Reviewing the Reviewers: a Study of Author Perception on Peer Reviews in Computer Science

Conny Kühne Klemens Böhm Jing Zhi Yue
Karlsruhe Institute of Technology (KIT), Germany
Email: {conny.kuehne|klemens.boehm|jing.yue}@kit.edu

***Abstract*—Peer reviewing is an important form of collaborative work that is used for quality assurance in science and in other domains like software development and knowledge management. Review ratings by authors have potential to improve the quality of peer reviews, by giving way to remuneration of good reviews. A significant problem, however, is that authors' perception is hardly neutral, but might be affected by the reviews. To gain insight into their perception of peer reviews, we have conducted a survey among the authors of papers submitted to a peer-reviewed computer science conference. One of our findings is that authors are satisfied with reviews whose comments they deem helpful, and when they feel that the reviewer has made an effort to understand the paper. Suprisingly, these results hold when controlled for the score given by the reviewer. Based on the study results, we discuss the suitability of author ratings to identify high-quality reviews. We describe a remuneration function for reviews based on author ratings that aims to neutralize the effects of review scores.**

I. Introduction

Peer reviewing is an important form of collaborative work that is used for quality assurance in science and in other domains like software development and knowledge management. Reviewing a scientific paper requires considerable intellectual effort and time. However, the incentive to write high-quality reviews tends to be somewhat low, as reviewers remain anonymous. While most reviewers do provide high-quality reviews, there is a non-negligible rate of reviews of lower quality, at least according to the perception of the authors. Personal communication with other scientists as well as numerous discussions regarding the pros and cons of peer reviewing in various scientific communities show this [1], [2].

We believe that feedback given by authors has potential to improve the review process. More specifically, we deem it promising to rely on review ratings to identify high-quality reviews and remunerate reviewers[1]. However, the specifics of such a remuneration mechanism are not obvious. For instance, assuming that accept/reject decisions affect the perception of authors, simply remunerating reviewers based on the ratings they receive from authors is not objective. To illustrate, a review of a paper that has been rejected will obtain lower ratings than a review of the same quality of a paper that has been accepted, should that assumption hold.

To gain insight into authors' perception of reviews, we have conducted a study with authors who had submitted papers to a peer-reviewed computer science conference. One important goal was to determine which criteria may be useful to identify good reviews and thus to determine an adequate basis for reviewer remuneration.[2] To this end, we have incorporated review ratings into the review process. Authors could assess each review they had received according to a broad selection of criteria, such as helpfulness of review comments. We have also asked them to rate the review scores they had received. For the sake of clarity, we refer to values used for assessment as *scores* when issued by reviewers and as *ratings* when issued by authors.

A promising way to improve the accuracy of scores are mechanisms for honest feedback known from the economic literature [3], [4]. These mechanisms reward truthfulness in scenarios where no objective truth criterion is available. Applied to our scenario, they are suitable to reward reviewers, contingent on how other reviewers have assessed the same submission. The mechanisms work based on the assumption that different opinions induce different estimates of the distribution of this opinion among others. A related objective of our study is to test whether this assumption holds with regard to author ratings. Our results are applicable to reviewers as well, for reasons discussed in Section II.

Having said this, our contributions are as follows:
Extensive Analysis. We have carried out a detailed analysis of author perception of peer reviews. Among others, our analysis addresses the following questions: How is author satisfaction with review quality distributed? How strongly do the characteristics of the review, in particular the review scores, as well as the accept/reject decision, affect author ratings? Which of the different assessments of the reviewer influence author perception of overall review quality?
Test of Validity of Assumptions behind Mechanisms for Honest Feedback. Mechanisms for honest feedback are promising to reward reviewers based on the assessments of other reviewers of the same submission, as explained earlier. These mechanisms rely on certain assumptions. We test whether a particularly crucial and important one holds in our setting.
Discussion of the Suitability of Author Ratings as a Basis for Review Remuneration. To our knowledge, we are the first to study how to remunerate peer reviewers based on author

[1]The form of the remuneration is not a topic of this paper. One possibility is to remunerate reviewers with specific awards, e.g., 'best reviewer award', as some conferences have done already.

[2]Based on our study, one might be able to derive other measures as well, e.g., re-design of review forms, or other measures which we have not come up with at this current point of time. In this article, we keep the discussion focused on reviewer remuneration as the core objective.

ratings. Given our results, we discuss how a suitable metric to remunerate reviewers could look like. This metric should neutralize possible effects of the review process, e.g., the effects of the accept/reject decision, on author ratings as much as possible.

Paper outline: We discuss mechanisms for honest feedback in Section II. Section III reviews related work. Section IV presents the questionnaire used for the study, its implementation and the statistical methods we use for the analysis. Sections V and VI present the results of the analysis. Section VII studies the suitability of ratings as a basis for remuneration of reviewers. Section VIII concludes.

II. Mechanisms for Honest Feedback

Honest feedback mechanisms reward truthful feedback in the absence of an objective truth criterion. Possible application scenarios include online product ratings ('How do you assess Product x?'), polls of expert judgments ('How likely do you deem global warming to occur?'), or psychological surveys ('Do you prefer red or white wine?'). In these scenarios, explicit rewards can improve the quality of responses by stimulating a respondent to take the time to respond accurately and truthfully. However, rewards are difficult to determine, because the objective truth is not available. This may be the case because the questions are inherently subjective, or because the truthfulness of a response can only be established at a much later point in time. And simple rewards, for example a remuneration based on the majority opinion, are unlikely to yield the desired results.

Honest feedback mechanisms solve this problem by rewarding answers depending on the answers made by peers. They compute rewards in such a way that honesty, not conformity, is the optimal strategy for respondents. They achieve this by exploiting correlations between opinions of different persons regarding the same question. The existing mechanisms differ in the computation of the rewards. [4] rewards a rating by comparing it to the rating of another randomly chosen rater called the reference rater. The rating is rewarded by comparing the likelihood assigned to the reference rater's possible ratings to his actual rating. [3] rewards answers that are "more common than collectively predicted". Truthful responses maximize the expected reward, given that all other participants answer truthfully as well.

The crucial assumption behind all these mechanisms is that respondents use their own opinion as information on the popularity of this opinion among others. More precisely, respondents who endorse a certain opinion deem it more popular than those who do not. For example, a red wine lover tends to estimate the ratio of people who prefer red over white wine higher than average. Various studies have confirmed this proposition, see [5] for an overview. A common explanation is that respondents use their own opinion as evidence to update (a hypothetic) common prior distribution. This is called the common prior assumption.[3]

It seems promising to apply honest feedback mechanisms to peer reviewing. By rewarding reviewer honesty based on scores of other reviewers for the same submissions, reviewers would have incentives to give accurate scores. As a prerequisite, we test whether authors who responded to our questionnaire act in line with the common prior assumption. More precisely, we test whether authors act accordingly to Bayesian theory when estimating the ratios of unfavorable ratings given by other authors. Since we envision applying honest feedback mechanisms to reviewers, it would be necessary to test the validity of the assumption among reviewers, not among authors. However, we assume that our results are generalizable to reviewers, for two reasons. The first one is that the group of reviewers and the one of authors overlap to a large degree, i.e., many authors are reviewers in (other) conferences. The second reason is that assigning ratings to reviews is very similar to assigning scores to papers.

III. Related Work

Criticism of peer reviewing has concentrated mainly on its efficacy and effectiveness. Some studies [7]–[9] have surveyed authors who had submitted manuscripts to journals. However, the results from the surveys differ from each other. Gibson et al. report on an online survey of 445 authors of research manuscripts submitted to the *Obstetrics and Gynecology* journal [8]. Authors were asked to rate six aspects of editorial comments and three aspects of the review process. One result is that authors of accepted manuscripts give higher ratings for overall satisfaction than authors of rejected manuscripts. Garfunkel et al. find a weaker correlation between author ratings and manuscript fate [9]. Gibson argues that the difference results from the number of survey items and the rating scales in the questions.

We see many exogenous factors which might influence author satisfaction with peer reviewing, for instance the organization of the review process, the selection of reviewers and the design of the review forms. In addition, the review process of a conference is different from the one of a journal. [10] has pointed out that, at least for experimentalists, conference publication is preferred to journal publication, and the premier conferences tend to be more selective than the premier journals. Hence, many conferences have huge numbers of submissions and tight time constraints. Publication in conferences needs shorter time to print (7 months vs. 1-2 years). However, there is a lack of studies on conference reviews.

There also are various proposals to increase review quality. Some proposals attempt to improve the review process itself, like allowing authors to submit feedback in the rebuttal phase or supporting a rather open review process instead of double blind. In the journal *Biology Direct* [11], to give an example, authors can select their reviewers from the editorial board, and

[3]The psychological literature has initially regarded this phenomenon as an egocentric error of judgment (a 'false consensus'). Dawes offered a Bayesian explanation [6].

reviews are not only signed, but also published together with author responses as part of each article. Analyses of different modes of peer-review activities, e.g., online vs. face-to-face reviewing [12], exist as well.

Others have proposed to train reviewers. *The British Medical Journal*, offers reviewers a workshop which gives them clear briefs, including guidance on what to include in the review etc. [13]. Callaham et al. try to improve reviewing skills by means of feedback from the editorial board. In their study editors write short feedback in text to the reviewers to comment on the quality of the reviews submitted [1]. However, the performance of reviewers is hardly improved, i.e., simple written feedback to reviewers seems to be inefficient as an educational means in this specific context. Another study finds that reviewer ratings given by journal editors are moderately reliable, and that they correlate modestly with the ability of reviewers to find flaws in a test manuscript [14].

Peer reviewing not only is an important instrument in the scientific community to pick good contributions, but also finds its usage in other disciplines. In software-engineering processes, to give an example, peer reviews are used to detect deficiencies in the code [15], [16]. Other studies investigate the effect of peer reviewing on student learning. In [12], students review papers written by their peers, and the results indicate that students take peer reviews seriously and provide constructive reviews. Finally, peer ratings have also been proposed in the context of the collaborative creation of structured knowledge. For example, Noy et al. discuss ratings for the evaluation of ontologies [17]. Hütter et al. evaluate ratings and rating based incentive mechanisms for the collaborative construction of structured knowledge empirically [18].

IV. Materials and Methods

We have carried out our survey by means of an online questionnaire. Survey participants were the authors of the CASES 2009 conference. In this section, we first describe details of the conference and its peer-review process which are relevant to our study. Then we describe the questionnaire and the implementation of the study. Finally, we review the statistical methods we use in our analysis.

A. Conference and Peer-Review Process

We invited the authors of the *CASES 2009 Conference for Emerging Technology in Embedded Computing Systems* to participate in our study. The conference is held annually and focuses on compilers and architectures for embedded systems [19]. Authors submitted 72 papers to the conference overall. 48 reviewers wrote 311 reviews on the submissions in total. The number of reviewers per submission ranged from 2 to 6 (avg.=4.38 reviewers/submission). The reviewers did not know about our study beforehand. Out of the 72 submissions, the conference rejected 47 and accepted 23 as full papers and 2 as short papers.

A review contained one *Overall Score*. Further, the reviewers had to assign the following detail scores: *Originality*, *Technical Contribution*, *Experimental Results*, *Description of Related Work*, and *Language and Clarity*. Additionally, reviewers provided a numerical self-assessment of their own expertise regarding the topic of the submission. Scores and self-assessment were based on the usual 1-5 scale, with 1 being the minimum and 5 the maximum score. Furthermore, reviews could provide written comments. The conference chairs based their accept/reject decisions mainly on the *Overall Score*. However, they revised some of the ranking based decisions during a one day physical meeting.

B. Questionnaire

The questionnaire consisted of two parts. The first one contained questions concerning each individual review the respective submission had received. The second part contained general questions.

Review Specific Ratings. Regarding individual reviews, the following issues were part of our questionnaire. The *Overall Quality* rating is supposed to summarize the overall satisfaction of the author with the review. Further, we elicited ratings regarding the appropriateness of the 6 scores. We also let the authors rate the expertise level of the reviewer on the same scale as the reviewers' self-assessment. Additionally, we asked questions referring to criteria which might influence review quality: helpfulness of the review comments for future work, appropriateness of review length, perceived effort of the reviewer to understand the paper, percentage of justified comments.

General Questions. To test whether authors act in line with the common prior assumption, we let them estimate the ratio of reviews rated 'very low' or 'low' among i) all authors, ii) authors whose submissions had been accepted, and iii) authors whose submissions had been rejected. Finally, we asked authors whether they deem ratings likely to improve review quality.

The response formats were mostly ordinal and differed depending on the question. Some questions elicited interval-level data. Table I gives an overview of the review-specific ratings. See [20] for an online version of the questionnaire.

C. Implementation of the Survey

We sent out invitations to participate in the survey immediately after the notifications. We invited the contact author, i.e., one author per submission. We did not invite multiple authors per submission to avoid that authors distort results by answering questionnaires for their co-authors. Moreover, we assume the opinions of co-authors to be highly correlated. We set up the questionnaire software so that the number of questionnaire items matched the numbers of reviews a submission had received. Authors had ten days to complete the questionnaire. We sent out one reminder eight days after the invitation. As an incentive to participate in the study, besides that of helping the scientific community, we raffled off six Amazon gift certificates of USD 20,– among all survey participants. We had announced the raffle in the invitation to the survey.

TABLE I
REVIEW SPECIFIC RATINGS AND RESPONSE FORMATS

Survey Rating for	# Choices	Choices
Overall Quality	5	'very low' to 'very high'
Perc. of Justified Comments	5	0%, 25%, .., 100%
Helpfulness for Future Work	4	'not at all' to 'very helpful'
Perceived Expertise of Reviewer	5	1-5
Effort of Reviewer	3	'low', 'average', 'high'
Appropriateness of Review Length	3	'too short', 'appropriate', 'too long'
Appropriateness of (each of the 6) Review Scores	3	'too low', 'appropriate', 'too high'

D. Statistical Methods

To quantify the effects the different variables, such as the characteristics of the reviews, the ratings, the accept/reject decision, etc., have on each other, we perform a correlation analysis. Because most of the variables are ordinal in nature, we use Spearman's rank correlation coefficient ρ to calculate correlations between two variables. In line with [21], we obtain ρ by applying Pearson's product-moment correlation coefficient to ranked data as follows.

The bi-variate rank correlation ρ of a series of n observations of the variables X and Y, written as x_i and y_i, where $i = 1, 2, .., n$, is calculated as

$$\rho = \frac{\sum_i ((R_i - \overline{R})(S_i - \overline{S}))}{\sqrt{\sum_i (R_i - \overline{R})^2 \sum_i (S_i - \overline{S})^2}}$$

where R_i is the rank of x_i, S_i is the rank of y_i, $\overline{R}$ is the mean of the R_i values, and $\overline{S}$ is the mean of the S_i values. In situations where observations are tied, the average rank is assigned.

The significance level is calculated assuming that, under the null hypothesis,

$$t = (n-2)^{1/2} \left(\frac{\rho^2}{1-\rho^2} \right)^{1/2}$$

is coming from a **t** distribution with $(n-2)$ degrees of freedom, where ρ is the Spearman correlation of the observations. In line with the common practice we refer to effects that have a significance level of $p \leq .05$ as (statistically) *significant*. Note that statistical significance does *not* refer to the size of the effect in question or its practical relevance. E.g., a weak correlation can still be statistically significant.

In some situations we are interested in removing the effect of a third variable on the correlation between two variables. To control for the effects of the third variable, we use partial correlation. We obtain the partial correlation between variables X and Y controlled for the effects of a variable Z by the following formula

$$\rho_{xy.z} = \frac{\rho_{xy} - \rho_{xz}\rho_{yz}}{\sqrt{(1-\rho_{xz}^2)(1-\rho_{yz}^2)}}$$

where ρ_{xy}, ρ_{xz}, and ρ_{yz} are the appropriate correlations.

We use Pearson's χ^2 test to compare differences in ratings and response rates between accepted and rejected submissions.

V. RESULTS

In the following we present the results of our statistical analyses. To begin with, we present the response rate and an overview of the author ratings dealing directly with review satisfaction. Then we analyze the effects of review characteristics on ratings. Finally, we examine whether author estimates on rating distributions are in line with the common prior assumption.

A. Response Rate

39 out of 72 authors of distinct papers we invited completed the questionnaire, resulting in an overall response rate of .54. Authors of accepted papers were significantly more likely to complete the survey than authors of rejected papers (odds ratio 8.46, $\chi^2(1) = 13.730$, $p < .001$). Nevertheless, 46% of the respondents were authors of rejected papers. [8] reports similar response rates. Overall, the authors assessed 175 reviews. The average number of assessed reviews per participating author is 4.49.

B. Distribution of Review Satisfaction among Authors

Figure 1 is an overview of the distributions of the 4 author ratings related to review quality, categorized by accepted and rejected submissions. The percentages are relative to the respective category. Authors find 39% of reviews to be of high or very high quality and deem 45% of review comments helpful or very helpful. Further, they think reviewers made a high effort to understand their paper with 34% of the reviews and deem 71% of the review comments to have an appropriate length. The mean value of the rating *Percentage of justified comments* is 63.67 (std. deviation 17.67). These findings suggest that authors are quite satisfied regarding the quality of their reviews. Nevertheless, there seems to be room for improvement, as authors rate 22% of reviews to be of low or very low quality and 15% of the reviews as being not helpful at all.

C. Influence of the Overall Score on Quality Ratings

Table II shows the dependency of the ratings concerning review quality on the *Overall Score*. All ratings show a statistically significant positive correlation with the *Overall Score*. In other words, authors tend to assign higher ratings to reviews that assign high scores. But the correlations are not perfect and vary between rating categories. The *Overall Quality* rating shows the highest correlation, helpfulness has the lowest one.

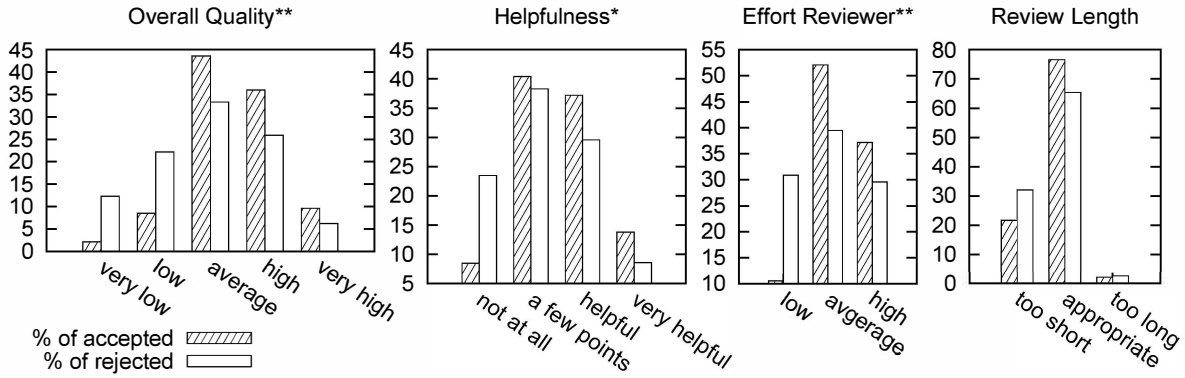

Fig. 1. Distributions of author ratings for review quality. Differences between ratings for accepted (filled) and rejected submissions are statistically significant (* $p < 0.05$, ** $p < 0.01$) except for ratings of review length.

TABLE II
CORRELATIONS OF QUALITY RATINGS WITH THE *Overall Score*
(** $p < .01$)

Rating	Correlation with *Overall Score*
Overall Quality	.591**
Justified Comments	.506**
Helpfulness	.360**
Expertise Reviewer	.417**
Effort Reviewer	.482**

D. Which Ratings Do Explain the Overall Quality?

The *Overall Quality* rating is the overall assessment of the review by the author. By comparing it to the other ratings, we can determine which criteria have the highest influence on review quality from the authors' perspective. However, Subsection V-C has shown that the *Overall Score* affects ratings. To remove this effect, we computed the partial correlations while controlling for the *Overall Score*. Figure 2 shows the results of both the bi-variate and the partial correlations. The light bars show the correlation between the respective rating of authors and their *Overall Quality* rating. The dark bars show the same correlation when controlled for the effect of the *Overall Score*. The difference between the respective bars shows how big this effect is. For instance, the difference for the helpfulness rating is relatively small. This means that the correlation of *Helpfulness* with *Overall Quality* is rather independent of the *Overall Score*. In contrast, the *Overall Score* strongly influences the correlation of the rating for *Technical Contribution* with *Overall Quality*. All ratings correlate significantly with the rating for *Overall Quality*. Ratings for *Effort of Reviewer*, *Helpfulness*, and *Expertise of Reviewer* show the highest correlation with perceived review quality – both in the bi-variate case and when controlled for *Overall Score*.

E. Influence of Acceptance Status on Ratings

Authors of rejected submissions assign lower mean ratings than those of accepted ones. This effect is statistically significant. Because acceptance is on the submission level, we computed averages of the review-specific ratings per submission to test for correlation with acceptance. The correlations of acceptance with the respective mean ratings per submission range from .06 to .253. In particular, the correlation of acceptance with the mean values of *Overall Quality* and effort of reviewer is $\rho = .236$ and $\rho = .182$, respectively. Thus, the effect of accept/reject decisions on author ratings is weaker than the effect of review scores.

This finding was unexpected to some degree, at least to us, as we had anticipated a stronger effect. However, in retrospect it is explainable by the following facts. In our study, authors rated individual reviews. Thus, they could differentiate between reviews that assigned scores in their favor and those that did not. Since reviews per submission vary in their scores, and authors apparently take this into account, acceptance has a weaker effect on ratings than the scores.

F. Influence of Review Length

Minimum, maximum, and mean length of review comments in characters were 0, 11604, and 1488 respectively (standard deviation=1213, median=1258). Review comments are very rarely perceived as too long. But in over one fourth of the cases, authors perceive them as too short (see Figure 1). The length of a review is positively correlated with its respective rating ($\rho = .501$, $p < 0.01$). The partial correlation controlled for *Overall Score* is slightly less ($\rho = 0.433$). Thus, authors appear to prefer longer reviews.

G. Expertise of Reviewer – Self-Assessed vs. Perceived

Authors rated the reviewer expertise on the same scale as the reviewer. The self-assessment and the assessment by the author are moderately correlated ($\rho = .360$, $p < .001$). This is the *only* non-negligible correlation of the self-assessed expertise with all other variables we analyzed. In particular, we do not

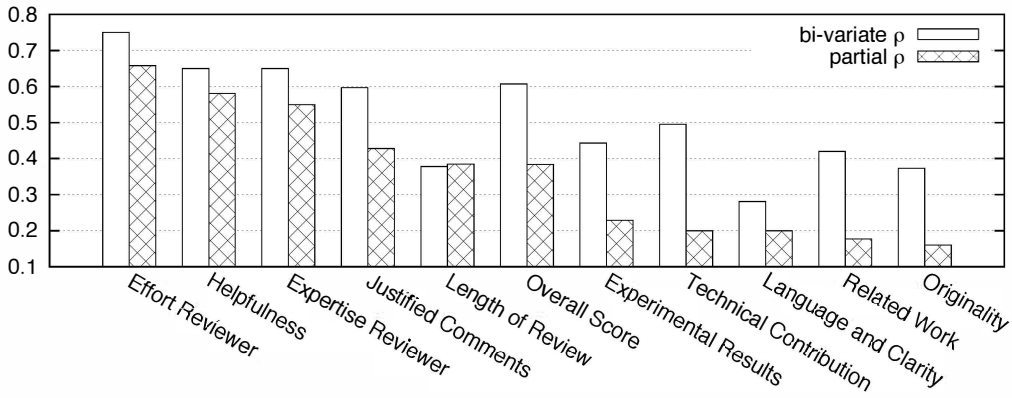

Fig. 2. Correlation of ratings with *Overall Quality* – bi-variate and controlled for *Overall Score*

TABLE III
CORRELATION OF REVIEW SCORES AND THEIR RESPECTIVE RATINGS
(** $p < .01$)

Score	Correlation with Respective Rating
Overall Quality	.612**
Originality	.596**
Technical Contribution	.672**
Experimental Results	.620**
Related Work	.568**
Language and Clarity	.655**

find any correlation with ratings for *Review Quality*, *Helpfulness*, and *Justified Comments*. On the other hand, perceived expertise is significantly ($p < .01$) partially correlated (controlled for the *Overall Score*) with *Effort of Reviewer* ($\rho = .571$), *Helpfulness* ($\rho = .523$), and *Justified Comments* ($\rho = .434$). Like other ratings, *Expertise of Reviewer* moderately depends on the *Overall Score* ($\rho = .417$, $p < .001$).

H. Rating of Review Scores

Authors rate the six scores their submissions have received per review mostly as adequate. The number of ratings per score with value 'adequate' ranges from 66% to 77% for the respective scores. Authors almost never perceive their scores as too high. Out of 175 ratings for *Overall Quality*, 4 had the value 'too high'. All 4 were assigned by different authors. Further, 18 of the 875 ratings on the five detail scores had the value 'too high', 8 of which were assigned in category *Language and Clarity*. Review scores are significantly positively correlated with their respective ratings (see Table III). This means that authors tend to rate high scores as adequate and low scores as too low. But considering that authors have rated scores directly, the correlations are lower than we had expected.

I. Authors' Estimations of Rating Ratios

To test the common prior assumption, we asked authors to estimate the ratios of reviews rated unfavorable ('very low' or 'low') among i) all authors, ii) authors whose submissions had been accepted, and iii) authors whose submissions had been rejected. Authors' mean estimates for the three ratios are higher than the mean values observed (Table IV). Furthermore, all three estimates have a relatively high standard deviation. However, the overall tendency of the estimated ratios is the same as in the ratios observed, i.e., accepted submissions yield less unfavorable ratings on average than all ratings combined, and all ratings combined yield less unfavorable ratings on average than rejected submissions.

TABLE IV
OBSERVED AND ESTIMATED VALUES OF UNFAVORABLE RATINGS FOR *Overall Quality*

	Observed	Estimated	
	Mean	Mean	Std. Deviation
Accepted	.106	.238	.165
Rejected	.346	.474	.174
All	.217	.354	.176

More importantly, regarding Bayesian updating, there is a statistically significant effect of an author's own *Overall Quality* ratings on his estimations regarding the *Overall Quality* ratings issued by other authors. We obtained this result by calculating the share of unfavorable *Overall Quality* ratings issued by an author and comparing it to his estimates. The respective correlations of this share with the three estimates are significant and range from .374 to .422 ($p < .05$). Put simply, the more unfavorable ratings an author issues, the more he expects others to do the same. This suggests that authors do indeed behave like Bayesian learners who use their own opinion to update a (common) prior.

VI. DISCUSSION

The analysis confirms our expectations regarding this variant of collaborative work to a large extent. Authors' assessments of reviews are biased. They depend on review scores (on overall as well as on detail scores), but only weakly on the acceptance status. We think that this is because the granularity of assessment was the review, not the submission. I.e., authors are not very much affected by a rejection per se, but differentiate between reviews in their favor/not in their favor.

The correlations of ratings with review scores are relatively moderate. We expected them to be stronger. In so far, authors appear to be 'decently honest'. Some ratings are relatively 'neutral' regarding the review scores. These ratings are the perceived effort of reviewer, the percentage of justified comments, and the helpfulness of the comments. They are, compared to the other ratings, relatively weakly influenced by the *Overall Score* of the review. Moreover, their respective correlations with *Overall Quality* hold when controlled for the *Overall Score*.

We are surprised to find that the reviewer's self-assessed expertise is not correlated with any of the ratings except for one: the assessment of the reviewer expertise by the author. Therefore, we speculate that the display of the self-assessment itself affects the opinion of the author. This is akin to the so called "Seeing is believing effect" discussed in [22]. To examine this issue further, future experiments could have two groups of authors, and the self-assessed expertise level could be displayed to one group only. Comparing the results of both groups would yield insights as to whether this is indeed the case. Next, in our study we provided the authors with 3 choices to assess review scores: 'too low', 'adequate', and 'too high'. We did this mainly to find out how many authors would choose 'too high'. For a real rating system, these choices appear to be rather inadequate, as we have learned from our study. The number of ratings being 'too high' is negligible, resulting effectively in a boolean rating scale. More importantly, with the exception of *Language and Clarity*, ratings for scores are relatively strongly affected by their respective score. As expected, they are not useful as a quality measure. Objective criteria to identify and remunerate high-quality reviews are difficult to find. In the end, quality and helpfulness can only be perceived and assessed by authors. Other parties that are assumed to be objective in their assessment are rather unsuitable [1] to increase the quality. On the other hand, authors are influenced by reviews. So their assessment is not objective either. How much of this influence is due to the scores and how much is due to the written comments is hard to determine. To examine this, a future experiment would have to introduce an experimental group of authors who only see review comments and do not see the scores. However, it is difficult to impossible in practice to split the group of authors into two groups which are then treated differently. One could, however, try to eliminate the influence of the scores on quality ratings. How this could be achieved is the topic of the next section.

VII. Remuneration for Reviews

One important objective of ours behind this study was to identify criteria that might be suitable to reward high-quality reviews. The main question in this context is: How to decouple incentives to write high-quality reviews from incentives to give accurate scores? We have shown that there is a positive correlation between review scores and ratings by authors. That is, authors like reviews that like their submissions. Thus, if one simply remunerated reviews based on how highly authors rate them, it would create incentives for reviewers to give inaccurately high scores. Consequently, we propose to remunerate *relatively* highly rated reviews, i.e., reviews that receive high ratings by authors despite assigning low scores. In the following, we formalize one possible function that achieves this. We explicitly write down this function for illustration purposes, and to indicate a potential direction of future research.

TABLE V
T APPLIED TO THE DATA OF OUR STUDY

T	Reviews receiving T
-2	1.1%
-1	24%
0	44.6%
1	24.6%
2	5.7%

Let $R \in \{1, .., k\}$ denote the value of the author rating of a given review. Let $S \in \{1, .., l\}$ denote that review's score. The remuneration function $T(R, S) = R - S$ removes the influence of the score on the remuneration. T can be further refined. For example, reviewers might be deterred from reviewing if threatened by penalties. So one could only remunerate good reviews, but refrain from any penalization. Further, one could normalize the rating scales if $k \neq l$.

In order to see whether our proposed remuneration indeed neutralizes the effects of scores, we apply T to the data of our study. Let R be the *Overall Quality* rating by an author and S the *Overall Score* of the respective review, and set $k, l = 5$ according to the number of different choices for ratings and scores in our study. Table V shows the results. The remuneration is quite symmetrically distributed. 44.6% of reviews would not be remunerated at all. Further, T is positively correlated with *Overall Quality* ($\rho = .526$) and weakly negatively correlated with *Overall Score* ($\rho = -.333$).

A further decoupling of the incentives from scores could be achieved by choosing a rating category for R that is only weakly dependent on S. One candidate is, for example, the helpfulness of the comments for future work, because, of all review ratings, its dependency on the *Overall Score* is the weakest one. To demonstrate this, we use a normalized variant of the remuneration function above and apply it to the data of our study: Let $R_{help} \in \{1, .., 4\}$ be the rating for the helpfulness of a given review and S be that review's *Overall Score*. The resulting remuneration for helpfulness

$$T_{help}(R_{help}, S) = \frac{R_{help}}{4} - \frac{S}{5}$$

is only negligibly dependent on the *Overall Score* ($\rho = -.126$, $p = .096$), while still being strongly correlated with helpfulness ($\rho = .761$, $p < .01$) and *Overall Quality* ($\rho = .591$). Thus, it decouples the incentive to give accurate scores and the incentive to write high-quality reviews to a large degree.

One problem that might occur with the remuneration functions above is that, all else being equal, reviewers could increase their chance of being remunerated by assigning lower scores. In the worst case, all reviewers would assign

minimum scores while still trying to write helpful comments. Clearly, this is undesirable. To counter artificially low scores, conferences could use mechanisms for honest feedback. In this case, some of the remuneration for a review would be based on its score in comparison to the scores of other reviews for the same submission. Reviewers would then face a trade-off between two factors: Some of the remuneration would be based on author ratings, some based on review scores. Studying the question how this trade-off influences reviewer behavior is beyond the scope of this paper, for several reasons. The specifics of the remuneration function, in particular the proposal how it might depend on review scores, are a result of our study. We did not foresee them prior to the study and hence had not incorporated them in the questionnaire. Next, the focus of our study is author perception. Discussing the behavior of reviewers based on our results would be highly speculative. Finally, for future work we deem experiments the most promising way to study reviewer behavior in presence of the trade-off described above. I.e., we would let reviewers know the remuneration function(s) and measure how this affects their behavior.

VIII. Conclusions

Selecting conference articles is an important instance of collaborative work. Today, this is typically done by means of peer reviewing. Review ratings by authors have potential to improve the quality of peer reviews. A significant problem however is that authors' perception is hardly neutral, but might in turn be affected by the reviews. To gain empirical insight into authors' perception of reviews, we have conducted a study with 39 authors of a computer science conference who rated 175 reviews they had received. The results of this study show that authors' satisfaction with review quality is good, but has some room for improvement. Review scores affect author ratings to different degrees. Authors rate reviews as good if they deem the review helpful for their future work, deem the review comments justified, and have the impression that the reviewer made an effort to understand the paper. By and large, these results hold when controlled for the overall score. Acceptance and self-assessed reviewer expertise only have a weak influence on perceived review quality. Finally, the common prior assumption, which is crucial for honest feedback mechanisms, holds with respect to authors. Given these results of the study, we have discussed suitable metrics to compute remunerations for reviews based on ratings and scores. Applied to the data collected in our study, they neutralizes the effects of scores to a large degree.

Acknowledgment

The authors would like to thank Jörg Henkel, chair of the CASES 2009 conference. Furthermore, we thank Björn Thier for his technical help in setting up the survey software.

References

[1] M. L. Callaham, R. K. Knopp, and E. J. Gallagher, "Effect of Written Feedback by Editors on Quality of Reviews: Two Randomized Trials," *JAMA*, vol. 287, no. 21, 2002.

[2] "Nature's peer review debate," http://www.nature.com/nature/peerreview/debate.

[3] D. Prelec, "A Bayesian truth serum for subjective data," *Science*, vol. 306, no. 5695, 2004.

[4] N. Miller, P. Resnick, and R. Zeckhauser, "Eliciting informative feedback: The peer-prediction method," *Manage. Sci.*, vol. 51, no. 9, 2005.

[5] G. Marks and N. Miller, "Ten years of research on the false-consensus effect: An empirical and theoretical review," *Psychological Bulletin*, no. 102, 1987.

[6] R. M. Dawes, "Statistical criteria for establishing a truly false consensus effect." *J. Exp. Soc. Psychol*, no. 25, 1989.

[7] E. J. Weber, P. P. Katz, J. F. Waeckerle, and et al., "Author Perception of Peer Review: Impact of Review Quality and Acceptance on Satisfaction," *JAMA*, vol. 287, no. 21, 2002.

[8] M. Gibson *et al.*, "Author perception of peer review." *Obstetrics and gynecology*, vol. 112, no. 3, September 2008.

[9] J. Garfunkel *et al.*, "Effect of acceptance or rejection on the author's evaluation of peer review of medical manuscripts," *JAMA*, vol. 263, no. 10, 1990.

[10] Computer Science and Telecom. Board, *Academic Careers for Experimental Computer Scientists and Engineers*. Washington, D.C.: National Academy Press, 1994.

[11] "Can 'open peer review' work for biologist? biology direct is hopeful." http://www.nature.com/nature/peerreview/debate/op1.html.

[12] C. Bauer *et al.*, "The student view on online peer reviews," in *ITiCSE'09*. Paris, France: ACM, 2009.

[13] "How can we get the best out of peer review? a recipe for good peer review," http://www.nature.com/nature/peerreview/debate/nature04995.html.

[14] M. L. Callaham, W. G. Baxt, J. F. Waeckerle, and R. L. Wears, "Reliability of Editors' Subjective Quality Ratings of Peer Reviews of Manuscripts," *JAMA*, vol. 280, no. 3, 1998.

[15] D. Galin, *Software Quality Assurance: From Theory to Implementation*. Harlow, UK: Pearson Education, 2004.

[16] K. E. Wiegers, *Peer Reviews in Software: A Practical Guide*. Boston, MA.: Addison-Wesley, 2002.

[17] N. F. Noy, R. V. Guha, and M. A. Musen, "User ratings of ontologies: Who will rate the raters?" in *Proceedings of the AAAI 2005 Spring Symposium on Knowledge Collection from Volunteer Contributors*, 2005.

[18] C. Hütter, C. Kühne, and K. Böhm, "Peer production of structured knowledge -: an empirical study of ratings and incentive mechanisms," in *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*. New York, NY, USA: ACM, 2008, pp. 827–842.

[19] "CASES 2009 - international conference on compilers, architecture, and synthesis for embedded systems," 2009.

[20] http://www.ipd.uni-karlsruhe.de/~ckuehne/cases_survey.html.

[21] A. W. Jerome L. Myers, *Research design and statistical analysis*. New York: Erlbaum, 2003.

[22] D. Cosley *et al.*, "Is seeing believing?: how recommender system interfaces affect users' opinions," in *CHI '03*. New York, NY, USA: ACM, 2003.