Analysis of Reviews for CVPR 2012

Aditya Khosla Derek Hoiem Serge Belongie

Abstract

1. Introduction

Peer reviewing is an important component of academic conferences used for quality assurance and maintaining high quality of published work. Reviewing papers can be a relatively challenging and time-consuming task. Unfortunately, given the anonymous nature of the reviewing process, there is little incentive to write high quality reviews. While most reviewers do spend consierable effort and write high quality reviews, there is a non-negligible number of lower quality reviews, at least in the authors' perception. In a post-CVPR 2011 survey, the quality of reviews was found to be of concern to a relatievly large number of people.

In order to better evaluate reviewers, we conducted an author survey as part of CVPR 2012. Through the survey, we have a measure of reviewing quality, which can help to identify systematic problems or to show that problems are rare. Further, it can be used to identify high/low quality reviewers, and potentially reward reviewers based on their performance. More specifically, authors could assess each review they had received according to three criterion, namely fairness, helpfulness and negligence. After accounting for review score, we found that $X\%$ of the reviewers received excellent ratings all-around, while $X\%$ of the reviewers were rated poorly by majority of the authors.

In this report, we explore the effect of various factors on the survey responses, such as review length and review decision. We also look at the relationship between the paper decision and review scores, and the predictiveness of the final decision from the review scores. Further, we also explore the problem of reviewer bias. Some reviewers may tend to assign higher or lower paper scores than others. For example, area chairs tend to rank papers by scores when making their final decisions and a bias between different reviewers could significantly affect the rank and thereby the outcome of the paper. We explore the extent of this phenomenon in CVPR 2012, and hope to include the option to rectify this bias when ranking papers in CMT for future conferences.

Overall, we provide recommendations to authors, reviewers, area chairs and program chairs for future conferences based on the author feedback. . .

2. Review Feedback Design and Collection

We designed an optional survey to collect feedback from authors about the reviews. This survey was included in the Conference Management Tool (CMT) as part of the paper rebuttal phase, and thus there was only one survey per paper. The paper decision was not available at the time of the survey. The survey responses were not visible to the area chairs to ensure that they did not affect the paper decision. We informed the authors that the survey was designed only to collect feedback, and had no impact on the outcome of the paper, in order to reduce review-score based biases.

The survey consisted of three questions per review (with a total of nine questions per paper):

- *Fairness:* Are you satisfied that Reviewer X carefully considered the paper and wrote a fair analysis? (Responses: 1=Very satisfied, 5=Very dissatisfied)
- *Helpfulness:* How much did Review X help you, in terms of writing a better paper or ideas for future research? (Responses: 1=Very helpful, 3=Unhelpful)
- *Negligence:* Was Review X negligent or abusive? (Responses: Yes or No)

The *Fairness* rating is aimed at understanding whether the authors agreed with the reviewers' assessment of their work. We can expect this to be highly biased against reviewers who gave less favorable scores, but it is useful for understanding the extent of the authors biases, and to get an overall summary of the authors' perceptions.

The *Helpfulness* rating is aimed at identifying whether authors thought the reviews could help to improve the quality of their work, and whether the reviewers provided useful feedback. We expect this to be less biased, and potentially useful for rating the quality of the reviewers.

The *Negligence* rating was used to flag inappropriate reviews that did not follow the guidelines or were generally abusive. We expected authors to identify reviewers that might not have read their papers in detail or used language

that was unwarranted. This was also used to identify particular reviewers that got marked negligent frequently by independent authors.

3. Analysis

3.1. Paper Ratings and Decisions

Here, we analyze the relation between the paper scores (assigned by reviewers: 1=definite accept, 5=definite reject) and final decisions (oral, poster, reject). We also examine the variance of the scores. We aim to elucidate the degree of randomness in the review process and the extent to which area chairs base their decisions on reviewers' recommendations.

Predictiveness of decision from paper ratings. Generally, we would hope that area chairs would accept reviewers' recommendations when all reviewers agree and otherwise use their judgement in making decisions. In Figure 1, we plot the probability of the average score given the decision and the probability of the decision given the average score. We can see that paper decisions depend strongly on the scores. Starting with an average score of borderline (3.0), each increment of 0.33 (e.g., one reviewer incrementing the score) yields roughly a 30% chance of a better decision. Our later analysis on reliability of scores supports the high reliance for the accept/reject decision but not the oral/poster decision.

Variance of paper scores. Given that paper scores strongly determine the final outcome (oral, poster, reject), we would like to know whether these scores are reliable. For example, if we submitted a paper through identical but independent review processes, how much would its average scores differ? There are many sources of variance: different reviewers may genuinely have different opinions of the papers; the quantization into discrete scores may exaggerate differences; differences in expertise may determine the likelihood of confident ("definite") ratings; reviewers may have different standards for acceptance; a paper may not be clear enough for a particular reviewer to understand the contributions or experiments; or a reviewer may not spend sufficient time to consider the paper. Some of this variance is intrinsic to the subjectivity of any review process, and some could potentially be reduced through improved procedures or reviewing effort.

The simplest analysis is to estimate the standard deviation of the scores for each paper. Given a sample $x_1 ... x_n$ drawn from a normal distribution, the maximum likelihood estimate (MLE) of the parameters are: $\mu_{MLE} = \frac{1}{n}\sum_i x_i$ and standard deviation $\sigma_{MLE} = \sqrt{\frac{1}{n}\sum_i (x_i - \mu_{MLE})^2}$. The MLE provides a biased underestimate of the variance because the mean is estimated using the same samples. The unbiased estimate of the variance is $\sigma^2_{n-1} = \frac{1}{n-1}\sum_i (x_i - \mu_{MLE})^2$. The estimate of $\sigma_{n-1} = \sqrt{\frac{1}{n-1}\sum_i (x_i - \mu_{MLE})^2}$ is a slightly biased estimate of standard deviation, for reasons that can be found in most statistics textbooks.

For simplicity, we assume that each paper has a different mean score μ_i but the same variance σ^2. Realistically, an exceptionally good or bad paper might get more consistent scores than one that has more balanced strengths and weaknesses, but our assumption of shared variance is necessary to have sufficient samples. Analogous to the expressions for a normal distribution, we can compute $\mu_{iMLE} = \frac{1}{3}\sum_j s_{ij}$ and $\sigma_{MLE} = \sqrt{\frac{1}{N_i(3)}\sum_i \sum_j (s_{ij} - \mu_{iMLE})^2}$ and $\sigma_{n-1} = \sqrt{\frac{1}{N_i(3-1)}\sum_i \sum_j (s_{ij} - \mu_{iMLE})^2}$, where s_{ij} is the jth score for the ith paper, N_i is the number of papers, and 3 is the number of reviews per paper. We obtain estimates $\sigma_{MLE} = 0.739$ and $\sigma_{n-1} = 0.905$, which indicates that the standard error of the mean estimate from three reviews is $SE_{MLE} = 0.427$ or $SE_{n-1} = 0.523$.

Variance analysis through score resampling. Another way to analyze the variance in paper ratings is by resampling a paper's scores given its observed scores. We have a set of 1740 score triplets that we can use to estimate the probability of a third paper score given the first two. Using this probability estimate, we can use each pair of observed scores to sample three new scores for each paper. By repeating this process, we can create a distribution of possible score triplets based on the observed score triplets. The sampled scores should have slightly higher variance than the observed scores because each score is based on only two observations. Indeed, the estimated standard deviation for the sampled scores is $\sigma_{MLE} = 0.843$ and $\sigma_{n-1} = 1.033$, slightly higher than for the observed score triplets. In Figure 2, we show the mean and standard deviation of the resampled score distributions given each original score triplet. One observed trend is a regression to the mean: if a paper scores very well (e.g., 1, 2, 2), we would expect a worse score if the same paper were reviewed again through an independent process.

We can make two important conclusions. First, the expected mean of good scores deviates from the observed mean much more than the expected mean of bad scores. In other words, a very poor rating would likely be repeated, but a very good rating would not. Second, the resampled distributions for average scores from 1.0 to 2.0 are extremely similar, yet the probability of a paper receiving an oral is strongly determined by its precise score (Figure 1). This confirms the

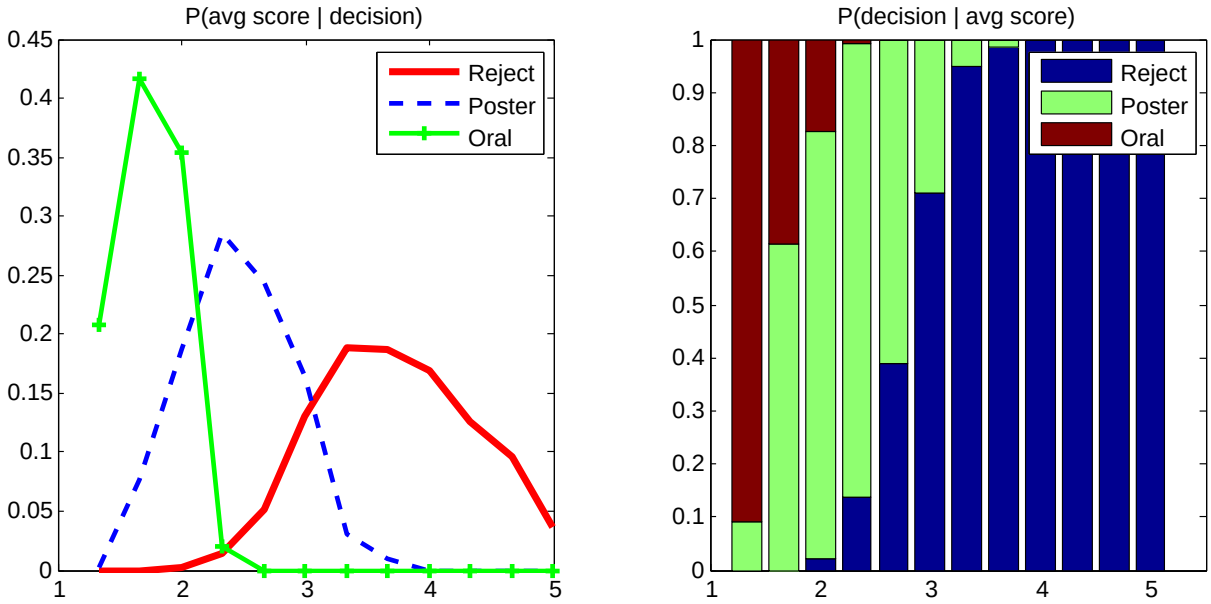

Figure 1. **Relation between paper scores and decisions.** There were only two papers with average scores of 1 (both were accepted as posters); these papers were merged with the group of papers with average scores of 1.333 for analysis.

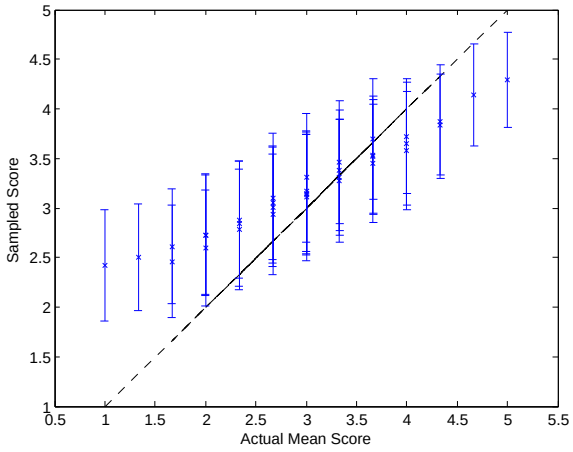

Figure 2. **Variance analysis through resampled scores.** Based on the empirical probability of paper score triplets, we sample a new distribution of score triplets for each observed triplet. Above shows the mean +- one standard deviation for each observed score. For example, a paper that had a score of (4,5,5; average=4.67) has a resampled mean of 4.14 with 0.52 standard deviation of the mean estimate. Some unique score triplets have the same mean; an error bar is plotted for each.

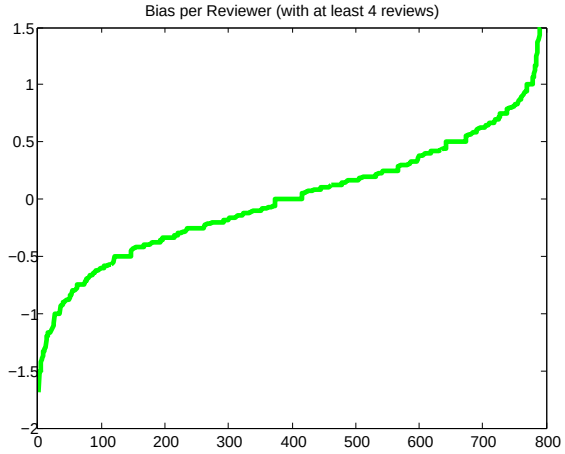

Figure 3. **Reviewer bias.** Some reviewers tend to be more positive or negative than other reviewers. Mostly, though, the bias is small, considering that a difference of "1" is a single rating increment and the sample size per reviewer is small. Experiments in correcting for bias show a small reduction in paper score variance.

conventional wisdom that the poster/oral decision is particulary unpredictable and suggests that area chairs should not use the average paper score as a determinant for the decision.

Reviewer bias. Some reviewers may tend to assign higher or lower paper scores than others. We consider a reviewer's bias to be the average difference between the reviewer's score and the average of the other two scores for each paper. In Figure 3, we show the bias for each reviewer with at least four reviews.

If these biases are significant, then it may be possible to improve paper score reliability by subtracting the bias for each reviewer. We tried this, estimating the bias for each review with a leave-one-review-out analysis. To rescore a review, the reviewer bias was computed based on all other reviews

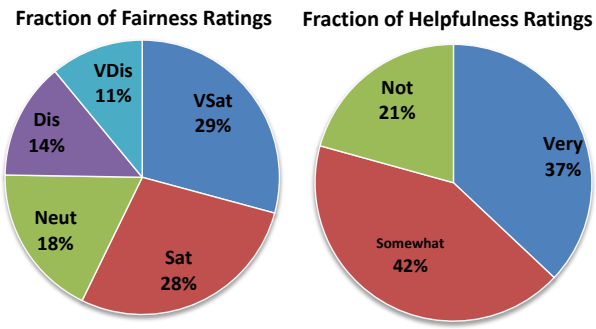

Figure 4. **Overall feedback ratings.**

from that reviewer, and the bias was subtracted from the assigned score. Because the sample size is small, we added a sample prior of 5 samples with zero bias (this helps and sample sizes from 3 to 20 yield similar results). Our bias-adjustment reduces the maximum likelihood estimated standard deviation σ_{MLE} from 0.739 to 0.711 and σ_{n-1} from 0.905 to 0.871. The largest change in the average score for a paper was 0.604, with an average (mean) change of 0.13. It may be worth computing bias-adjusted scores as an additional column so that the area chair is aware if a reviewer has a tendency towards negative or positive ratings.

3.2. Helpfulness and Fairness Ratings

In the rebuttal form, authors had the option to rate each review in terms of fairness (scale 1 to 5) and helpfulness (scale 1 to 3). The overall statistics are shown in Figure 4. We name the helpfulness categories "Very" (very helpful), "Somewhat" (somewhat helfpul) and "Not" (not helpful). We name the feedback categories "VSat" (very satisfied), "Sat" (satisfied), "Neut" (neutral), "Dis" (dissatisfied) and "VDis" (very dissatisfied). 75% of the fairness ratings and 79% of the helpfulness ratings are neutral or better. As shown in the analysis below, the paper rating is a major determinant of the perceived fairness or helpfulness of a review, though review length is also positively correlated with perceived helpfulness.

Relation between paper rating and perceived review quality. As shown in Figure 5, more positive reviews are widely considered to be more fair and more helpful. Nearly all reviews with "weak accept" or "definite accept" scores are considered fair. Most "definite reject" ratings are considered very unfair. A strong correlation between how much the reviewer likes the paper and how much the authors like the review is unsurprising, but these results confirm that some of the perceived problems with review quality are due to disagreement between the authors and reviewers about the submitted paper quality.

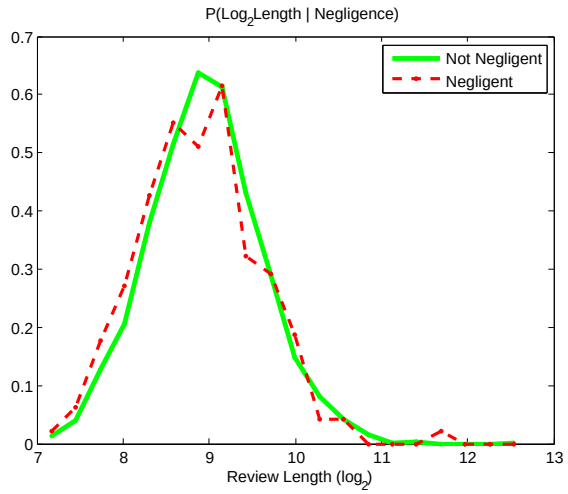

Figure 8. **Comparison of review lengths as a function of negligence flag.** Whether a review is considered "negligent" seems to be independent of its length.

Relation between length and perceived review quality.

One might think that a longer review is likely to be considered more helpful and fair, as length may indicate effort. We examine this hypothesis in Figures 6 and 7. Length does have a moderate impact on perceived helpfulness and a minor impact on perceived fairness, but the influence of the paper score is much greater. A very short positive review is much more likely to be considered helpful and fair than a very long negative review.

3.3. Negligence

The authors also had an option to flag reviews as negligence. Surprisingly, 10% of reviews were flagged as negligent, but these reviews did not seem to differ from others, except in the paper rating (mean of 4.2 for negligent, 3.0 for not). For example, Figure 8 shows the nearly identical distribution of paper lengths for reviews considered negligent and non-negligent.

4. Rating Reviewers

We would like to rank reviewers according to the helpfulness and fairness ratings provided by the authors and the predictiveness of their reviews. Such rankings could be used to nominate reviewers for awards, along with feedback from the area chairs.

Helpfulness and Fairness. To rank reviewers by helpfulness or fairness, we need to normalize the ratings according to the paper scores. Our approach is to consider the helpful-

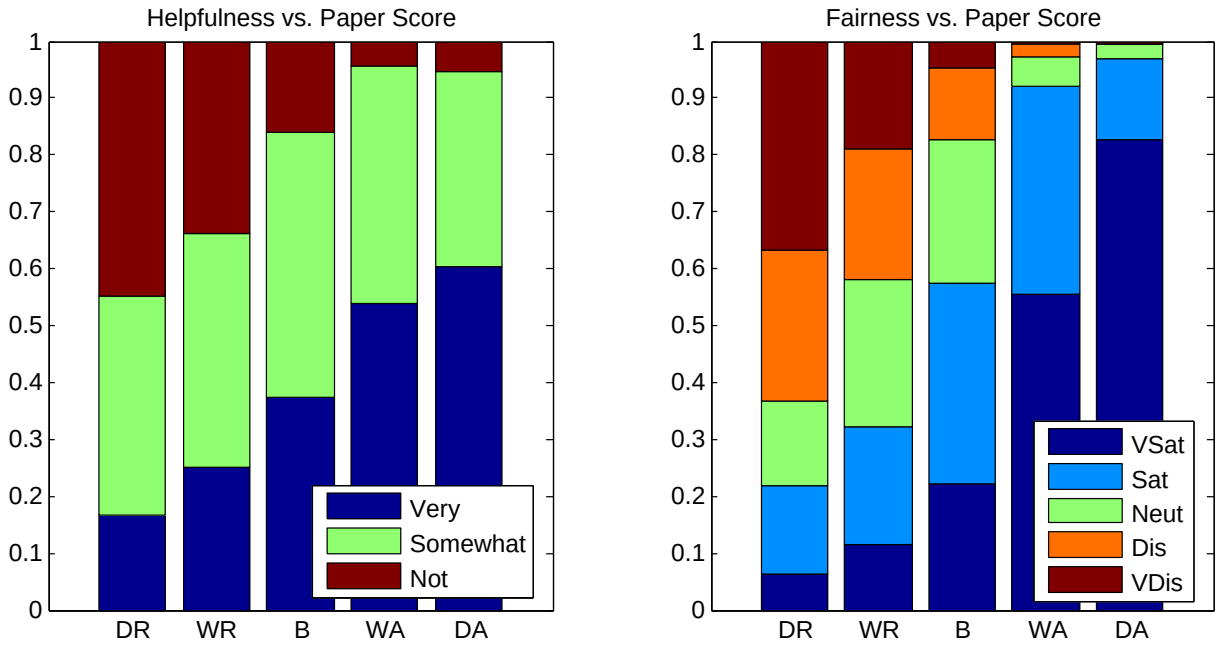

Figure 5. **Paper Score vs. Fairness/Helpfulness.** Perceived fairness and helpfulness as a function of the score that the reviewer assigned to the paper.

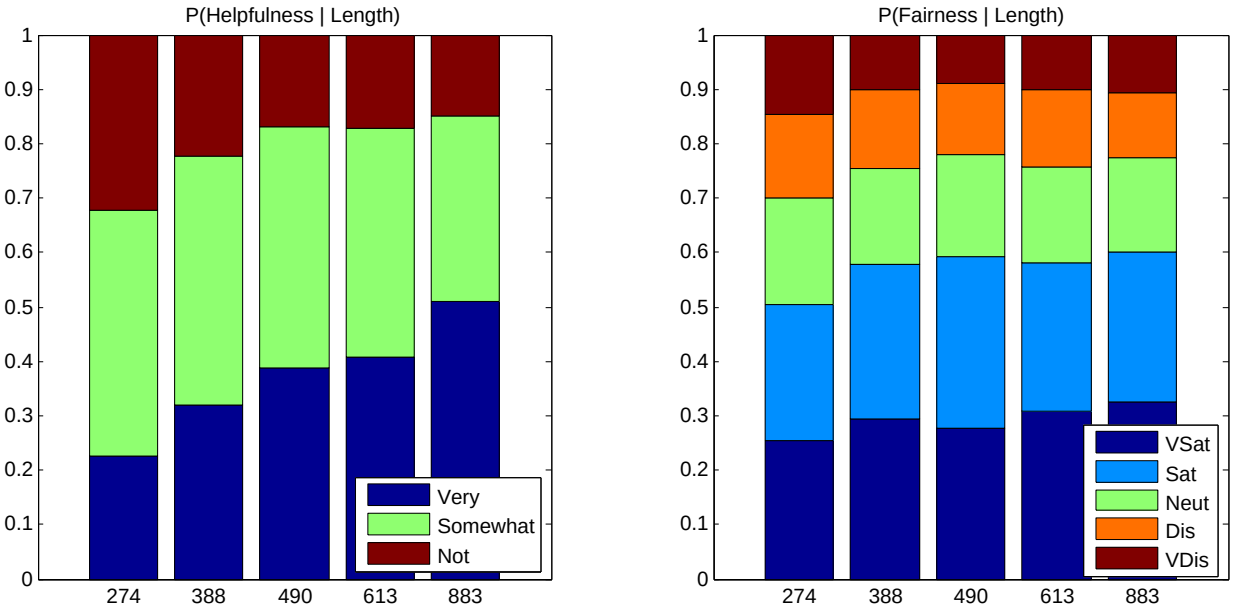

Figure 6. **Review Length vs. Fairness/Helpfulness.** For each quintile of review length (e.g., shortest 20%, next longest 20%, etc.), we show the fraction of each helpfulness and fairness rating. Longer reviews are likely to be considered more helpful. But, unless the reviews are very short, review length does not impact perceived fairness. The labels on the x-axis show the median review length for each quintile (approximate number of words calculated by number of characters minus characters in form, divided by 6).

ness of a reviewer as

$$score_h = \frac{1}{n_r + n_p}(0.5n_p + \sum_i P(helpfulness < h_i|s_i) + 0.5P(helpfulness <= h_i|s_i)) \quad (1)$$

where $score_h$ is the helpfulness score of a reviewer, s_i is the paper score for the ith review, n_r is the number of reviews, h_i is the helpfulness rating for the ith review ($h_i \in \{1, 2, 3\}$), and n_p is a prior sample. Lower numbers here mean less helpful. This rating assumes that half of the

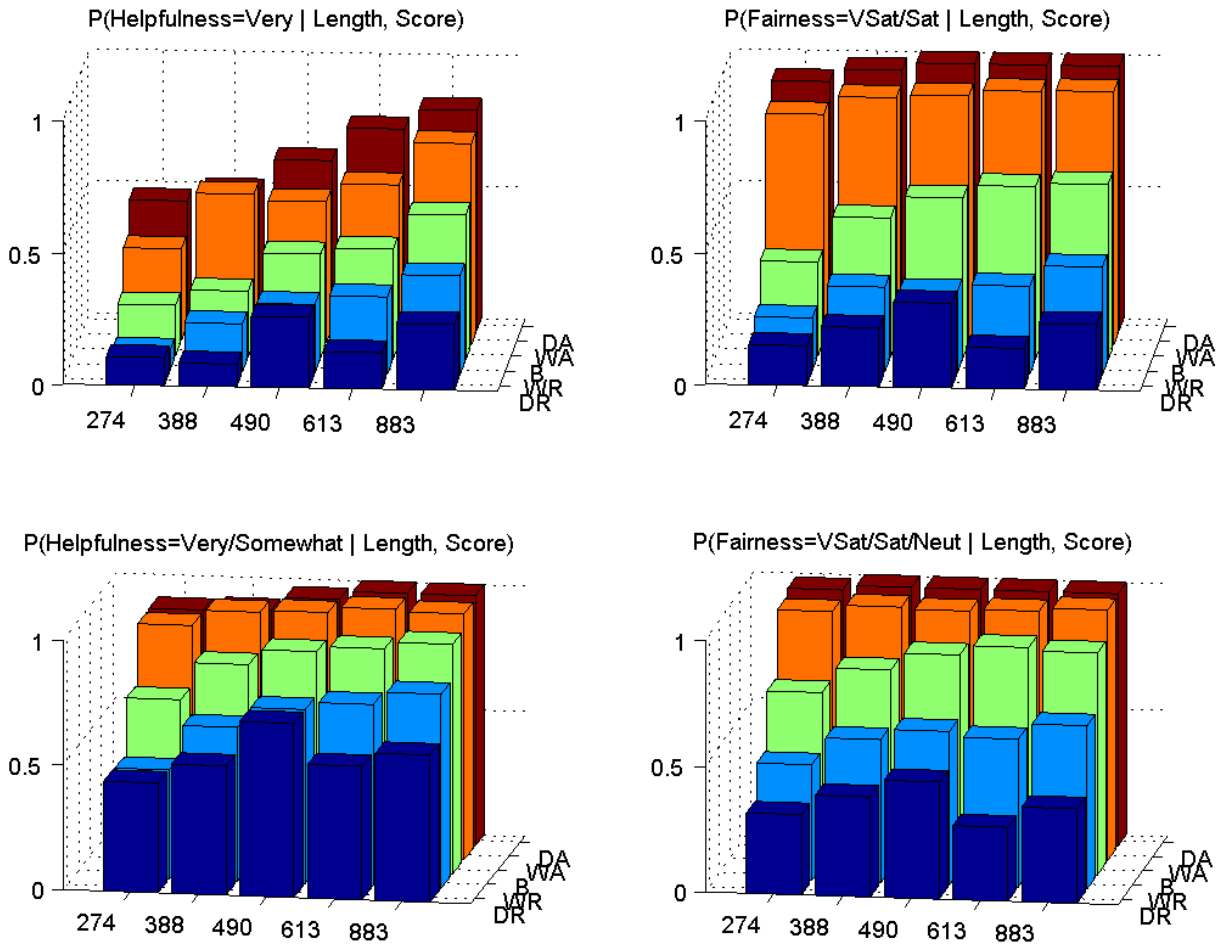

Figure 7. **Relation between Fairness/Helpfulness and Paper Score, Review Length.** The top two bar charts show the probability that a review's helpfulness or fairness rating is positive given the review's length and paper score. The bottom two charts show the probability that the review's rating is at least neutral. There is a large effect for paper score, and the helpfulness rating depends strongly on review length when conditioned on score (except for definite reject scores). The fairness does not depend strongly on review length, except that the shortest reviews are typically considered less fair, and border line reviews exhibit a stronger relation between fairness and length.

reviews with the same helpfulness rating are more helpful (and half are less helpful). Our score represents the average probability that a person's reviews are considered more helpful than another with the same paper score. To account for small sample sizes, we use a prior sample $n_p = 3$.

Similarly, our fairness ranking is

$$score_f = \frac{1}{n_r + n_p}(0.5n_p + \sum_i P(fairness < f_i|s_i) + 0.5P(fairness <= f_i|s_i)) \quad (2)$$

where the terms are analogous to the helpfulness case. Our helpfulnes and fairness rankings are not significantly correlated with number of reviews or with the positivity of recommendations.

Figure 9 shows the range of helpfulness/fairness scores for reviewers and the scatterplot of helpfulness and fairness scores. There is a large correlation (coefficient 0.68) between helpfulness and fairness.

Predictiveness. Good reviews should not just be helpful and fair but should help the area chair decide whether a paper should be accepted. We do not know which papers really should have been accepted, but we can use which papers were accepted as a substitute. To rate predictiveness, we separately compute scores for accepted and rejected papers. A review gets a score of +1 if it agrees with the decision (e.g., "definite accept" or "weak accept" for an accepted paper), 0 for a borderline rating, -1 for a "weak"

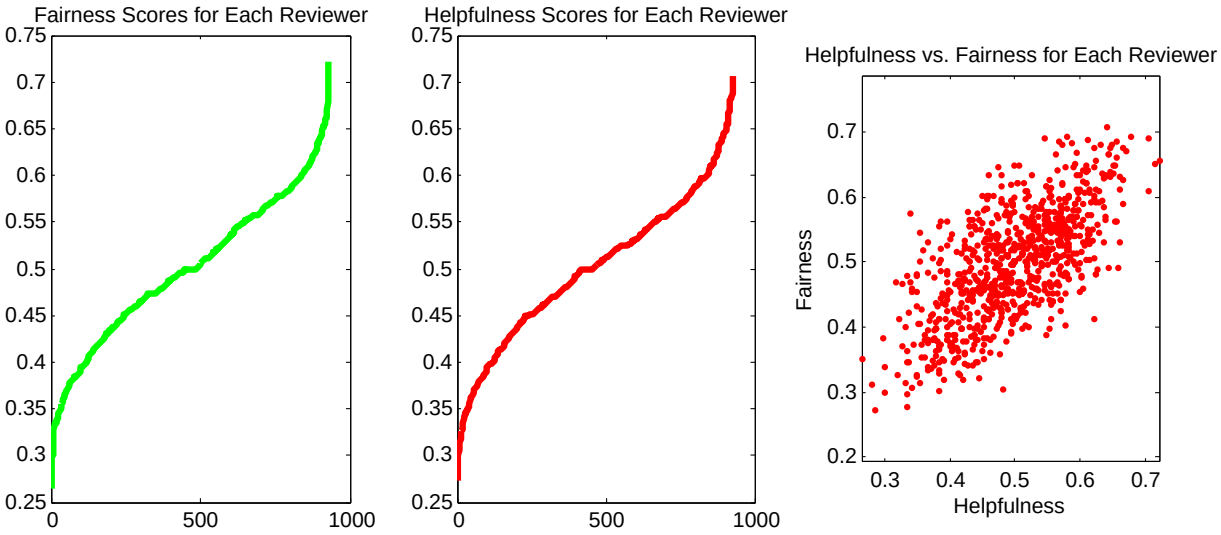

Figure 9. **Helpfulness/fairness reviewer scores.** A small number of reviewers are considered especially helpful or unhelpful, fair or unfair. Perceived helpfulness and fairness are strongly correlated. These scores are normalized to discount the effect of the paper score for each review.

rating that disagrees, and -2 for a "definite" rating that disagrees. We assign the same score to a "definite" and "weak" rating that agrees with the outcome because a "definite" rating will influence the outcome more. The scores are computed as follows

$$score_p = \frac{n_p\mu_{accept} + \sum_{i\in accept} agree\_score(s_i, d_i)}{2(n_{accept} + n_p)} + \frac{n_p\mu_{reject} + \sum_{i\in reject} agree\_score(s_i, d_i)}{(2(n_{reject} + n_p)} \quad (3)$$

where $score_p$ is the predictiveness score, n_{accept} is the number of reviews for accepted papers, $n_p = 3$ is a prior sample, $agree\_score$ is the function described above is based on the agreement between the recommendation s_i and the decision d_i. The terms μ_{accept} and μ_{reject} are computed as the average accept/reject scores when $n_p = 0$.

Our predictiveness ratings are not correlated with the number of reviews, but they are correlated with the paper scores (more negative reviewers tend to be more predictive with correlation coefficient of 0.27). The top three rated reviewers had the following recommendations/decisions: (accepted: 2 DA, 2 WA; rejected: 2 DR, 3 WR); (accepted: 1 DA, 2 WA; rejected: 2 DR, 4 WR); (accepted: 2 DA, 1 WA; rejected: 3 WR, 2 DR). The median rated reviewer had (accepted: 1 B; rejected: 1 B, 2 WR, 1 DR). The two worst rated reviewers had (accepted: 1 DA, 1 WA, 1 WR; rejected: 2 DA, 1 B) and (accepted: 1 DA, 1 WA, 2 WR; rejected: 2 DA, 2 B, 1 WR).

Figure 10 shows the range of predictiveness scores for reviewers and the scatterplot of predictiveness and helfpulness/fairness scores. There is no significant correlation between predictiveness and helpfulness/fairness.

5. Discussion

5.1. Recommendations for Authors

5.2. Recommendations for Reviewers

5.3. Recommendations for Area Chairs

5.4. Recommendations for Program Chairs

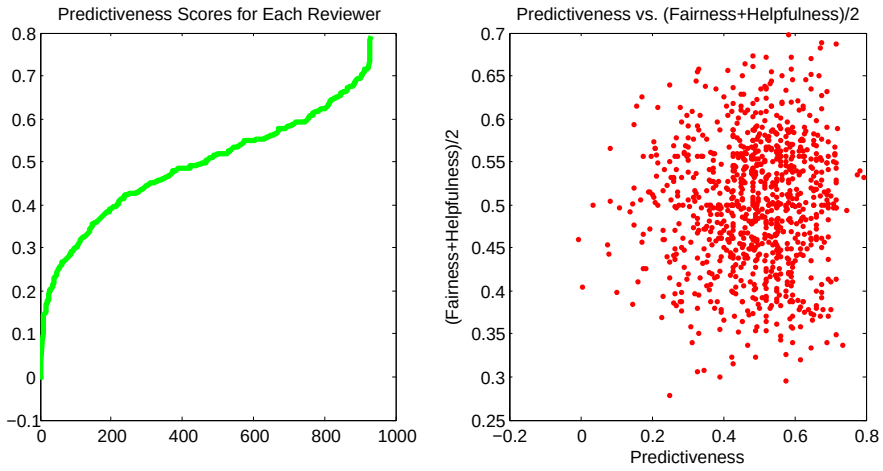

Figure 10. **Reviewer predictiveness scores.** Most reviewers' recommendations are predictive of the decision, but some are weakly predictive. Interestingly, there is no correlation between the predictiveness of a reviewer (how often their recommendation is adopted) and the fairness/helpfulness score (the quality of the review perceived by the authors, controlled for paper score).