Insights and Lessons Learned from Analyzing ICSE 2014 Survey and Review Data

Lionel Briand, André van der Hoek

Introduction

This document reports on survey and review data collected during and after the ICSE 2014 review process. Its goal is to provide full insights and lessons learned about the reviewing and decision making process that led to the selection of 99 scientific articles (out of 495) for the technical research program of ICSE 2014. Such data collection and analysis is a first at ICSE and is particularly important this year as significant changes were made in the call for papers and review process.

For the first time this year, ICSE used a two-tier review committee composed of a Program Board and Program Committee, in a way similar to the RE and MODELS conferences. The details of the implementation, however, differed somewhat on a number of points, as discussed in the detailed guidelines that were made public earlier this year[1]. The tasks and prerogatives of the PB and PC members were defined in a precise way to avoid an overdue influence on the decisions by PB members, a threat that was mentioned by people involved in the RE review process.

The primary advantage of the Program Board model is that it scales. With ICSE receiving a significant increase of submitted papers, year-by-year, the traditional Program Committee model started to not be able to handle all of the submissions in a satisfactory manner. Individual PC members had to review a great many papers, and the discussion time in the PC meeting per paper shrunk significantly over the years.

We believe that the quality of reviews depends, to a large extent, on assigning a reasonable number of papers to each reviewer, within their areas of expertise. The PB model enables this through a much larger program committee. Another advantage of this model is that PB members can dedicate themselves entirely to monitoring the review process, moderating online discussions, and making final decisions on papers that could not be resolved online. With almost 500 papers, monitoring the review process and moderating online discussions is task that becomes impossible to be performed with the same level of scrutiny just by the PC chairs.

A second important difference this year is that all submitted papers were classified by the authors, according to the nature of their primary contribution. The various categories in the classification provide criteria by which different

[1] http://2014.icse-conferences.org/research

kinds of contributions should be judged, in an attempt to make the overall review process more systematic and consistent.

We also stressed to the Program Board and Program Committee that no conference paper can achieve perfection on all aspects within the given size constraints. The review guidelines sent to PB and PC members were, also for the first time at ICSE, made public[2].

As mentioned above, this report represents the first time review and survey data are collected, analyzed, and made public. As a result, one limitation of our analysis in this report is that there is no baseline of comparison. We hope, however, to start a practice that will facilitate continuous improvements to the ICSE review process based on objective data and careful analysis. We also hope that maximum transparency with the review process will result in an increased trust and involvement of the wider research community into ICSE events, as authors, reviewers, organizers, or simply attendees.

This report addresses a number of questions regarding the quality of the review and decision process. The interpretation of the results requires some degree of subjective judgment, which we tried to keep balanced and to a minimum. We also intentionally keep the analysis and diagrams presented in this report as simple as possible.

As described in our guidelines, reviews were handled by two committees, the Program Board (PB) and the Program Committee (PC). We collected survey data from the PB (20), the PC (58), and the Authors (185), in addition to the review data that were readily available in the submission management system (CyberChair). These questions covered the quality and expertise of the reviews, the effectiveness of the online discussion and PB meeting, and the defined paper categories. Below, we present our analysis. Verbatim, anonymous comments by the Program Board and Program Committee are included in the appendix.

1. Some basic data

The total number of papers that was submitted was 495. 29 were desk rejected or withdrawn before the review process began. 167 papers were rejected after 2 reviews by PC members (DD and CD papers were rejected by default, unless the PB member in charge requested a third review, for instance to compensate for lack of expertise; CC papers were included in the next round for an additional review, unless the PB member requested it not be, for instance because the reasons provided were stronger than the actual letter grade of C). An additional 41 papers were rejected after all three reviews were in. These rejections, again, were performed based on letter grades and associated reviews, with PB members speaking up if they disagreed with the default.

[2] http://2014.icse-conferences.org/research

The online discussion was meant to lead to one of three outcomes: suggest accept, suggest reject, and undecided. Undecided papers, as well as seventeen suggest accept and sixteen suggest reject papers were discussed at the PB meeting. The inclusion of these suggest accept and suggest reject papers were crucial to help the PB establish a 'bar' of which papers should be accepted and which papers should be rejected. Moreover, the PC member comments for some of these papers aligned well with papers with similar comments but opposite decisions in the online discussion, warranting these papers to be included in the deliberations. Overall, 95 papers were discussed in the PB meeting, out of which 23 were accepted, 15 were accepted by confirming the outcome of the online discussion (i.e., suggest accept), 40 were rejected, and 14 were rejected by confirming the outcome of the online discussion (i.e., suggest reject). Only for three papers, the result of the online discussion was overturned (two papers were rejected that were a suggest accept, and one paper was accepted that was a suggest reject). This last fact is important, meaning that PB members did not overstep their responsibilities by overturning large numbers of PC decisions.

Overall, 163 papers (60 accept, 103 reject) were dealt with satisfactorily through the online discussion overseen by the PB members. These papers, thus, did not need to be discussed at the PB meeting, leaving significant time for the papers that needed to be discussed: those receiving opposing reviews. This represents a significant savings. Note that for quite a few papers the online discussion was extremely thorough, with a clear outcome. For some, little discussion was needed because reviewers agreed or quickly reached consensus once understanding the respective points of view.

Outcome	Number
Accept	60
Desk reject/withdrawn	29
PB Accept	23
PB Confirm Accept	15
PB Confirm Reject	14
PB Overturn Accept	1
PB Overturn Reject	2
PB Reject	40
Reject	103
Reject (2 reviews)	167
Reject (3 reviews)	41
Total	**495**

In terms of acceptance rates, exactly 20% of the submitted papers were accepted. Of the papers firmly decided online (not further discussed at the PB meeting, 163), 37% were accepted and 63% rejected. Of the papers decided in the PB meeting (95), 41% were accepted and 59% rejected. These numbers suggest a balanced process, with the decisions being distributed consistently regardless of how the decisions were arrived at (online or in the PB meeting).

In terms of letter grades and acceptance/rejection, the table below provides the raw data. We first note the large number of papers receiving A and B's, which we believe is due to a better distribution of papers over expertise of the reviewers, as well as (somewhat) due to the reviewers' lighter load (it is difficult to stay positive with 20+ papers to review).

	Accept	Reject
AAA	7	
AAB	9	
AAC	5	
AAD		1
ABB	17	
ABC	15	8
ABD	1	5
ACC	3	11
ACD		6
ADD		1
BBB	17	4
BBC	16	28
BBD	4	9
BCC	4	52
BCD		21
BDD	1	4
CC		10
CCC		31
CCD		15
CD		79
CDD		3
DD		78
DDD		1
Total	99	367

In terms of reviewer expertise, the table below summarizes the distribution of X, Y, and Z across papers (X-Y-Z being the standard scale used at ICSE and most other conferences). On a number of cases, we have been surprised to see PC members assessing themselves as Y while we thought they were experts. After careful analysis, we realized that many papers cover several areas of expertise or fall into application domains that the PC members may not have been familiar with. X is typically used for perfect and complete expertise. We alleviated this problem by assigning PC members with complementary expertise, leading in many cases to two Ys.

We note the very high percentage of papers with at least one X (77% of the 466 reviewed papers) and with a minimum of at least two Ys (96%). Only two papers had Z only expertise, and only 15 a single Y. Compared to ICSE 2013, these

represent improvements, which is significant since more papers were submitted and needed to be reviewed. For example, in 2013, the percentage of papers with one X was lower at 72.4%. Further, the number of papers with at least one Z was higher in 2013 at 22.8%, compared to 18% in 2014. We attribute this to the larger size of the PC (naturally).

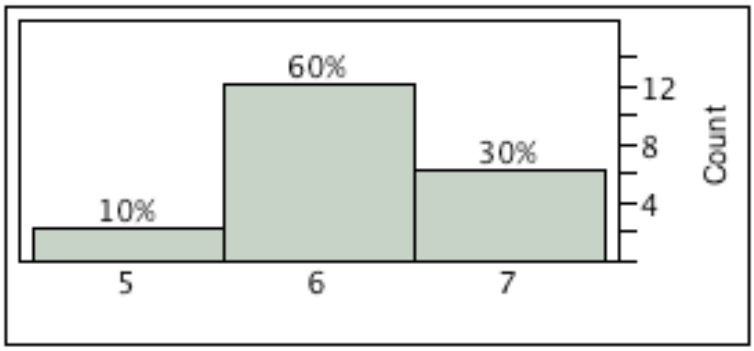
Expertise	Papers 2014	Papers 2013
XX	52	50
XXX	27	17
XXY	81	39
XXZ	10	11
XY	68	95
XYY	86	66
XYZ	25	29
XZ	7	15
XZZ	3	2
YY	29	47
YYY	39	29
YYZ	22	26
YZ	6	14
YZZ	9	4
ZZ	1	2
ZZZ	1	0
Total	466	446

2. How adequate was the reviewing expertise and quality?

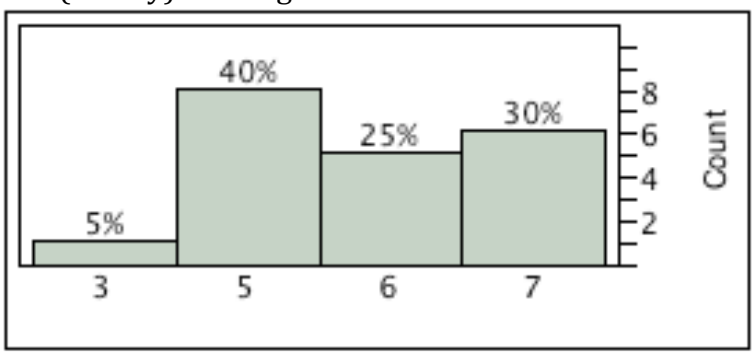
2.1. What was the Program Board's perception?

There were three questions related to this aspect in the survey, all defined on a Likert scale from 1 (Strongly disagree) to 7 (Strongly agree). Following standard practice, we selected a Likert scale as it has been widely studied in psychology, which showed that it was an unbiased subjective measurement scale and that it could be used as an interval scale during analysis (e.g., computing and comparing average scores).

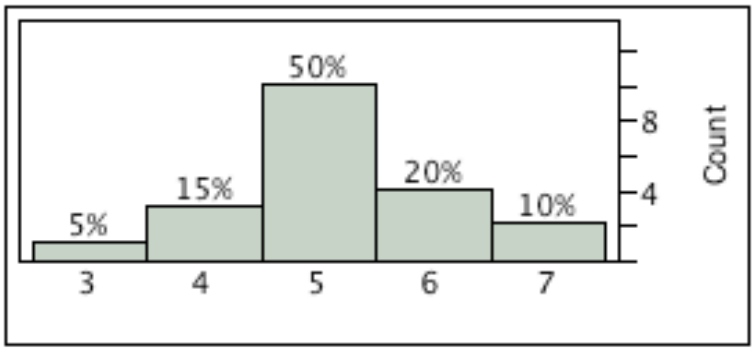
Q: I mostly managed papers within my area(s) of expertise.



Q: Reviews were (mostly) thorough.



Q: Reviews were (mostly) constructive.



From the figures above, we can conclude that all PB members were confident about the adequacy of their expertise for most papers (scores range from 5 to 7). All but one PB member were positive about the thoroughness of reviews (5 to 7). Note that (most) PB members carefully monitored the reviews and asked, on many occasions, that some of the PC members perform changes to their reviews.

Here is what the only negative PB member stated:
"Given the number of submissions, ICSE needs a hierarchical organization of the PC. I think nobody is able to provide significant reviews for more than 15/20 papers, and thus there must be a proper distribution of the work. I thus liked the idea of having many more reviewers than usual, and I am sure the reviewers appreciated the new load. The challenges associated with this new system, or with any system that involves a significant number of reviewers, is that the quality of reviews must be controlled properly to avoid too high differences between one review and another. Another problem is that some reviewers, who were not supposed to show up at the PC meeting, did not feel they were really

part of the process, and thus they provided pretty dry reviews, did not really discuss them, and at a given point they disappeared."

The positive comments were also numerous, and are in some ways contradicting the above comment. They can be summarized as follows (with frequencies of occurrences in parentheses):

- The new model enabled a quality assurance of reviews at a level not possible for two PC chairs alone (3)
- Reviews are of better quality: relevant, complete, respectful (2)
- There were a significantly lower number of papers to review per reviewer (3)
- The larger number of PC members led to a better match of papers with appropriate expertise (3)

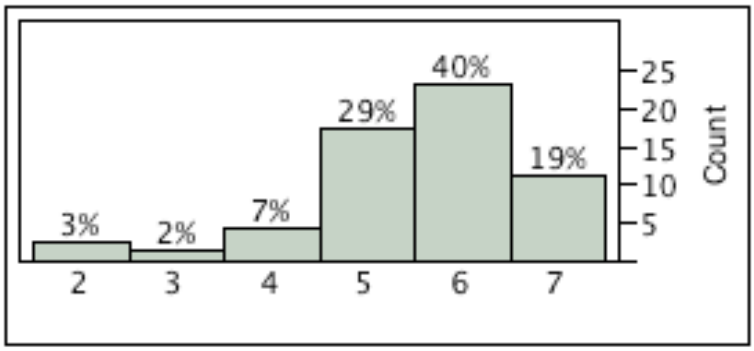
Other negative comments about reviews included:

- Controlling for uneven reviews was challenging (2)
- Reviewers were not involved in the physical meeting (1)

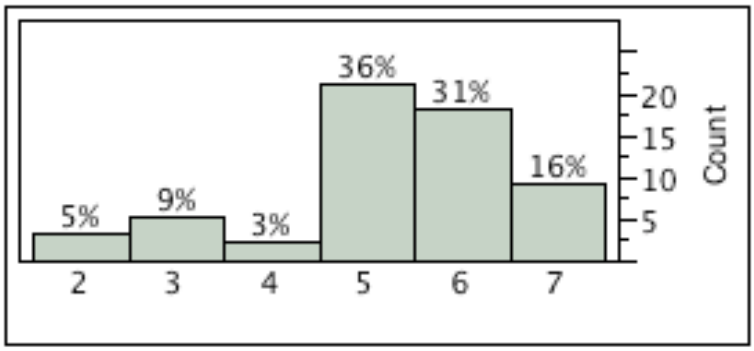
Regarding our last question as to whether reviews were constructive, the scores are still positive but less so than for the other two questions (there is a notable shift to a score of 5 as opposed to 6 or 7). In addition to the same PB member who scored 3 in the previous question, two PB members gave a score of 4 for this specific question, which is neutral. None of them provided any explanation for their scores. While the overall sentiment is still that reviews are constructive (95% for a score of 4 and higher), it is worthwhile for future Program Chairs to pay attention. There is a significant role here for the Program Board members. We experienced some variability in how much they shepherded the Program Committee members (most PB members gave feedback on many reviews, others simply took them in). We believe if all PB members equally scrutinized the reviews and made suggestions on how they could be more constructive, this percentage would go up.

2.2. What was the Program committee's perception?

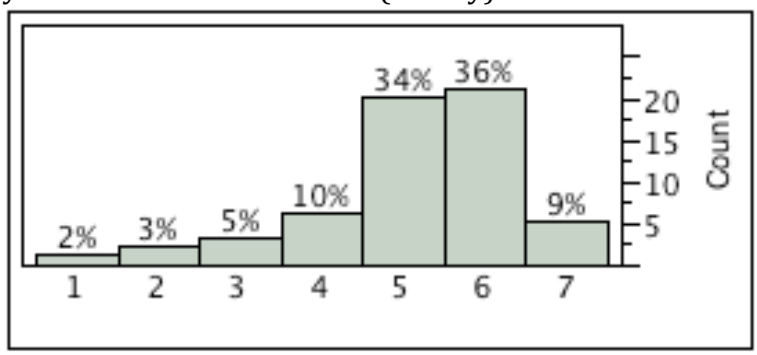
Q: I mostly reviewed papers within my area(s) of expertise.



Q: Reviews by the other reviewers were (mostly) thorough.



Q: Reviews by the other reviewers were (mostly) constructive.

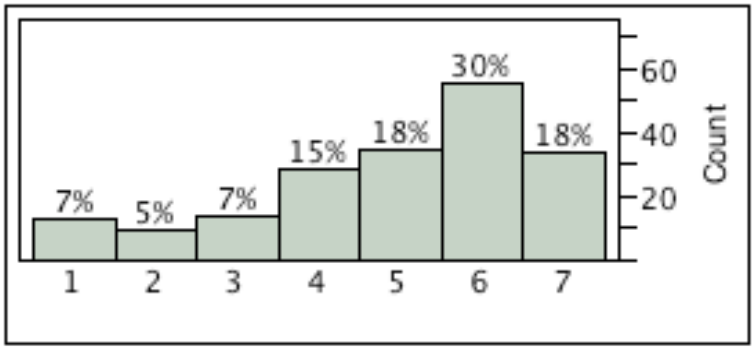


As opposed to the PB member distribution, they were seven PC members with neutral (4) or negative (2,3) scores regarding expertise. Neutral and negative scores represented 7% and 5% of the PC members who filled out the survey, respectively. It is difficult to interpret such results as we are the first ones to capture such data at ICSE. Given how wide the field of software engineering has become, predicting the expertise that will be required, at the right level of accuracy to enable a balanced distribution of reviews, will probably remain a challenge. In such a context, 12% of negative and neutral scores is probably not that bad. We note, however, the distribution of X, Y, and Z presented earlier. It seems PC members are slightly more pessimistic in their views expressed in the survey than in their self-assessed experience levels for each paper individually. A Y sometimes is seen as 'insufficient expertise'.

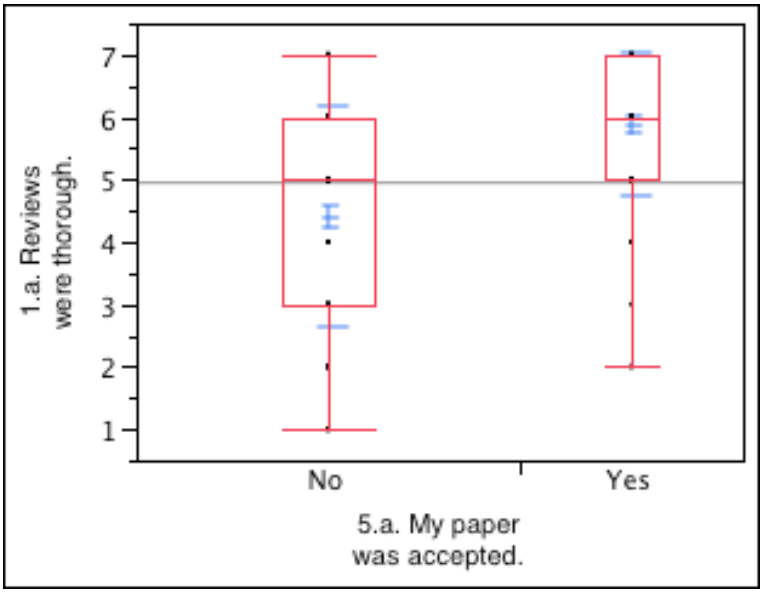
Regarding the perception of PC members regarding their fellow PC members and their reviews, the scores are also mostly positive, with 83% and 79% of positive scores (5-7) for the two questions about how thorough and constructive were the other reviews, respectively. It is interesting to note that PB members were more positive about reviewers' thoroughness, but distinctly less positive about the level of constructiveness. This is perhaps less of a surprise, since PB members had to make judgments based on the reviews, and the more constructive the reviews were on what a paper needed to do to improve, the clearer typically its contribution, strengths, and weaknesses were.

2.3. What was the Authors' perception?

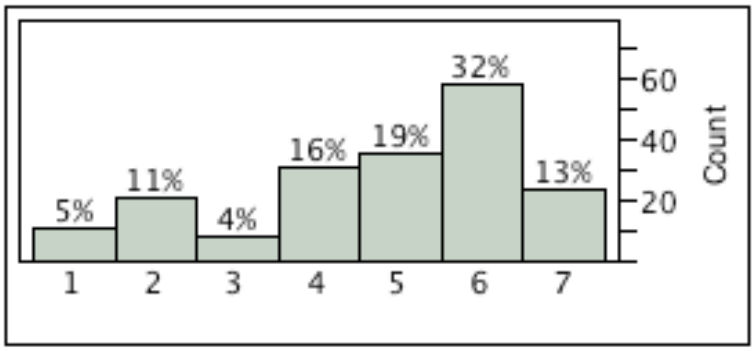
Q: Reviews were thorough.



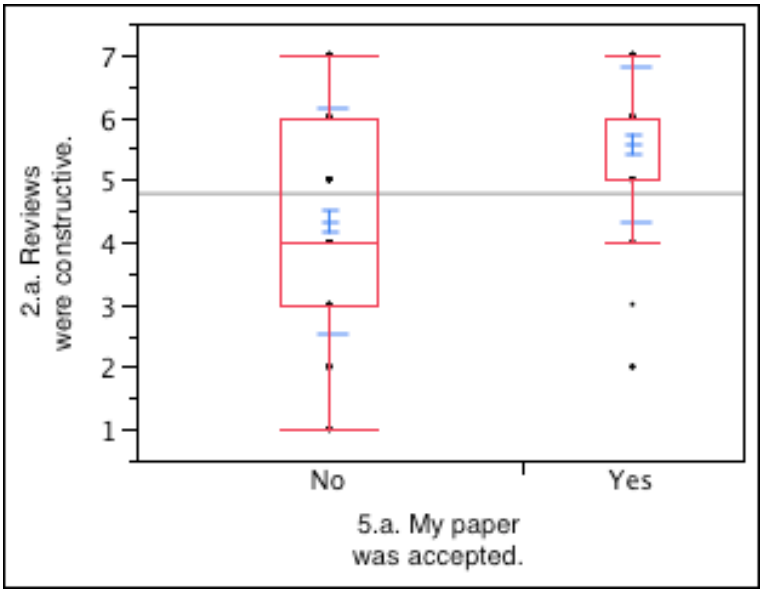
Regarding the thoroughness of reviews, 66% of authors were positive, 15% were neutral, and 19% were negative. The variance in scores is very high, covering the entire range. However it can largely be explained by whether a paper was accepted or not. As we can see in the box plots below, there is a statistically significant association (Wilcoxon Rank Sums test, $p < 0.0001$) between thoroughness scores and paper acceptance. Accepted papers show an average score of 5.9 versus 4.4 for rejected papers. Authors of accepted papers are therefore clearly more positive about reviews whereas the other authors are rather neutral, on average, with widely varying opinions.



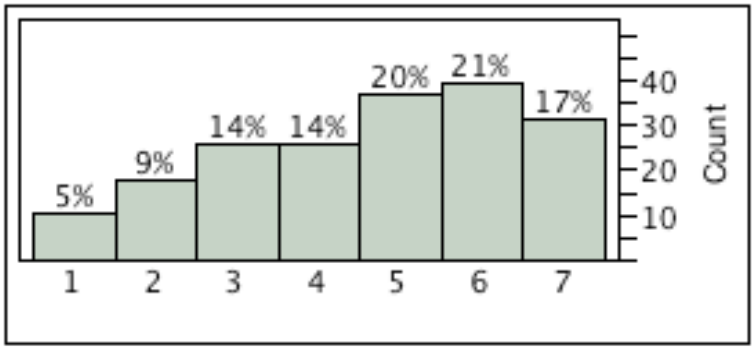
Q: Reviews were constructive.



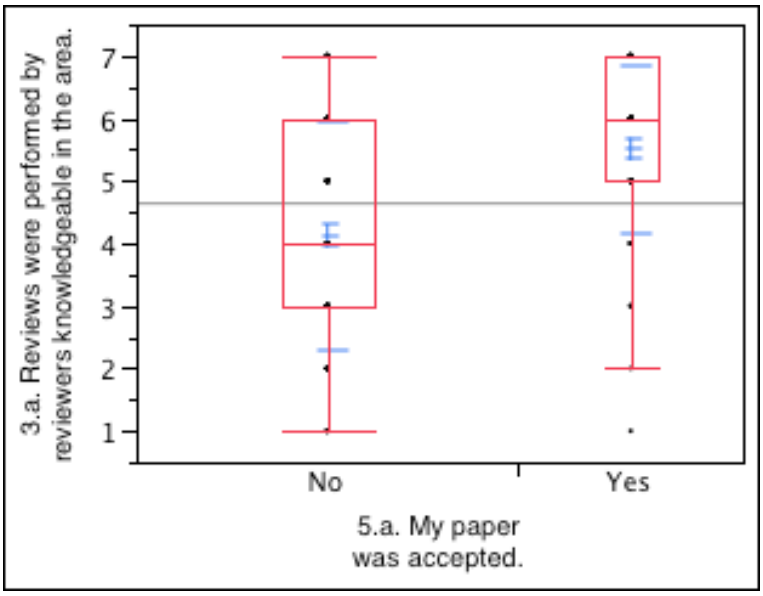
Regarding whether reviews were constructive, 64% of authors were positive, 16% were neutral, and 20% were negative. The variance in scores is once again very high and can, as seen in the box plot below, once again be largely explained by the acceptance or rejection of the papers. Indeed, the average score for accepted papers is 5.56 versus 4.32 for rejected papers. The variance in scores for rejected papers is once more much higher than that for accepted papers, as shown by the box plots. These scores are also slightly more negative than thoroughness scores (though, we note still positive on average!)



Q: Reviews were performed by reviewers knowledgeable in the area.



Results regarding how knowledgeable the reviewers were, as perceived by the authors, are overall less positive: a clear majority of positive scores (58%), but also more negative scores (28%), and the usual large variance that is largely explained by whether a paper was accepted or not, as depicted below in the box plots. This result seems to contradict the much more strongly positive point of view of the reviewers themselves, as well as the Program Board.



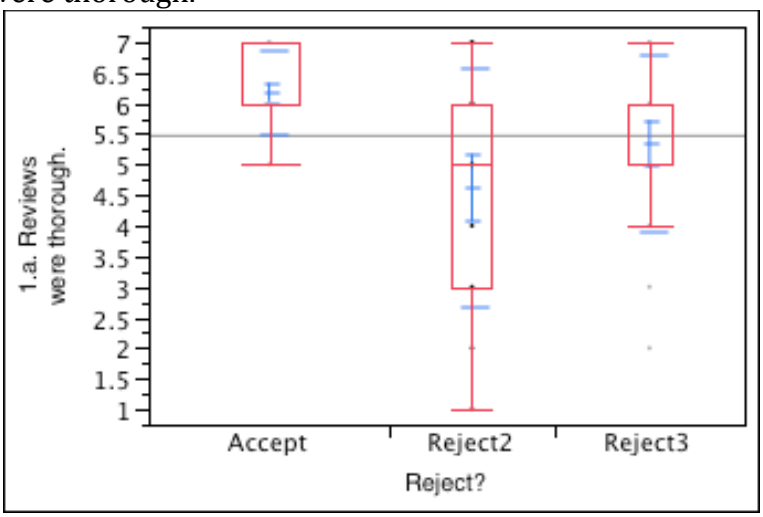
Overall, we conclude that authors are (understandably) partly affected in their judgment by the decision on their paper. Furthermore, it is our historical experience that reviews for papers of poor quality tend to, on average, be less complete and detailed, which we suspect also contributes. The high variability in scores for rejected papers might also be due to authors who received only two reviews. Informal feedback we have received over the years, indicates that receiving two reviews is often a strong disappointment and could, at least partially, explain such results.

Only a clear link between survey and review data would allow us to explore the above issues in more detail. Unfortunately, because the survey data collection was anonymous, and names were only be provided on a voluntary basis, we do not have complete information in that respect. However, for 47 papers, all with at least one X or two Ys in terms of review expertise, we were able to link survey and review data because the survey respondents provided their names. Below is a box plot showing the relationship between how thorough the reviews were perceived by the authors and whether the paper was accepted, rejected with three reviews, or rejected after the first round with two reviews. The trends clearly show that the authors' perception was much more negative for the latter category and suggests that papers rejected with two reviews explain, to a large extent, the low scores and wide variance for rejected papers. Note that we only show the plot for the first question about review thoroughness but that similar results were obtained for the other questions. The difference is statistically significant ($p < 0.05$), despite the small number of observations. Whether this difference stems from knowing their paper was considered poor because it just got two reviews, from the reviews perhaps not being as thorough because the

reviewers chose to spend more time with the better papers, or from feeling their paper should have received three reviews regardless, we do not know.

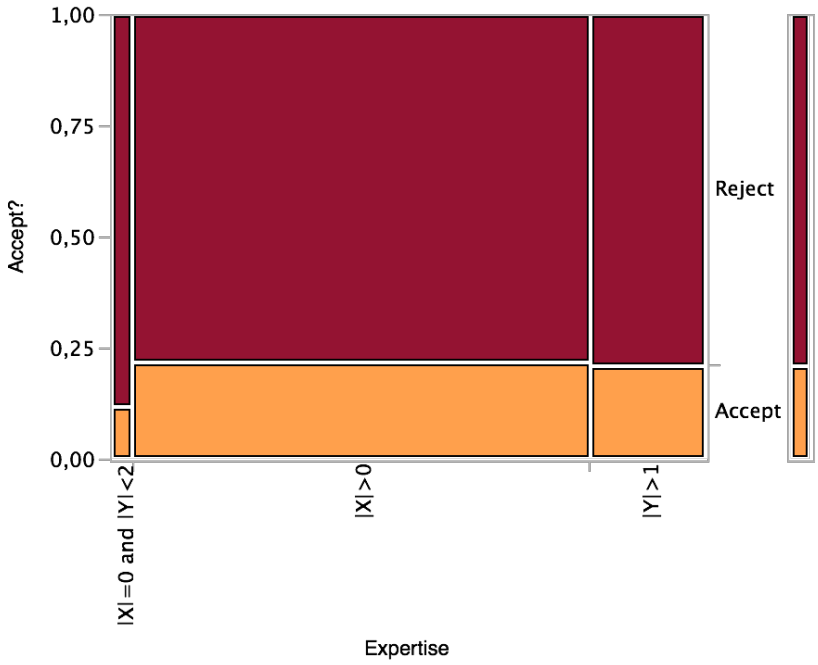
One possible conclusion is that, for the sake of improving the perception of ICSE among a larger number of researchers, we should consider providing three reviews for each paper. It would require a larger PC though, a solution made possible by the Program Board model. Yet, for many of these papers, the extra review entails limited additional work, and it is up for discussion whether this is worth the improved perception.

Q: Reviews were thorough.



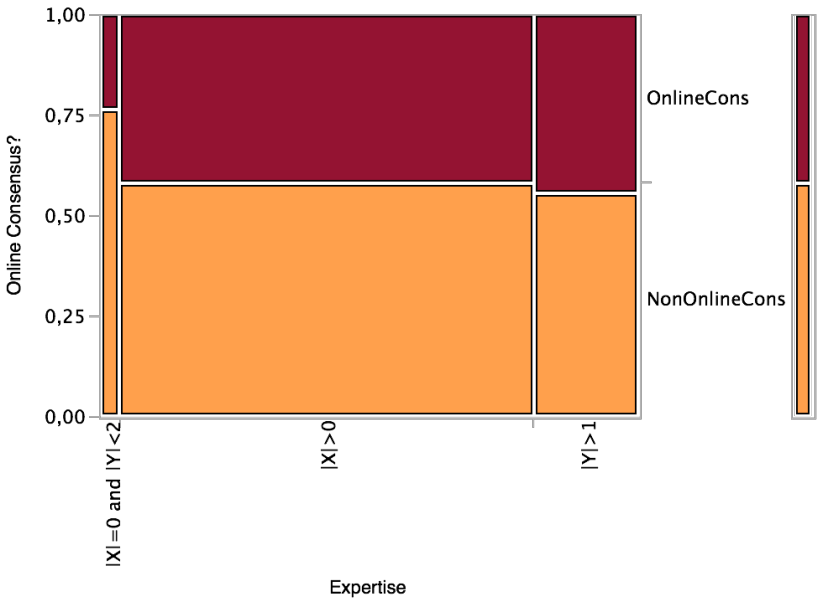
2.4. Was there a difference in acceptance rates between papers with at least one X vs. no X and two Ys?

Acceptance rates may be one indicator of differences in decisions across papers with different levels of expertise. In the analysis above, we considered papers with one X or two Ys to be papers with sufficient expertise to lead to a reliable final decision. But is there a difference between the two? It turns out that papers with at least one X ($|X| > 0$) or no X but two Ys ($|Y| > 1$) have a similar acceptance rate around 21%. Other papers, which only account for 17 of the total number of papers, show a lower acceptance rate around 12%. This is depicted in the Mosaic plot below. These results indicate that expertise, overall, is tremendously important; if papers for which the expertise is insufficient are accepted at a lower rate, this implies that review models that bring additional expertise to the review process should be strongly encouraged.



2.5. Were papers with two Ys less likely to achieve consensus during the online discussion than papers with at least one X?

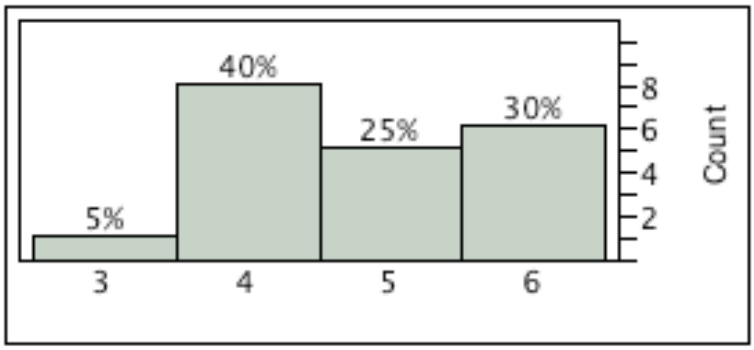
A trend that could be expected is that the presence of higher expertise might lead to a higher percentage of online consensus decisions. For the few papers with no X and less than two Ys, the online consensus rate is as expected lower (23.5%) than for papers with more expertise (> 40%). The question is now whether there is a difference between papers with two Ys and no X, and papers with at least one X. For both categories of papers, we get a similar percentage of online consensus decisions among papers, 42% and 44%, respectively, as depicted in the Mosaic plot below. This further supports the claim that papers with two Ys and no X lead to similar decisions as papers with at least one X.



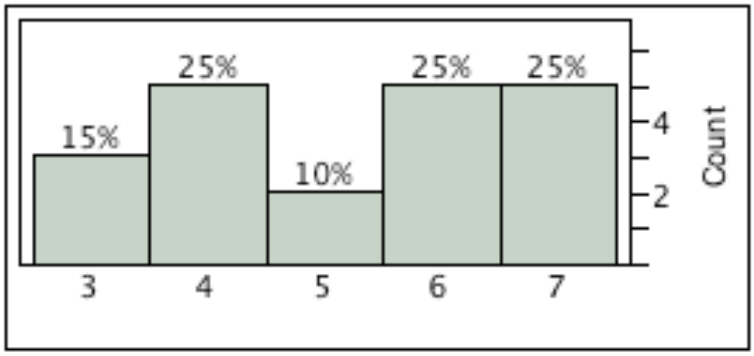
3. How effective was the online discussion?

3.1. Program board perception

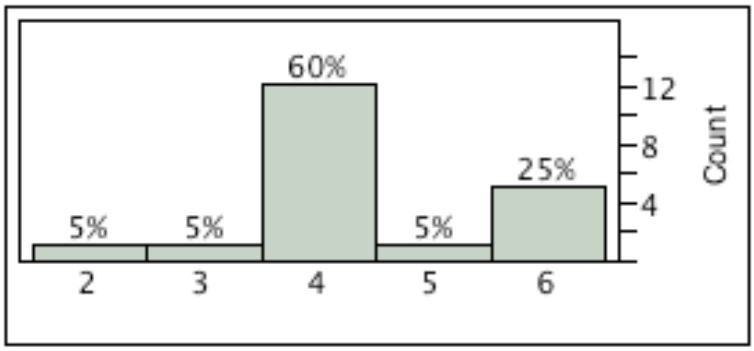
Q: Reviewers participated fully in the online discussion.



Q: Most online discussions were insightful and useful.



Q: Length of online discussion.



Regarding participation of PC members to the online discussion, 40% of PB members were neutral and 5% negative. Negative scores rise to 15% when asking PB members about the usefulness of discussions. There is therefore room for improvement regarding online discussions, according to the perception of PB members. Once again, it is difficult to fully interpret such results as we do not have past ICSE data about PC meeting discussions. Part of the problem is likely due to the fact that this process was entirely new to PC members and that the specific expectations, despite our precise guidelines, were not necessarily understood in the same way by all. We also witnessed different PB members begin and engage in the discussion somewhat differently; encoding their best practices in, for instance, a series of standard e-mail messages that PB members can use might well help.

As for the length of online discussions, 60% of the PB members were neutral, with a larger proportion (30%) thinking they were too long (5, 6) rather than too short (10%). Hence, opinions seem to be widely distributed on that topic and it is hard to conclude anything in that respect.

PB comments related to online discussions can be summarized as follows:

Positive:

- As a PB member, it was not difficult to discuss on the behalf of PC members as online discussions were sufficiently informative
- Online discussions were overall more detailed and thorough than what typically happens at a physical PC meeting (2)

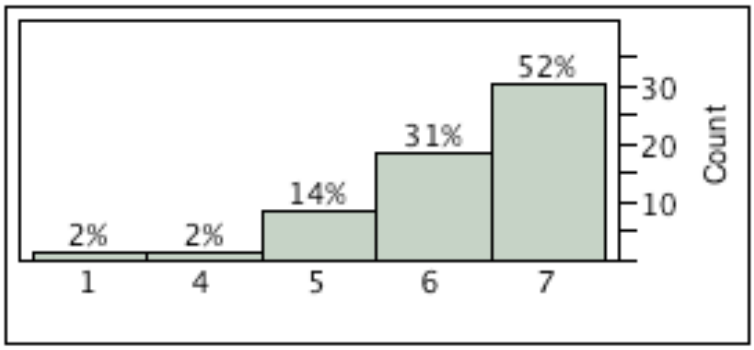
Negative:

- Some "power games" took place among reviewers and we should consider keeping reviewers anonymous
- Some borderline papers would have benefitted from face-to-face discussions and for some papers we should consider teleconferencing

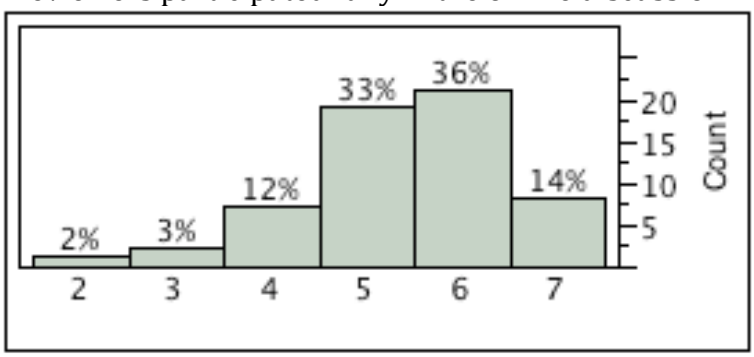
Note that we did consider such teleconferences, but preferred for all papers to be treated equally, and for the actual discussion to be captured so it was available during the PB meeting.

3.2. Program committee perception

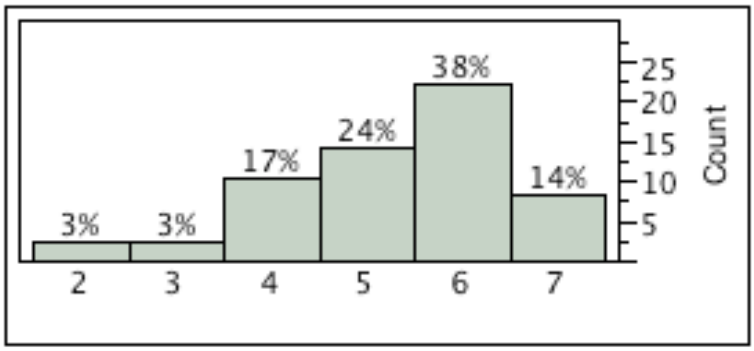
Q: I participated fully in the online discussion.



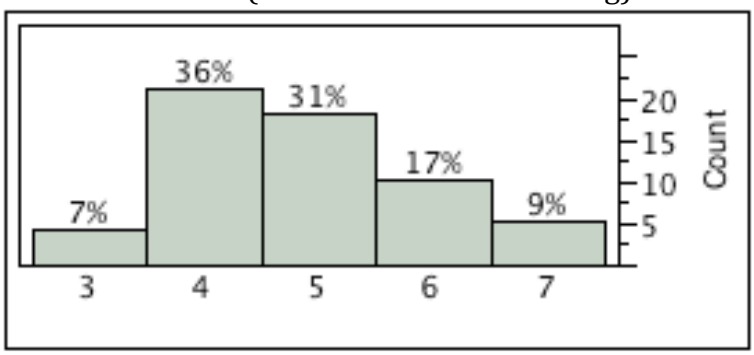
Q: Most other reviewers participated fully in the online discussion.



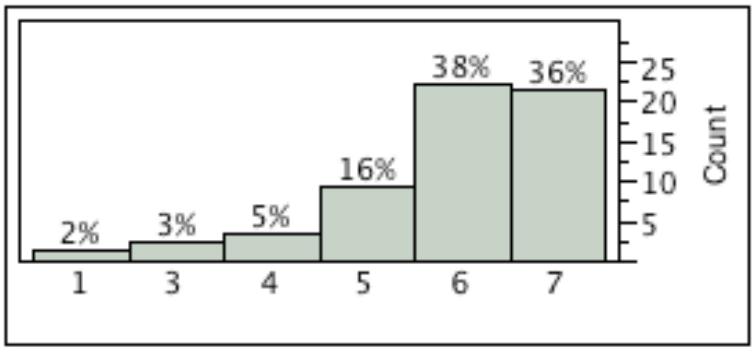
Q: Most online discussions were insightful and useful.



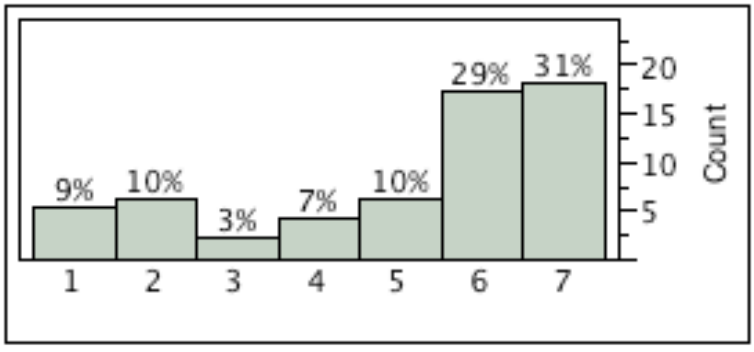
Q: Length of online discussion. (Scale: Too short - Too long)



Q: I was comfortable making (some) accept/reject decisions online.



Q: I was comfortable leaving (some) accept/reject decisions to the program board.



96% of PC members were positive about being sufficiently involved in online discussions. But their perception of other reviewers was more negative, with 5% and 12% of negative and neutral scores, respectively. Such scores are overall highly positive, but it is interesting to note that PC members are significantly more positive about their own online discussion involvement than that of their

colleagues (the latter perception is more in line with how the online participation of PC members was perceived by PB members, as discussed in Section 2.1).

Generally PC members viewed the insightfulness and usefulness of the online discussions more favorably than PB members (76% positive versus 60% positive). Of primary concern here, we believe, is the somewhat inherently uncertainty of when reviewers participate in the online discussions. First, with 78 PC members, different reviewers will have different schedules and travel. Second, whether or not someone participates and the extent to which depends a lot on the individual. Though PB members led, actively prompted, and monitored the online discussions, this problem may be inherent to peer reviewing (e.g., with people overcommitting their time and underestimating the required effort) and once again, we do not have a baseline of comparison.

A large majority of PC members found the online discussion to be adequate (36%) or too long (57%). This is somewhat different from PB members who were far less numerous to find the discussions too long. An important point to make in this regard is, once more, the fact that, with a large PC, different reviewers will have different schedules and it would not be possible to have just a meaningful online discussion period of just 1 or 2 weeks.

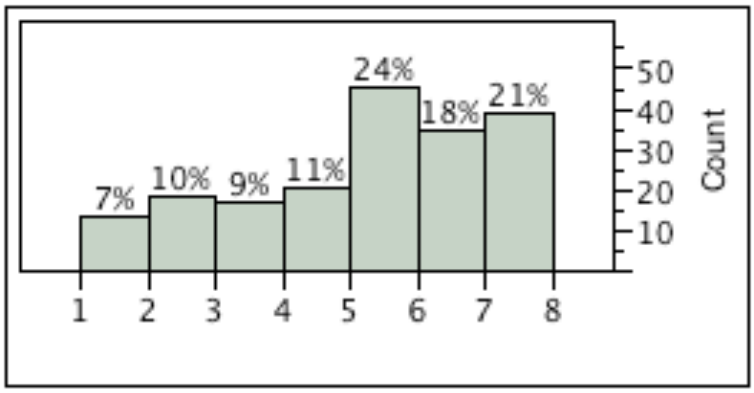
Based on online discussions, a large majority (90%) of PC members were comfortable making decisions about the papers on which a consensus was reached (95% percent if we include neutral). A smaller majority (70%) was positive about letting the PB make decisions for papers on which no consensus was reached (77% if we include neutral). A higher degree of doubt (though with still a very positive trend) about the capacity of PB members to make final decisions is understandable, as there is no history and experience with PB meetings at ICSE. After all, it is a departure from tradition for PC members, who have read papers carefully, to not be able to directly influence the final discussion. However, it was interesting to see that more than a few online discussions quite comfortably ended with the PC members agreeing to disagree and handing off the decision to the PB meeting 'because they will set the bar'.

4. How effective was the PB meeting?

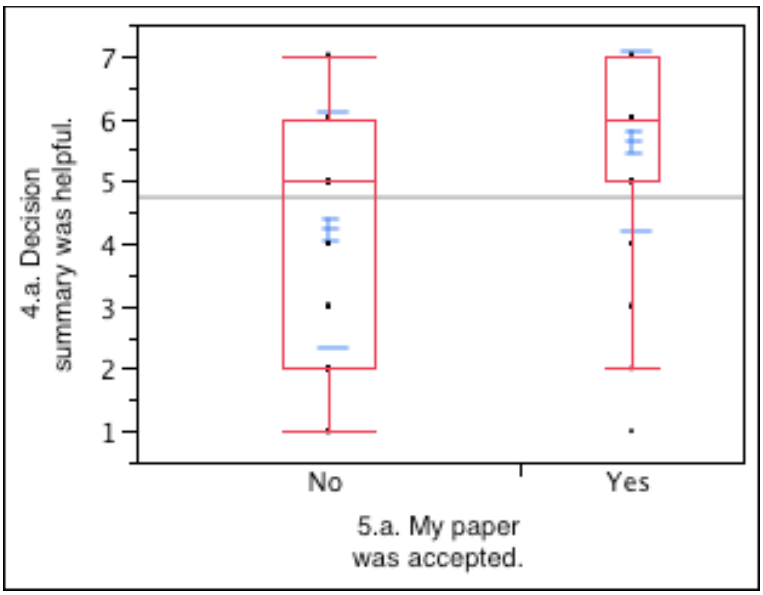
We have not collected data directly related to the PB meeting (other than the acceptance rates discussed in Section 1). However, one additional change we made to the process was decision summaries[3]: each paper that was discussed (online or in the PB meeting) received a decision summary capturing the primary reasons why a paper was accepted/rejected as well as key suggestions for improving the paper.

[3] Decision summaries were introduced by Gail Murphy and Mauro Pezze at ICSE 2012 but were not used at ICSE 2013.

Q: Decision summary was helpful.



Results regarding decision summaries are similar to results regarding the quality of reviews. 63% of authors found the decision summaries useful, 11% were neutral, and 26% were negative. We expected decision summaries to be particularly useful in borderline cases, so we should not be overly surprised if a quarter of the authors did not find them too helpful, since for many rejected papers the decision summary was short and simply reconfirmed the key points of the individual reviews. Possibly confirming this, a large variation can be observed for accepted and rejected papers regarding decision summaries, where the latter are significantly more negative (though still neutral on average) than the former, as depicted by the boxplot below. Overall, results seem to suggest that discussion summaries are useful and should be continued in the next editions of ICSE, regardless of which review model is employed.



Discussion summaries were not commented on very much by PB members except for one comment, stating that writing summaries for mediocre papers, for which no or little discussion occurred, is not particularly useful. At the same time, it is not particularly onerous either, and it ensures that each paper is treated in the same way.

There were, however, a number of comments regarding the PB meeting itself that can be summarized as follows (again, all comments are included verbatim at the end of this report):

Positive:

- The meeting was more thorough, thoughtful, fruitful, enjoyable, professional and less emotional than typical ICSE PC meetings (3)
- The small size of the meeting was more manageable and made it easier to stay engaged (2)
- It encouraged acceptance of truly worthy papers, more so than the previous model did
- The program board members were active during the meeting and got involved in the discussions of papers they were not in charge of

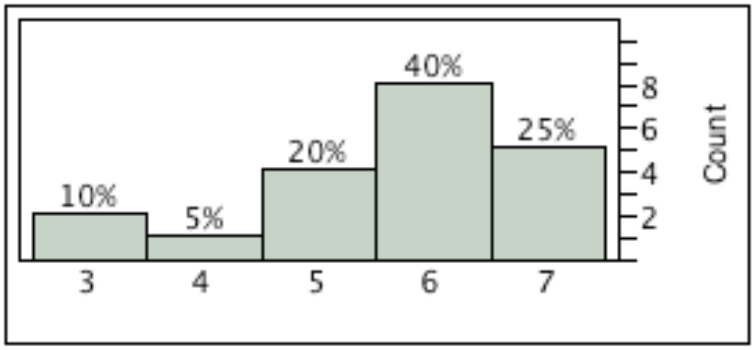
Negative:

- There was inconsistency among PB member attitudes, either acting more as a reviewer or a meta-reviewer
- The role of PB members in the meeting should be clarified (2)
- The role of the second PB reader was challenging and should be clarified (3)
- There was not enough participation among PB members during the meeting
- Some PB members were not sufficiently familiar with some of the papers.

One important lesson learned is that, despite precise and detailed guidelines, some PB members were confused about the role of PB members and second PB readers. Because the comments above are contradictory, it is also difficult to conclude much with certainty. Some degree of confusion may be partly due to the novelty of the process, or a lack of clarity in the detailed guidelines that were provided.

5. How reliable was the decision process?

Q: Final decisions of acceptance/rejection were balanced and justified.



A large majority of PB members (85%) found the final decisions to be balanced and justified (scores 5-7). An additional PB member was neutral and two PB

members were mildly negative (3). Below are relevant statements from these more negative PB members.

- "Finally, my last problem is about the discussions during the PB meeting. I found strange the fact that I had to discuss a paper, but I was not really allowed to express my idea and my judgment. Sometimes it happened, but in theory each PB member had to discuss a paper as if s/he were his/her reviewers. I found this very difficult especially when I did not really agree with the reviewers assigned to the paper. I also did not really understand the role of the second reader."
- "I expected more participation on the PB meeting (in most papers we had 2 participants + chairs which made it difficult to determine where to set the bar), and often had a sense that PB members did not really know the papers inside out (but maybe that is ok with this model)"

There were a number of positive comments as well regarding the decision process:

- Because of the new model, PB members had a better ability to consistently accept or reject papers
- Directly accepting papers with strong scores was very effective (2)
- The discussions were open and unbiased

Note that PB members were not told to refrain from expressing their own opinions, but were told to also account for the opinions of the reviewers and present a full picture. That point should perhaps be further clarified in the future, though the novelty of the process was also part of the challenge, as noted by a PB member.

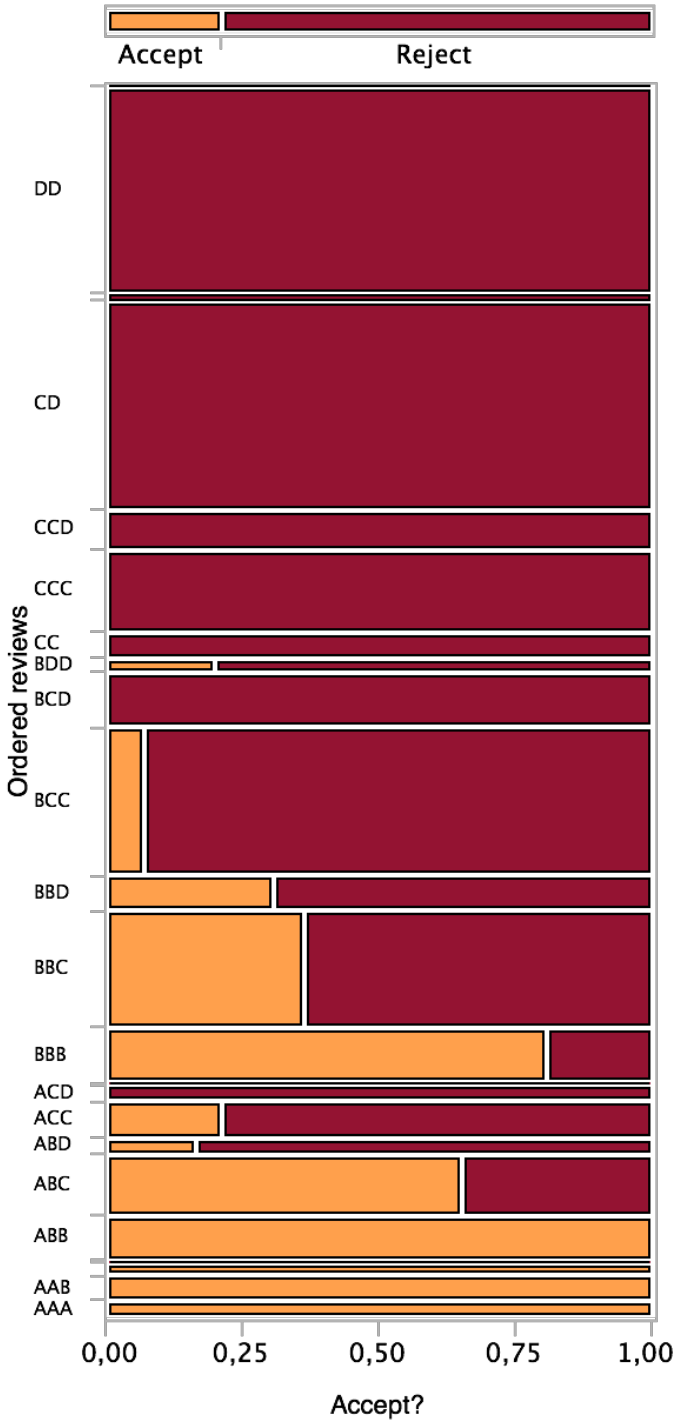
Regarding the comment on limited participation of PB members in the discussions of papers they were not in charge of, this is a matter of perception. One PB member collected data during the meeting and told us such discussions occurred in a third of the cases. This is not unsurprising, since the PB meeting was seeded with papers that the PC co-chairs knew would help set the bar, but were otherwise uncontroversial. Overall, we felt that, for the papers that needed it, more extensive discussion took place (indeed, multiple papers could be discussed for 20 or even 30 minutes, involving several PB members, something no longer possible in the conventional ICSE PC meeting).

An important change we wanted to achieve was for the discussion at the PB meeting to be able to focus on an emerging 'bar', rather than each paper in isolation. That is, we wanted to explicitly discuss 'the rules' by which papers were judged and eventually accepted/rejected, in order to treat each paper fairly. This required time, which we achieved by reducing the numbers of papers to be discussed in person through the online discussion, and careful consideration during deliberations and putting papers in light of previous discussions. Some PB members felt somewhat uncomfortable, because some papers were accepted or rejected that they felt differently about, but had to concede in context of other papers. We suspect this is partly the source of some of the discontent (PB

members do not 'own' the decision over a paper). At the same time, we strongly believe this is the right way to treat all of the papers: equally.

6. What should be the score threshold above which a paper should go to the second round or be discussed?

The Mosaic plot below shows the proportion of papers with different scores and their acceptance rate (raw data in the table on page 4).



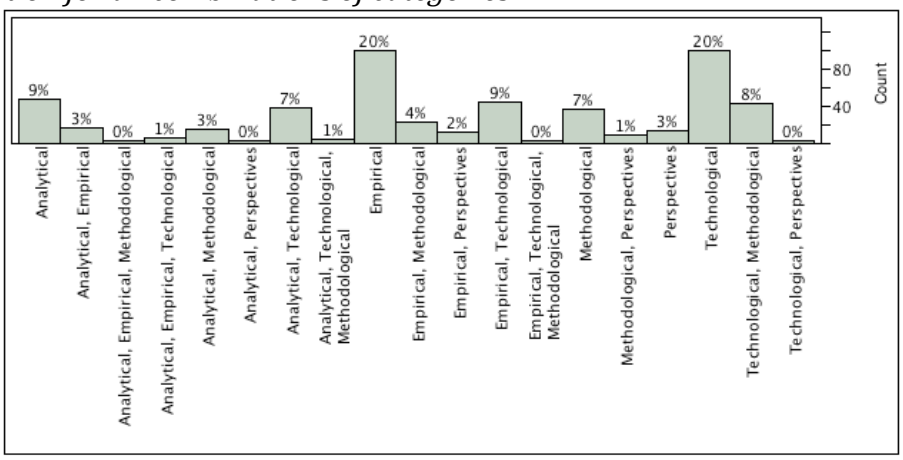
Out of 116 papers with two Cs, seven got accepted (ACC and BCC). This is a 6% acceptance rate. In our case, these six papers amount to 7% of the total papers we accepted whereas the 116 additional reviews that were needed to consider these papers represent roughly 9% of our total number of reviews. Whether this is worth it depends on the review load of PC members and therefore if the additional reviews can be afforded. Even one BDD paper, out of five, was accepted in the end.

A discussed earlier, our policy was for the PB to oversee whether CC, CD or DD papers needed an extra review after the first round. CC papers by default did, CD and DD papers did not. PB members were quite active in this regard.

7. What are the lessons learned from using paper categories?

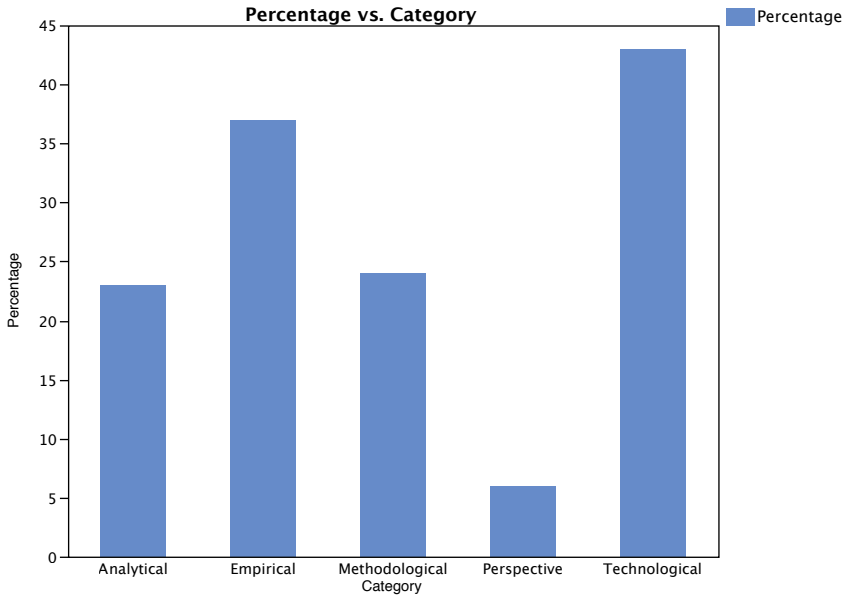
7.1. What are the relative proportions of paper categories?

Distribution for all combinations of categories



Distribution for each category

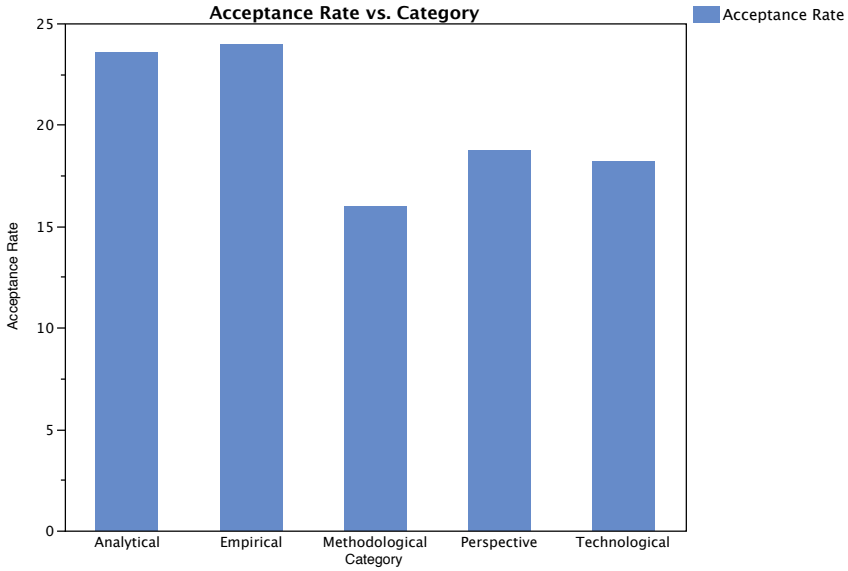
Since many papers have listed two or more categories, we look at the distributions of how often a specific category was listed, alone or combined with others.



Most papers have one category (59%) or two categories. In the future we recommend not allowing more than two categories, as multiple categories limits the usefulness of the categorization in the review process (and it is unlikely a paper makes primary contributions in three different categories).

The most represented categories are, from higher to lower percentages: Technological (43%), Empirical (37%), Methodological (24%), Analytical (23%), and Perspective (6%). Having a large proportion of technological papers in an engineering discipline should not come as a surprise. For many years now, the empirical nature of software engineering research has been recognized and the large percentage of empirical papers confirms this trend. Unfortunately, we have a very small number of perspective papers. Because we believe such papers are important for the community, we suggest to keep promoting such papers in the future until it becomes a natural part of publication practice. Further clarifying acceptance criteria for such papers is also recommended.

7.2. What are differences in acceptance rates across categories of papers?



Based on the distributions of accepted papers across categories, the only categories that are significantly higher (~24%) than the average (20%) in terms of acceptance rate are Empirical[4] and Analytical. Though Perspective papers are about average, the picture changes considerably when examining the thirteen papers classified exclusively in the Perspective category (pure Perspective papers): only one of them was accepted in the end.

[4] Though we do not have data to support this, there was anecdotal evidence during online discussions and the PB meeting, that the review of papers reporting human studies -- in particular those including qualitative analysis results -- was in many cases challenging. Such papers have been submitted in increasingly large numbers over the years and, in the future, we need to ensure sufficient competence to review them in the PC and PB.

Conclusions

We summarize here our main conclusions and personal reflections based on the above analysis. We provide what we think are the most plausible interpretations for the data and complement them with recommendations.

Overall conclusion

Overall, the feedback regarding the new review process was very positive. This includes the Program Board model, the classification of papers, and the use of discussion summaries. However, we have identified a number of improvement points described below. As for any new process, we expect it to converge over time towards a more stable and refined implementation.

Scalability

An important property of the Program Board model is that it scales, and could handle a doubling of the number of submissions without requiring yet another model to be adopted. To handle the actual reviewing, the size of the PC can be increased and/or PC members could review more papers. To handle the decision making process, the size of the PB can also easily be increased while still being able to hold a manageable, physical PB meeting. The current PC-only model, or slight variants thereof, simply cannot.

Equality

A second important property of the Program Board model is that all papers are treated equal. Except for papers that are rejected after two reviews, all other papers are reviewed by an equal number of reviewers, are given the chance to be decided upon by those reviewers in the online discussion, and are discussed in the PB meeting if no consensus was reached in the online discussion. A key element here is that PB members should not be reviewers (otherwise, these papers receive five reviews – three PC reviewers, one PB member overseeing the paper, and one PB reader). With an increasing number of reviewers, the chance of a paper being rejected increases. This is why PB members are to discuss the bar of ICSE, and put the papers and their reviews in context of this bar. Generally, PB members should refrain from strongly asserting their own review opinion – it would create inequalities among the papers and how they are reviewed. We get back to this point below.

Expertise

Expertise on papers seems to be key to providing balanced decisions (papers with Y's and Z's tend to not be resolved in the online phase and overall are rejected more frequently). While some of those papers may be out of scope altogether, it is the responsibility of the review process to treat every paper fairly. In that regard, a larger review committee with broader expertise is important, especially in the context of ICSE ever expanding its topics of interest. For reasons of equality, it is important that this expertise be present in the PC

itself, and from the start; assigning fourth reviewers begins to treat some papers differently and, from our personal experience, it seems to be the case that the higher the number of reviewers, the greater the chance for a paper to be rejected.

Roles of program board members

As discussed, some PB members were wondering to what degree they could or could not serve as reviewers. We view the primary role of the PB member first to moderate the discussion and guide reviewers to a consensus, second to present this consensus at the PB meeting, and third to discuss at the PB meeting how the particular suggested decision 'fits' with the bar that is being established. This is similar to the case of journal associate editors, for good reason: it is difficult to both be a reviewer and to objectively represent and analyze other reviews.

What happens, though, when, after reading a paper, a PB member disagrees with the suggested decision in reading a paper? This question came up a few times during the PB meeting. Our strong suggestion is for such disagreements to be voiced during the online discussion. This gives the PC members a chance to reflect and respond, and typically deepens the discussion to arrive at a more satisfactory consensus.

For the PB reader, the role is slightly different. Since they only read the papers once the online discussion is closed (this is to reduce the number of papers that PB members have to read), they cannot interject if their opinion is different. Our view of the PB reader is therefore as someone who provides a second look at the reviews and discussion, and helps create an overall sense of balance and can act as a knowledgeable interlocutor for the PB member in charge of a paper. While frequently the PB member in charge and the secondary PB reader agreed in their reading of the reviews and discussion, in some cases, they held different views, which led to useful discussions.

Clearly, these are different expectations from the traditional role of being a PC member and, as such, we believe time is an important factor: experience must be built with this model of review, with appropriate refinements over time.

PB meeting

The PB meeting was remarkably free of 'heated discussion' emerging from strongly dissenting views on papers and PC members 'digging in' to defend their positions. PB members, in representing three other reviewers and their opinions, tended to have much less of a personal stake. Discussions were more objective and generally more cooperative as a result.

Consensus decisions from online discussions were reversed in only three cases during the PB meeting (1 to accept, 2 to reject). Such occurrences should be rare, but they are unavoidable given that a critical activity is 'bar setting' and judging decisions of acceptance and rejection on one paper in the context of all of the other papers.

Given that this may happen, it is important to clearly convey to the PC members that their decisions remain recommendations to the Program Board, not final decisions.

Quality of reviews

Overall, the feedback from authors on the quality of review was very positive, however much more so for people who got their paper accepted.

Given the very negative perception of some of the authors who received only two reviews, and the relatively small gain in reviewing effort combined with significant loss of time in the two-round review process, it might be worth it for all papers to be given three reviews in a single round of review. It would shorten the review period (and in the process perhaps enable rebuttals to take place, another factor that may improve satisfaction of the authors even if their papers are rejected) and perhaps increase the perception on the quality of reviews. Typically, reviews of poor quality papers are relatively quick to perform. However, there is no guarantee that providing three reviews to poor quality papers will significantly improve how the authors of these papers perceive the quality of the feedback they receive. It is nevertheless worth experimenting with, if we want to create a strong sense of community at ICSE.

Though we had sufficient expertise on the vast majority of papers, there were a few cases where that was not so (only 17 papers had at most one Y, and just two papers had ZZ or ZZZ expertise only). This is to be expected in such a wide field as software engineering. Overall, the larger PC reduced the number of these occurrences compared to the year before, even with more submissions. In a number of cases, the PB member could provide additional expertise, which helped steer the discussion (a PB member would ask the reviewers to look at a relevant piece of related work, or ask how a paper's contribution related to the state of the art from a certain perspective). This usually eliminated the need to take other measures – and kept equality in the review process.

One possible idea to experiment with might be for the PB and PC members to briefly look at the abstracts of papers when the assignment is made and get back to the PC chairs within two week if they suspect an extreme case of insufficient expertise. We could consider involving additional PC members after the initial assignment based on gaps in expertise. Such a process is only realistic, in terms of timeline, with a one-round review process.

Online Discussions and Program Board Meeting

PB members are responsible for ensuring the quality of the review process, moderating online discussions, and bringing discussions to a decision, either online or at the PB meeting. They are of course entitled to have their own opinions, especially since they usually hold a high level of expertise, but the philosophy that we followed was to avoid an overdue influence of the PB members on the decision process. We must, however, clarify that their opinion should be heard, while exercising care in not being too forceful in expressing it.

This is indeed a subtle exercise, requiring experience. Changing the nature of the PB discussion to be focused on the bar, and how individual papers manifest themselves in its context, is an important shift.

There was a concern that some PC members did not fully participate in online discussions, despite clear expectations from the start. This may be partly due to a lack of close monitoring by some of the PB members. Nonetheless, because the PC is now much larger, we recommend that perhaps a list of reviewers providing substandard reviews should be kept, in order to guide the selection of PC members for subsequent conferences. If this is adopted, the existence of this list should be made explicit in the PC invitation.

We recommend that discussion summaries be made available to all PB and PC members regarding papers for which they have no conflict. This requires a significant modification of Cyberchair, if this is to remain the system to be used. It probably suffices if summaries are only written for papers which have undergone an online or PB meeting discussion.

In order to clarify the roles of PB members, writing detailed guidelines may not be sufficient. We should identify a list of Frequently Asked Questions that could be refined over time until the PB model becomes familiar within the ICSE community, which should take a few years. Moreover, some standard template e-mail messages to be used by PB members might well help the process as well.

Classification of papers

The classification of papers, based on informal feedback, was clearly found to be useful, particularly from the perspective of identifying commensurate evaluation criteria that help orient the online discussion and discussion at the PB meeting, as a shared set of expectations is crucial. However, we recommend that authors do not classify their papers in more than two categories. Even though the contributions may partially spread to more categories, the authors should focus on where the paper's primary contributions fall.

Perspective papers, in a research community, can play an important role as they provide insights into key questions that shape the direction of research and the general perception of 'how the field is doing'. Reviewing them, though, is not an easy task. When is a perspective paper insightful enough to be accepted? There are no hard criteria that, like in the case of Analytical and Empirical papers, are widely accepted in terms of what constitutes a proper Perspective paper. To a lesser extent, this is also true with Technological and Methodological papers.

Surveying after the review process is over

We strongly recommend ICSE adopts a process of surveying authors and review process members (PC, PB, or other model), to build up a body of knowledge that helps it steer the conference and helps communicate to authors what happens with their papers.

One recommendation we want to strongly make is for feedback to tie to papers and the data concerning its review process. This can still be done anonymously and would help strengthen the data analysis and the conclusions that can be drawn.

A final word – an ICSE culture of data and experience

While we have attempted to be as objective as possible in the above, it should be noted that all the above interpretations and recommendations are our best guesses at this moment in time, given that we do not have much comparative data from other review years and models. Only sustained data collection over several years will enable ICSE to provide firmer answers to the questions posed in this report. We strongly believe the ICSE SC should move the conference into this direction.

Acknowledgements

We are very grateful to the program board and program committee of ICSE 2014 for their help and support in implementing the new review process. This would not have been possible without their advice and recommendations.
We would also like to thank Jo Atlee, Domenico Bianculli, Margaret Burnett, and Willem Wisser for their comments on earlier versions.

Program Board comments (Verbatim)

What, to you, were the main advantages and challenges of the new review process (open question)?

More reasoned, thoughtful, and less emotional in-person discussion. Much more manageable meeting. The small size also made it much easier to sty engaged for the entire meeting. Not having the reviewers available made us have to interpret more than we normally would (though I think this only happened in a very small number of cases).

Advantages: Better big-picture ability to consistently accept/reject papers. I liked this model for other reasons too: it's more scalable, and seems to encourage acceptance of truly worthy papers more than the previous model did.
Main advantages: PC meeting with less people, more manageable Quality assurance is distributed among PB (instead of Chairs) Main disadvantage: Reviewers do not make the ultimate decision, this was problematic when online discussions were not rich or deep enough or just lacked enough participation and timely exchanges, and also when discussion direction and outcome defer. I think this led to some poor decisions.

Surprises: + Directly accepting papers with strong scores and no major objections was very effective. This reduced enormously the number of papers we discussed, and I believe it increased the acceptance rate quite a bit. This is a practice to be implemented independently of the model. - I did not sense an improvement on the expertise of reviewers (I do not have a baseline, but ~20% of papers discussed at the PC meeting did not have an X, it would be great if we could get more data on this from previous years) - I expected more participation on the PB meeting (in most papers we had 2 participants + chairs which made it difficult to determine where to set the bar), and often had a sense that PB members did not really know the papers inside out (but maybe that is ok with this model)

Pros: review load was reasonable; review quality was assessed; up to five people looked at borderline papers; online discussion was very helpful; following a staged process was also very useful; a reasonable number of papers was left for discussion at the physical meeting. Cons: reviewers were not involved in the physical meeting.
Advantages: - More effective overall handling of papers. - More effective PB meeting. - Better scalability: PC can be larger, can cope with more submissions, and papers can be assigned to subject area experts. - Reviews are of significantly better quality (more relevant, more complete, more respectful of the authors) - Less travel, pollution and energy consumption. Challenges: - Some say speaking on behalf of others is difficult. My experience is different. I tried to be as neutral as possible when doing so, and didn't feel uncomfortable, mostly because the suggestion was discussed online with the PC members beforehand.

++ The final reviews were of much higher quality. It helped to have 25 PB members looking over all of the reviews and following up with reviewers when their reviews weren't complete. This is too much for two PC chairs to do well, for all reviews. ++ P

A primary challenge is uneven reviewing. This is always a problem but it is hard to know how to get feedback to the reviewers where their reviews were: a) not sufficiently detailed, b) too grounded in their own view of the world, c) lacked clear determination of strengths and weaknesses. A main advantage is reducing the number of reviews per reviewer. Another challenge is educating the program board to focus on the reviews from the committee first before their own review. I think this did in some cases affect acceptance/rejection although we all know whatever the process it has noise.

Given the number of submissions, ICSE needs a hierarchical organization of the PC. I think nobody is able to provide significant reviews for more than 15/20 papers, and thus there must be a proper distribution of the work. I thus liked the idea of having many more reviewers than usual, and I am sure the reviewers appreciated the new load. The challenges associated with this new system, or with any system that involves a significant number of reviewers, is that the quality of reviews must be controlled properly to avoid too high differences between one review and another. Another problem is that some reviewers, who were not supposed to show up at the PC meeting, did not feel they were really part of the process, and thus they provided pretty dry reviews, did not really discuss them, and at a given point they disappeared. Finally, my last problem is about the discussions during the PB meeting. I found strange the fact that I had to discuss a paper, but I was not really allowed to express my idea and my judge. Sometimes it happened, but in theory each PB member had to discuss a paper as if s/he were his/her reviewers. I found this very difficult especially when I did not really agree with the reviewers assigned to the paper. I also did not really understand the role of the second reader.

thorough, deep and interesting discussions during the PB. Much better shared vision of what is a good ICSE paper.
1. Much more detailed online discussion than what typically happens at a PC meeting. 2. PB meeting was more engaging since more people contributed to the discussion. 3. Even though the authors might never know their papers were reviewed very well.

Representing three reviewers, most of program board members seemed less prejudiced, less taking-it-personal, and more fair, calm, and professional. Because of these (it seems), the PB meeting was more fruitful and enjoyable comparing to my previous PC meeting experience. There was a bit of inconsistency among the PB attitudes, either as a meta reviewer (the one who is informed and empowered with PC reviews) or as a review organizer/representative (the one who is summarizing and representing PC opinions). I thought we were instructed to be the latter, but may be not (it still was not clear even after the PB meeting). Then, there were some cases when PC members wanted to bring their paper to PB meeting to make a decision. In that

case, what is really the role of the PB? Merely a rep for the assigned reviewers? This still confuses me.

In the online discussions, I think I notice some "power games" among some reviewers. It could have been a bit more different if the reviewers had been anonymous (among the PC members). But I am not sure if it would have been better or not. I had a couple of papers that I would have strongly argued for rejection/acceptance, but because the three reviews were in agreement to take one way or the other, I did not bring it up. If I had been a meta-reviewer, I would have done it differently. This is not directly related to the new review process, but It bothered me a little bit among PCs and PBs that they become like a bit of examiner, who decides to pass or fail examinees in an entrance exam. I personally believe in diversity and if papers benefit to "some" community members, it still is a worthwhile paper to be presented at a conference.
The online discussions in most cases were much more thorough and thoughtful than we might have had in a face-to-face meeting.

Overall, I _really_ liked the new review process. The multiple rounds worked well to control the load on both the program committee and the program board members. The much smaller PB made for a much more effective face-to-face meeting than large PC meetings. The addition of a second PB member to every paper that was discussed at the PB meeting was also very valuable for ensuring the right outcomes. I strongly support continuing with this new review process and organization model. My most significant issue with the new process was the offline discussions. They were certainly adequate for some of the papers. For others--particularly those on the borderline--I think we really needed discussion. If I had it to do over again, and if I'd had more time, I would have set up teleconferences with the reviewers to discuss the borderline papers--and the papers where a key reviewer did not participate in the online discussion. Including a short period of time in which such papers can be discussed would be really useful.

Advantages + A larger set of PC members that helps to match papers with expertises + A smaller set of persons participating to the physical meeting Challenges: - Refining the role of PB members at the meeting (leader and reader). Reporters, reviewers or both? Some papers have got 5 reviews instead of 3, and this naturally brings them to be rejected (as we all know).
Overall I think this model worked very well. + Involvement of larger set of people in decision making process + Having fewer people in the physical meeting made the meeting much more thorough and in depth + Good to have not all accepted papers discussed, but just a small subset. + The pre-meeting symposium actually helped to create a good atmosphere during the PB meeting. I was surprised how much influence I had as a board member, even in the on line discussions. This puts a HUGE responsibility with the board. - I did not enjoy writing summaries for mediocre papers. This may be inevitable though. - There was some confusion about whether summaries were needed. I did not like writing them, but yet I think it is good to have the key reason and key suggestions for improvement described for every paper.

The program board meeting was active, not only on the specific papers which I was reviewing and managing, but involved many others in the discussion. The size of the PB allowed interaction and discussions across the table. I also got the impression that the discussion was open and un-biased with the goal to select the most feasible papers, in contrast to what many bigger PC meetings end up in defending or fighting publication of one's favorites. The second reader was a somewhat challenging role, but n most cases we managed to keep it as a second reader, and not a fourth reviewer. To me, the PB/PC organization resembles what most journal have, but with the additional value of having the chance to adjust across PB (cf editorial board members). I think the process worked extremely well.

main advantage: existence of PB alone probably raises the quality of the reviews; having s.o. explicitly in charge of the discussion also is good. Question above re: reviews were constructive: That was the case after the discussion and explicit requests to be more constructive. Good. We have had that discussion: I am still not sure the roles of a second reader and a fourth reviewer can clearly be separated. That's not really a problem, but I realized you guys are making a big point of that distinction. Proposal for the future: Maybe explicitly write the expectations w.r.t. two roles down somewhere. During the PB meeting, I sometimes felt that rather than being too critical, we may have now been too generous. But if one has to choose among these two, I find it better to be too generous. Hitting the middle grounds would be great, of course.

The instructions could have been clearer. For instance, I thought that I was expected to have read the papers I was assigned, but it was not clear from the instructions, and clearly not all PB members interpreted the instructions the same way. Having two people in the room who have carefully read the paper is almost essential.
Shorter discussion period. Consider bringing back a limited form of rebuttal (eg, answers to specific questions posed by PC or PB member). For example, one paper might have benefited if authors were asked to summarize succinctly to how their paper differed from a previous publication.

Discussion leaders may prepare a one-slide to be projected at the physical meeting, showing paper summary and discussion summary. I think this would speed up the discussion and would keep it more focused. Maybe reviewers could be involved over Skype during the physical meeting.
I am not sure anything must still be improved. All models have their problems. Perfection doesn't exist in this world. However, I think this is the closest we can get to maximum efficiency in the ICSE context. I my opinion, 20 reviews per capita as in the 'traditional' format, significantly decreases review quality. The best reviewers might even decline the invitation due to the workload. More than 20 reviews is even less acceptable. 10-12 reviews per member is the maximum we should aim for, and I think the PB+PC model is the only one that allows that together with an effective PC management. Maybe a few things can still be improved: - the PB members can be asked to do a final check on the papers before the CRC is sent to the publisher: this is to take action against a tendency nowadays of authors removing critical parts of their papers once they are

accepted to warrant further publication at another venue. - it time allows, a rebuttal can be added, but I am not a strong supporter.

I believe that the two-tier PB/PC model is the way to go. I just don't see how the flat PC model can deal with the numbers of submissions. The biggest problem (with both models) is how to handle ZZZ papers. It might make sense for the PC chairs to consult the entire PB at that point (or more of the PB), to see if anyone has any suggestions for a fourth reviewer among the PC or for a second reader.
Consider adding a meta-review from the program board member written prior to the program board meeting. The discussion period might have been too long. It was hard to get reviewers to respond (despite personal email) and then with a late response, the discussion often ends prematurely. Consider a rebuttal or revision phase. ICSE may be left behind as other communities change. I used to hate rebuttals but have seen several cases now (and no, not a lot) where the rebuttal has helped open the conversation enough that a better balance might be struck between the wide and varying reviewers and the authors intent. Good job overall!

PB members should use the reviews to better understand a paper, but then they should decide about it. At least, there should be a real discussion between the PB member and the second reader at the meeting. PB members should also be allowed to ask for a further review if the paper is very debated, the expertise is low, or a review is not good enough. I would really like to have three people at the PB meeting that can discuss a paper, but I am not sure I have a concrete idea on how to implement it.
reduce length of online discussion
1. PC was too large it would seem, since there was definitely a feeling that some PC members might have been subpar. Picking a smaller PC with only people you know (maybe not possible) would help. 2. Better defined deadlines for PB actions. 3. More time between the PB reviews and the meeting so that reviewers can still respond to the PB opinions. 4. Should probably allow PB members more say in the final decision, since ultimately they do have some say anyway, This with #3 above that gives the reviewers and PB more time. 5. Papers that nobody can review is out of scope, end of story!

I think the key to the success of this model is to how to mediate people who are taking different roles. I think I brought it up in prep for the review process that we need different sections to (a) communicate only for PCs, (b) communicate only for PBs, and (c) communicate for authors. We had (a) and (c) but not (b). And some of the PB members used "discussion summary" section for the dual purpose for (b) and (c). I was not able to access the discussion summaries of other PB members during the PB meeting, which I wish I could.
Have access to the discussion summaries other than those that we wrote. There were a few PC members who had disappointing (lack of content) reviews. And a few people still had "co-reviewers", who wrote really harsh reviews (and sometimes unprofessional reviews) that the PC member should have gone through and cleaned up.

I had a strong sense that the proximity to the end of year/Christmas holiday/finals somewhat reduced the amount of online participation. If at all possible, having the online discussion occur earlier or later might be beneficial.
The main issue I see, but I don't know how to solve it, is that this model weakens the role of champions/detractors. In traditional PC meetings, I've seen people strongly fighting in favor or against a paper (well, we may argue this may or may not a good thing), while for obvious reasons this did not happen in the PB meeting.

See challenges above. Probably two readers at the meeting are too many, because this reproduces a situation of 2 reviewers that easily override the previous reviews and discussion. The expertise at the meeting should be mostly used to disambiguate situations instead of overriding sometime very expert reviewers.
* For ZZZ / ZZY set of reviewers a 4th reviewer is needed. The long time between paper submission (September) and presentation (June) slows down the entire field. * Either: Shorten the review period by 2 months by having 1 reviewing round only (and a larger PC) * Or: send out 2-review rejects immediately. * ICSE needs some stability: It is a pity that the 2015 scheme is a bit too different. * There were a few papers in which a rebuttal could have helped to address reviewers' concerns. I'd be in favor of having a rebuttal phase. * Cyberchair is too outdated, and does not support this process. ICSE should take a stake in a long term solution for this. * Discussion summaries should not be called discussion summaries. Often they are long, hopeless, and hard to summarize. Instead, they should reflect the single one reason why the paper got accepted/rejected, and actionable advice on what needs to be done with the paper. These should be different fields: (1) Key reason for decision; (2) Key suggestions.

The order of papers in the PB meeting somewhat confused me. Sometimes, I would like to know the rationale behind the order. On the other hand, I think we were less biassed compared to if we explicitly were told that "we handle the papers in decreasing rank" or something.

I am not sure the PB meeting is necessary. (And I am not sure this is good or bad!) Great job, gentlemen!

Program Committee comments (Verbatim)

Advantage: democratic, and I like the emphasis on acceptance and a more positive attitude. Challenge: sometimes, discussions are lengthy and are "too democratic", with the PB member refusing to jump in with opinions (leading to longer discussions than needed).
Challenge: Getting the paper in the hands of area experts. I felt quite a few papers have not been assigned to individuals who understand the breath and depth in the given subarea of software engineering. Challenge: Significantly different reviewing criteria between the reviewers. These range from some (many) reviewers finding the slightest possible reason to reject, to others who advocated very weak papers. Advantage: Larger program committee and a clear attempt to improve the process. Advantage: Two stage consideration, PC and PB. Advantage: Careful coordination of reviews by PC chairs. Given the circumstances, job very well done!

Advantages - lighter reviewing load - better oversight of reviews and discussion - review summaries give authors a clearer picture of why their paper was accepted and how it could be improved - I thought the identification of paper categories was useful, because no paper can satisfy all expectations; I would be tempted to think that it is conducive to creating a more rounded and diverse program Challenges - interactions with PB member (to, e.g., revise reviews or provide feedback on review summaries to be sent to the authors) increased load somewhat (but only to an acceptable, worthwhile degree)
It's unsatisfying to do all the work of reviewing papers but then not be present when actual decisions are made in order to argue my points for/against a paper. It looks like good decisions were made, but I was not at all confident that this would be the case.

Most people who actually read and reviewed the papers didn't participate in the in-person meeting to make the final decisions.
I think that the program board can be eliminated and all decisions can be made after the online discussion, like many conferences do already
Personally, I think the division of a PB and PC takes power away from the PC. I worry that final decisions on some papers are made without the input from PC members that reviewed the papers. It relegates PC members to little more than reviewers, which is a shame, given the expertise available.

The challenge was just the sheer size of the task, but in terms of the number of papers and the fact that the process didn't end with submission of the review. So it was a big commitment. The advantage, though, I would hope is less random final decisions. I think the quality of the review process was definitely higher than a simple one-cycle review process.
Lack of expertise on a few papers. Not just me, all reviewers as well as PB member.

I felt the editorial board members did not have the same grasp of the paper contents and issues as the reviewers. It is hard for the reviewers to make decisions for on the fence papers without seeing the whole set of papers on the fence, being accepted, being rejected to make sure they are treated similarly to others. I felt like I was asked to make accept/reject decisions out of context of other papers in that situation and not enough information to compare.
Pros: a modest number of the papers to review due to the large number of PC members

I enjoyed the online discussion phase because I felt that it was an opportunity to further discuss the paper with fellow reviewers but we did not have to immediately make a decision. I liked that we had a chance to discuss the merits of the paper without having to constantly think about a vote up or down. The PB members did a very good job (for the most part) not to force us into a consensus. I also liked that the number of assignments were reasonable. Thank you for running a great process!
I thought it worked very well from a reviewer's perspective.
+) scale in number of papers, reviewers, expertise coverage by large team -) missing feedback loop for PC: the final decisions were not conveyed to the PC by the PBs

The combination of Program Board and Program Committee looks an interesting process to ensure that the reviews are thorough and also being closely monitored. This certainly helped in increasing the quality of reviews and overall effectiveness of the review process. All board members were active and gave constructive feedback to reviewers.
Advantage: perhaps it was more efficient in terms of PC chairs' resources this way.
Disadvantage: results of the PB discussion were sometimes surprising (sometimes followed the recommendation of the PC and sometimes surprisingly not).

Advantages: Understanding the other reviewers' perspective. Careful moderation by colleagues who were seeking consensus. Challenges: The process doesn't really provide a way to address fundamental clashes - and the moderators tried to maintain a relatively neutral stance for too long. The process implies that all reviewers are 'equal', when there are times when that is not true, and it's up to the committee to recognise those times. The process drags on too long. We were asked to reiterate positions already stated. I found it uncomfortable to be asked to alter my considered position (although I respect the committee's right to disagree with my judgment), when actually an editorial decision was required. The Programm board needs to take a stronger role - after all, the committee should be providing the editorial direction and priorities. The biggest threat is some reviewers are looking for an almost perfect paper which militates against something very innovative but perhaps not perfectly executed being accepted. Conversely there's a tendency for something safe and boring to be accepted. This isn't the result of the new process but I don't think it helps much.

I did not participate in the old review process, so... In what I did experience, I was surprised at the various kinds of negativity presented by some reviewers. I was very concerned that some reviewers rely on their mere skepticism (i.e., a distrust of the paper's presentation), or their disagreement with the approach (i.e., its not the approach they would prefer to see) as a basis for rejection. I also found that PB members frequently left the responsibility to the champion to defend the paper as a disincentive to be a champion. Why don't the PB members challenge the detractors to describe what needs to be done in specific terms to fix the paper and make it acceptable? I only recall one PB member, Jo Atlee, who took this step without prompting. I had to encourage other reviewers to update their reviews with specific suggestions to fix papers, which is a natural role for PB members. The role of the PB member, in my opinion, is to identify misunderstandings on the part of reviewers -- which does happen for all of us -- to clarify ambiguities in reviews, and to keep us fair and balanced. Some PB members were too disengaged, in my opinion. Underpinning the review process is a set of recurring themes that determine acceptability. These themes must be made explicit during discussions, but this is difficult. I tried to raise this in my review of one paper, but the PB member appeared to ignore this point and instead push for a decision. See my discussion of the risks to our community of requiring comparative evaluations for new approaches; did that ever come up during the PB meeting? If so, I didn't hear about it. PB members should reconnect with their reviewers after the PB meeting for contentious papers or where more foundational concerns were raised. One paper, for example, had four reviews -- two in favor, two against -- but what happened when that paper was discussed at the PB meeting? All in all, the model is a very good model, but ICSE should always have at least two junior members on the PB every year. I nominate XXX or YYY. They're both good eggs.

The different categories of paper are a welcome addition. I'm not sure reviewers are all equally experienced with these categories, however. For example, an empirical paper can be exploratory research in which case the outcome is a validated description of the state of practice. Some reviewers conflate their opinion with an empirical finding (e.g., "this result is obvious, why should we accept this paper!?"), in which case they discount the value of such results. Such papers could easily be rejected without experienced reviewers who guard their opinions.

I think the new process is very efficient. I found it brilliant, very effective, and I think it is mainly due to the professionalism and timeliness of the program chairs as well as the relevant contribution of the PC members. People are crucial to make it successful.
All depends on how the program board member acts. If she/he is active, the model is very good.
Main advantages: - the discussion time has been long, producing a good amount of constructive comments. Challenges: - quite tought to keep a focussed discussion in a so long period. Since the discussion window was very long, in many cases I had to read again the all reviews and previous discussion, to get back to know where I left it previously - in some weeks, there have been a lot of papers to discuss, with a big amount of effort.

While program boards can monitor the review quality and provide constructive directions to improve reviews, I believe that it adds another level of indirection and seems to take the responsibility and opportunity for program committee members who actually read the paper to discuss with other PC members.
I'm a newbie to ICSE reviewing, so can only comment that the process appeared to go smoothly. Some variability in the degree of shepherding of the discussions by the board members - some proactive, some not, otherwise I was happy with how it ran.

The quality of the reviews was clearly lower this year than previous years. While most reviews were high quality, there were many more poor reviews than in the past. I believe that because the PC members did not attend a meeting, and did not have to stand up for and answer for their reviews, some felt less responsible and put in less effort.

Advantages: - Not having to travel - Larger PC; less clicks Challenges: - There was no insight into what went on at the PB meeting. - It was awkward to cut the people that knew most about the paper (i.e., PC members that read the paper) out of the decision making process for borderline papers - The paper category was not considered by most reviewers, probably because they were somewhat rigid and confusing to begin with

No need to make a long trip and to attend full two-days meeting. This is a great advantage. Some PB members worked hard, and we got a good consensus.
Advantages: - scales very well - program board members can initiate and shape plenty of offline discussion (something that is quite a burden for PC chairs faced with hundreds of submissions) Disadvantages: - reviewers may not be as accountable or committed as if they show up for a physical meeting - program board members may not be as committed as if they were reviewers - process makes the review period even longer (IMHO, 4-5 months is unacceptable)

I think the main disadvantage is that too many papers are left undecided and it is essentially up to the PB member, who does not have the time to read each paper, to defend or to shoot down a paper. Dont recommend for future ICSEs.
As PC member (not PB) I did not get to witness the PB meeting, but from the point of view of interacting between the PC and PB, it was all ok. I like the idea. I think the main challenge is to make sure that the expertise of the PB is representative enough. Maybe the PB could be selected or updated after the submissions come in? For example, if 50% of the submissions are on testing, then maybe the PB should be updated to have 50% testing people.
This was my first time on the PC, so I am unable to comment on this.
Advantages: the time and opportunity available to debate and discuss, both at a fine-grained level (e.g., reviewers on a specific paper), and on a coarse-grained level (e.g., with PB members and chairs). The two-phase review process (papers with sufficiently supportive reviews received a third review, others did not). Disadvantages: the length of time available for the entire review process meant that some discussions drifted. Most reviews still seemed to be done at the last minute.

The communication limitations of the online discussion will always be there, but for the most part it went well. The schedules of everyone impact how quickly an online discussion proceeds, which can make it hard (a flurry, then a long silence). Overall the outcomes were good, I thought.
The online discussion allowed me to take the time to look back at the papers and internalize others' comments more as well. That's something that there simply isn't time for in the PC meeting. It also made sense to accept/reject papers where everyone was in agreement without needing to physically discuss it. I thought that was a big advantage. One main challenge was that as a PC member and not a PB member, it was unclear to me what would happen at the PB meeting for the papers that didn't achieve consensus since only one PB member was in charge of the paper and it wasn't clear to me that each PB member actually read each paper. So by my understanding, at most, one person at the physical meeting had read each paper under discussion. How did that work? There just wasn't much transparency about the process to the PC.
reduced reviewing load moderated online discussion involvement of PB members

Main advantage: The number of paper we have reviewed this year. Less than others. Main advantage: No need to attend the PC meeting Main disadvantage: The role of the PB members. I felt less in control of the decision that the other times I have attend the PC meeting. A few times I have felt the PB member had his own opinion, not always based on what the reviewers have said.
Overall, I think this new process is considerably better than the previous system. I really liked the enabling of online discussions and settling the bulk of the papers through the online discussion. I think that is a great idea and much fairer for the authors since we (as reviewers) have more time to have a discussion. In some prior ICSEs, the online discussion was never enabled and we only had a few minutes during the PC meeting for such discussion. So in short I think this is super great and even greater that we are taking the decision right there online in the comfort of our own homesoffices ;-) The expanding of the PC is a great idea. a) it makes ICSE feel more inclusive with more people being part of the process b) it ensures that there are more people qualified (the right expertise) to review papers.

(Program Board issues) 1. It was not clear what are the roles of the PB members. In particular, it was not clear even from the instructions that that papers accepted through online discussion cannot be over ruled during the PB meeting. I think that is a good thing (ie., that they cannot over rule PC decisions). I just think the instructions were not too clear on this point. 2. I was not clear on what exactly did the PB have to do during the online discussion. For papers that are clear accept it seemed like an easy job. But for papers that generated lots of discussion, I would have expected the PB member to step in and actually read the paper. In the papers I was on, I did not feel that happened. Same holds for the 4th person (the PB reader) assigned to the paper. It just was not clear what was the value of having two additional people just reading the reviews and asking the reviewers questions about the reviews. I think the PB should probably be reading the papers. This bothered me especially since the PB member and reader

were supposed to be the ones who are presenting the paper's case at the meeting - I find it alarming that in many cases neither of them might have read the paper at all. (Cyberchair issues) 1. I wish cyberchair would send an email when a new review shows up instead of requiring us to go in and check it. 2. Would be nice if one can just reply to the discussion update emails and cyberchair can just attach it to the discussion instead of requiring us to log in. (communication issues) I am not on twitter..I am too old ;-) And there were updates being sent on twitter without the PC members being aware of them. Might be worthwhile to make sure that PC members are notified along with the twitter sphere just so it looks like we are aware of things (e.g., how many papers have been accepted ahead of the PB meeting - this was posted on twitter only)

Advantages: + allowed for reasonable reviewing load + didn't need to travel to meeting + majority of papers can be decided ahead of time Disadvantages: - the most difficult decisions were left to program board meeting, but the main reviewers are absent at that meeting

PROS - the PB members who incrementally check the quality of the reviews while they are entered into the system could have (potentially) improved the quality of the reviews (like the adoption of code inspection makes programmers be more careful with their code simply because they know that their code might be inspected) - the online discussion gives to the reviewers plenty of time for going back to the paper, checking the paper according to comments from the other reviewers, thinking some more time about the paper, etc PC discussion has time constraints which might affect the quality of the decision - the PB summary is useful feedback for authors to understand the main reason for rejection/acceptance - no need to travel for discussing the papers - good matching between expertise and papers
CONS - several reviewers (and PB members) were not responsive enough in the online discussion, even if the online discussion phase has been quite long - the accept/reject decisions taken by different groups of reviewers (i.e., the PB member + 3 reviewers) could be based on very different expectations in terms of quality of the papers, empirical results, etc. than the accept/reject decision taken by another group of reviewers - it is unpleasant to review a paper but not being part of the group who will take the final decision (I refer to the Undecided papers). Even if I acknowledge that the satisfaction of the reviewers is not a major objective of a reviewing process and it makes perfect sense to make reviewers less happy if this implies taking good decisions - different PB members have different attitudes. Some PC members really pushed for being positives, e.g., keep instilling the idea that every problem could be potentially fixed in the camera ready so the paper could be accepted, and some others do not behave in the same way. This could have influenced the decisions. On the contrary in a standard PC the way the papers are discussed is influenced by the chairs who overlook and manage the discussion of all the papers, thus likely not introducing big differences in the way the papers are discussed. I am sure the chairs thoroughly monitored the process and make sure that unbalanced decisions were not likely to happen, but the reviewers got little direct evidence of this. For instance having the chairs participating to the online discussion of all the papers

(although it is very time consuming) would have produced a better feeling that the decisions have been taken in a balanced way.
The larger PC allowed for more expert reviewers for each paper. The online discussion, lead by the PB member was very useful with many relevant issues coming to light.

The best thing this year was external to the new process: Andre and Lionel's continual reminders to stay positive and look for good contributions. As the for the system, the main advantage is a more reasonable review load, which means more time can be given to each paper. There are two fixable downsides to the new system. First, the software is clunky, particularly the discussion features. Second, the discussion period was MUCH too long, which meant that discussion was bursty and I kept losing mental context for each paper. I think the clunky-ness and bursty-ness create real disincentives to the online discussions.
This is my first time as an ICSE reviewer, so it is difficult to compare with previous ICSEs, but as compared to other conferences, I felt that the PB members I worked with did a nice job of shepherding discussion and so the online discussion was better organized than I expected.

+ reduced workload, allowed me to spend more time on each review and on discussion + my assignments were more targeted to my interests and skills - Penultimate PB/PC discussions were desultory and spanned an unreasonable amount of time. Responses ofte

I think there still the issue of having reviewers with the right expertise for the papers. In more than one instance, the other reviewer was not an expert in the topic of the paper. Similarly, in more than an instance, I was not an expert in the topic of the paper. In both cases, I believe the authors might have had a not so fair evaluation.
Advantage: Saving reviewing effort by having two reviews per paper in the first phase.

I think that the review process was very well adapted to the number of submissions. Two reviews are generally enough to sort out terrible papers, and a third review fosters discussions for potentially acceptable papers. In the end, I was happy with the selection made on the papers I had read. I was also happy with the level of involvement that the PC chairs and the PB members had. In the end, I had the feeling that the organization was flawless.
It saved me a lot of time and money not to have a physical meeting. I really appreciated this. I really don't want to travel just because of a review meeting. My review load was acceptable. The review board members did a good job in moderating the reviewer discussion. The instructions given to the reviewers were clear.

Main advantage: using a layer of PB members facilitates the discussion among PC members, by reducing the load of the PC chairs.
I didn't know the previous one. I would say this is efficient, but strongly dependent from the willingness of Program Board to push and control the quality of the reviewing process. They did a great job!

Advantages: The thorough online discussion, albeit a tad too long, definitely helped making informed decisions on (most of) the papers I reviewed. In fact, I must confess I was way more comfortable than I thought I would have been making some decisions online. Challenges: I found the new review process a bit frustrating, as I invested a great deal of time reviewing and discussing papers without having a chance to defend my opinions at the PC meeting--a part of the process I truly enjoy.