increase for 551 accepted papers; 489 of those had no decreased scores and no late high scores. Disregarding those with no counter-consensus movement leaves 320, 68% of which would have been accepted regardless. That leaves 103 presumably positively impacted papers, or 3.6%.

Rebuttals presumably provide other benefits but nonetheless appear to leave much to be desired.

As already mentioned, the CHI 2016 and 2020 committees expressed skepticism of the worthwhileness of the rebuttal stage.[9,11,c] We note that CHI has since abandoned rebuttals and moved to the more journal-like "revise and resubmit" process.

FSCD 2022. There were 59 submissions to FSCD in 2022, 31 of which were accepted and 28 rejected. Each submitted paper underwent three reviews, except for four submissions with an extra one, and one with an extra late review. Overall grades were given on a scale of –3..3. The highest average grade of a rejected paper was 1.0. Of the 33 papers with a rebuttal and with a pre-rebuttal score not above 1.0, nine saw their average referee score increase, and seven of those were accepted, while five suffered a post-rebuttal decrease and were rejected. Of those seven increases, three were clear consensus building, with the lowest score simply raised to the next lowest. Of the four that manifested increases that were not mere consensus building, one or two would likely have been accepted regardless of the rebuttal. So, it is fair to say that about 3%–4% likely benefited from the opportunity to rebut, while perhaps 2%–3% were rejected and—if anything—suffered from an unsatisfactory rebuttal.

Like for other conferences, reviewers sometimes update their reviews—for various reasons—without revising their scores, which can still impact acceptance. So, it bears keeping in mind that numbers alone do not tell the whole story.[4,d]

SAT 2022. There were 70 submissions to SAT 2022, 31 of which were accepted. Authors of all but four took advantage of the opportunity to supply a rebuttal. In addition, some authors were explicitly requested to submit a revised paper addressing specific requests for re-evaluation. (Six out of seven such were accepted after the requested revision.) The mean overall score of 16 papers increased after rebuttal versus nine that had their scores lowered. The critical evaluation score was a weighted sum of four subscores, with the reviewer's confidence level factored in as well. Based on the mean overall score before rebuttal, 10 might have been rejected but were eventually accepted after their average score increased. (Four of these 10 rebuttals included a revised paper, as requested by the program committee.) Of those increases, four were clear moves toward consensus (lowest outlier was raised), one other was also a consensus-building change, another received a high score from an extra reviewer, and another one or two would likely have been accepted anyway. So for this conference, only one or two submissions (1%–3%) derived actual benefit

c https://chi2022.acm.org/2021/05/26/moving-to-a-revise-and-resubmit-review-process-for-chi-2022

d We did not have access to the FSCD reviews themselves.

Table 4. Summary statistics for the conferences. N: number of submissions; Acc.: percentage of N accepted; Reb.: percentage with rebuttals; Up/Down: nontrivial post-rebuttal overall score changes (> 0.1) as percentage of papers with rebuttals—in both directions; Impact: estimated positive/negative impact on outcome, as percentage of submissions.

Conference	N	Acc.	Reb.	Up/Down	Impact
ACL '18	1,545	25	78	15/14	1.0/0.9
CHI '20	3,125	24	73	36/19	3.7/1.4
CHI '21	2,844	26	74	37/18	3.6/1.4
FSCD '22	59	52	55	23/9	4.4/2.6
SAT '22	70	44	94	24/14	2.8/1.0

from the rebuttal (or revision) and the concomitant reconsideration. The (sometimes dramatically) decreased post-rebuttal scores were mostly in reaction to unsatisfactory rebuttals, and a third of the time due to failings noted by other reviewers.

Alternatives

There are several alternative models of review. In some programming languages, software engineering, and security conferences, a reviewing style called "identify the champion" is used.[e] In this setup, reviewers are asked to declare whether they will champion each paper. With rare exceptions, only papers that are championed by a committee member have a reasonable shot at acceptance.

There is also an "open" review model, which is currently popular among machine learning conferences.[f] Reviews and author responses are published online anonymously. The idea is that reviewers will be more careful in their reviews knowing they will be posted online.

In some conferences, rebuttals are restricted to queries from the reviewers or to correct factual inaccuracies. ACL, however, and many other conferences allow open-ended rebuttals within some length limit. We have also seen cases where authors are given the opportunity to revise their contribution in light of specific complaints and resubmit for reevaluation.

The Upshot

Many factors go into the peer evaluation of conference submissions. In ACL-like conferences, an average score below the median or mean (3.2 in ACL's case) is typically considered a "lost cause" for a paper. Rebuttals—when available—play only a small part in the evaluation process. We estimate that fewer than 1% of the ACL rebuttals achieved their presumed goal of leading to acceptance by clearing up misconceptions or clarifying matters. For the other conferences we examined, the percentage for which the rebuttal might have been a positive factor—but often not the sole factor—in acceptance appears higher (3%–4%). Table 4 summarizes the estimated impact of rebuttals at the conferences analyzed. The negative impact is the estimated percentage of papers likely to have been accepted were it not for the (presumably disappointing) rebuttal, using analogous criteria for positive impact. Confounding factors in our analysis include: revising the review text, but not changing the score when it is warranted.

The recourse of shepherding, as in CHI, in all probability increases the success rate of rebuttals. Likewise, the option of requesting a revised submission, as in SAT, probably leads to an increased chance of eventual acceptance. Such procedural allowances make conferences more journal like.

Rebuttals presumably provide other benefits, such as helping keep reviewing fair and raising the standard of committee discussions, but nonetheless they appear to leave much to be desired, at least from the point of view of their impact on acceptances and rejections.

Side by side with the occasional positive effect of rebuttal, it is likely that some disappointing rebuttals lead to rejection of papers that might have otherwise been given the benefit of doubt. Furthermore, for competitive conferences, acceptance is a zero-sum game. For each paper accepted thanks to a convincing rebuttal, another paper—judged weaker—is rejected. In this sense, one person's gain is another's loss.

Another moral to consider: Do not homogenize scores. Why mask the frequent diversity of evaluations, presenting a false facade of relatively consistent judgments? It would be more informative if a reviewer did not change her or his score to match those of colleagues, but rather only if and when others raised salient points that lead to reconsideration. Perhaps conferences should insist that reviewers choose from a list of reasons to justify any change of score.

There are diverse approaches being taken by different computer science conferences in an effort to improve the overall quality of the refereeing process. These include on open reviewing process, queries and author responses, open-ended author rebuttals, and the opportunity for shepherding or for revision in equivocal cases. Some conferences specify rigorous review criteria.[g] Some program chairs prod referees to explicitly refer to rebuttals. Each approach has its plusses and minuses.

Wholesale rebuttals, analyzed here, should be carefully reconsidered. All in all, it remains far from clear that the minimal favorable impact justifies the frequently enormous investment of effort entailed by across-the-board rebuttals. Analysis of reviewing data from five conferences shows the minimal impact of rebuttals on acceptance versus rejection. C

e http://scg.unibe.ch/download/champion

f As of 2022, ACL has moved to this model.

g See https://doi.org/10.48550/arXiv.2010.03525 for such a 20-page document.

References

1. ACL Committee. *A Report on the Review Process of ACL 2018*; https://bit.ly/3DHmaAw
2. Church, K.W. Emerging trends: Reviewing the reviewers (again). In *Natural Language Engineering 26.2* (2020); DOI: 10.1017/S1351324920000030
3. Freyne, J. et al. Relative status of journal and conference publications in computer science. *Commun. ACM 53*, 11 (Nov. 2010).
4. Gao, Y. et al. ACL-18 Numerical Peer Review Dataset (2021); https://bit.ly/3QlKbEz
5. Gao, Y. et al. Does my rebuttal matter? Insights from a major NLP conference. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*. Minneapolis, MN: Association for Computational Linguistics (June 2019); DOI: 10.18653/v1/N19-1129
6. Gurevych, I. and Miyao, Y., Eds. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics (July 2018).
7. International Conference on Learning Representations. The ICLR Open Reviews Dataset (Nov. 2020); https://bit.ly/47d44DZ
8. Kang, D. et al. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, LA: Association for Computational Linguistics, (June 2018); DOI: 10.18653/v1/N18-1149
9. Kaye, J. Do Rebuttals Change Reviewer Scores? A Tumblr for SIGCHI (Dec. 8, 2015); https://bit.ly/3qdBcL2
10. Littman, M.L. Collusion rings threaten the integrity of computer science research. *Commun. ACM 64*, 6 (June 2021); DOI: 10.1145/3429776
11. McGrenere, J. et al. CHI 2020—The Effect of Rebuttals. CHI 2020 blog (Dec. 15, 2019).
12. Price, E. The NIPS Experiment. Moody Rd. Blog (Dec. 15, 2014); https://bit.ly/47jdCgA
13. Shah, N.B. Challenges, experiments, and computational solutions in peer review. *Commun. ACM 65*, 6 (June 2022).
14. Stelmakh, I. et al. A Large Scale Randomized Controlled Trial on Herding in Peer-Review Discussions (2020); arXiv: 2011.15083
15. Wang, G. et al. What Have We Learned from OpenReview? (2021); arXiv: 2103.05885

Nachum Dershowitz (nachum@tau.ac.il) is a professor in the School of Computer Science at Tel Aviv University, Tel Aviv, Israel.

Rakesh M. Verma (rmverma2@central.uh.edu) is a professor of computer science in the Department of Computer Science at the University of Houston, TX, USA.