Report on the Technical Track of ICSE 2017

Alessandro Orso and Martin Robillard
Program Co-Chairs

Thomas Zimmermann
Data Chair

28 August 2017

Version 1.0

Contents

1 Introduction

In elaborating the ICSE review process we revisited many of the process parameters and, with the approval of the steering committee, settled on the following decisions. We made these decisions based on our experience as members of the ICSE reviewer community and the data available in the chairs' reports of ICSE 2014–2016. We revisit these decisions in the light of experience and discuss their impact in Section 7.

After countless projections and simulations, we decided to **retain the existing board model** despite its known imperfections. The ICSE bylaws require that the final decisions be made in a face-to-face meeting, and none of the alternative scenarios we explored revealed any substantial opportunities for improving the cost/benefit balance of the current model. The details of the process we followed are reported in Section 2. One of the main policy differences in our implementation of the model is that we **kept the identity of the evaluation committee members hidden from one another** throughout the entire process. The motivations for this decision were mainly to (1) avoid any potential for bias due to status or social relations between reviewers and (2) decrease the opportunity for side communication channels related to the evaluation of submissions. A second policy difference is that we **introduced the use of structured reviews** for evaluating submissions. The motivation for using structured reviews was to help clarify the expectations and unify the evaluations styles across a large pool of reviewers. We also publicly released the reviewing guidelines as part of the Call for Papers.[1]

Although we judged the basic mechanics of the existing reviewing process to be adequate for a steady state of submissions, we had serious concerns about the **scalability** of the process. The number of submissions to the ICSE technical track reached 530 in 2016, which constituted a 17% year-over-year increase, which coincidentally happens to be the same increase over the average of the previous 5 years. At the same time, the workload for ICSE reviewers historically centered around 20 papers, despite increases in the size of the program committees. In the more general academic context, some critiques of the randomness of review processes for conference submissions were becoming visible in various media. The collection of these factors made us doubt the ability of the current ICSE review process and pool of qualified and available reviewers to support the reliable evaluation of an unbounded number of submissions. After observing from the ICSE 2015 and 2016 submission data that the submission per author metric followed a so-called power law, we proposed **a cap of three on the number of submissions per author**. After extensive debate within the ICSE Steering Committee of the merits and limitations of this policy, it was adopted by majority vote and integrated into the ICSE 2017 Call for Papers.

A secondary decision regarding the call for paper is that we **did not include an explicit list of research topics** in the document. One general concern with the inclusion of a list of topics for a conference is that it may appear biasing or be an inaccurate reflection of the current relevant work in the field. However, our main deciding factor was the additional realization that the goal for the topics list on the call for papers is different from that of the

[1] `http://icse2017.gatech.edu/technical_research/reviewing_guidelines`

one used to effectively classify reviewer expertise. Instead of managing two inconsistent taxonomies of the field of software engineering, we focused on the expertise mapping for the evaluation committee, which is only visible to the authors at time of submission.

As in recent instances of the conference, we incorporated a **data collection** element into the review process. Based in part on the controversy generated by the cap on the number of submissions, we decided to step up the community component of the data collection effort and include a **post-submission survey** to study the demographics of the ICSE community. To help in this effort, and with the approval of the general chair, we recruited a **data chair** to join the organization committee of the conference. The data chair is a co-author of this report. As part of the process-monitoring effort, we also implemented an anonymous **peer evaluation** or review quality. The results of this exercise are summarized in Section 6.1.

Historically, ICSE had used the CyberChairPRO conference management system, with the exception of ICSE 2012. For ICSE 2017, we decided **to switch to EasyChair** to build on our experience with the system, to benefit from the additional flexibility offered by the system, and because the professional version of the system is licensed to ACM-sponsored conferences. However, to support the special board process model described in Section 2, it was necessary to order a special plug-in for the system.

Since around the spring of 2015 there has been discussions in the ICSE community of moving the conference to a **double-blind model**. However, ICSE 2016 retained the single-blind model, and we decided to retain it as well for an additional year to focus on the transition to EasyChair and the other initiatives we implemented.

2 Reviewing Process

This section describes the reviewing process from the time of initial submission until the time of final decision.

2.1 Overall Principles

The ICSE 2017 process followed the two-tiered **board model**, in which a *program committee (PC)* reviews papers, and a *program board (PB)* generally helps coordinate the reviewing and meets in person at a *board meeting* to arrive at a set of final decisions. The board model was first adopted for ICSE 2014 and also used for ICSE 2016. The responsibilities of the program board vary slightly between instances of the process.

Some of the minor variations we implemented for 2017 were aimed to ensure that there would be at least three members of the PB able to discuss each paper considered at the PB meeting, and that authors would have a chance to respond to additional reviews submitted after the rebuttal phase (in case these additional reviews introduced new elements that changed the overall sentiment for a paper).

In the process, PC members do the main part of the reviewing, whereas PB members play three roles:

Soundness	Claimed contributions should be supported through the rigorous application of appropriate research methods. The claims should be scoped to what can be supported, and limitations should be discussed. Note that a score of "Sound" should also be used for submissions that only have very small issues easily fixable through editing.	Flawed; Major problems; Minor problems; Sound
Significance	Contributions will be evaluated for their novelty, originality, and importance with respect to the existing body of knowledge. Submissions will be expected to explicitly argue for the relevance and usefulness of the research and discuss the novelty of the claimed contributions through a comparison with pertinent related work. Note that your assessment should take into account how well significance is explained and argued in the paper, taking into account appropriate discussion of the related work.	Done before; Incremental; Mostly new; Completely new
Verifiability	The evaluation of submissions will take into account to what extent sufficient information is available to support the full or partial independent verification or replication of the claimed contributions. This rating should take into account the nature of the work and the space limitations imposed on submissions.	Leap of faith; Some cross-checking possible; Results partly verifiable; Results mostly verifiable
Presentation quality	Submissions will be expected to meet high standards of presentation, including adequate use of the English language, absence of major ambiguity, and clearly readable figures and tables. Please do not count redacted parts of the paper as presentation problems.	Terrible; Major problems; Minor problems; Impeccable
Overall evaluation	There is no exact formula for combining the above criteria: your overall evaluation should be based on your assessment of their relative importance for each paper, as the case warrants.	Strong reject; Weak reject; Weak accept; Strong accept
Level of expertise	Please indicate your level of expertise on the topic of the paper.	Expert; Knowledgeable; Informed outsider; No expertise

Figure 1: Excerpt of the ICSE 2017 structured review form.

Overseer: Moderate on-line discussions about submissions under review;

Reviewer: Review submissions for which a strong consensual decision does not emerge from the PC;

Discussant: Read additional submissions discussed at the PB meeting, the reviews for these submissions, and the corresponding discussions.

From the standpoint of the authors, each submission receives at least three reviews, all submissions gets a chance at a rebuttal (and at an extra rebuttal when applicable), and all submissions receive a summary of the reviews and discussions.

One novelty that we introduced in the reviewing process involves the review form. We used a structured review form that, besides the usual level of expertise and overall evaluation, required reviewers to also provide a score for their assessment along four additional dimensions: Soundness, Significance, Verifiability, and Presentation quality. Figure 1 provides the details of the ICSE 2017 structured review form including the structured evaluation criteria, their definition, and the possible ordinal scores for each. This format for structured reviews had been successfully used as part of the evaluation process for the 31st IEEE International Conference on Software Maintenance and Evolution (ICSME 2015).

2.2 Phases of the Process

Figure 2 illustrates the different phases of the reviewing process, and Appendix A details the process timeline. After the submission deadline, the program chairs inspected all submissions to identify submissions unsuitable for review. These submissions were *desk*

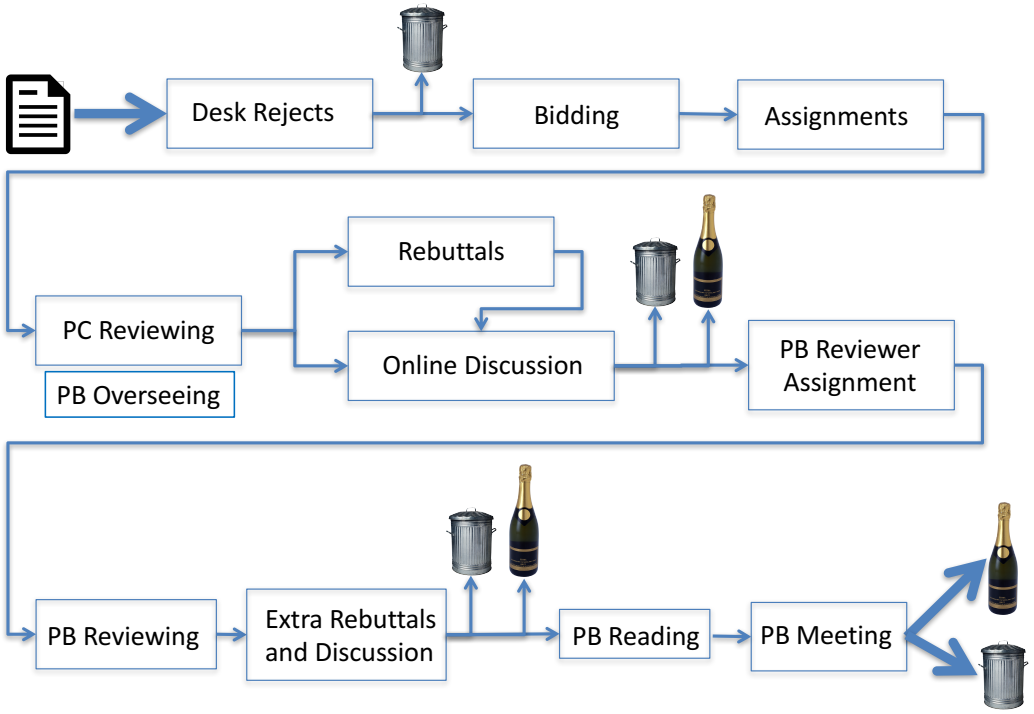

Figure 2: Phases of the ICSE 2017 process.

rejected, the corresponding authors notified, and the submissions removed from the conference management system.

The PB and PC members were then asked to submit *bids* on the remaining papers, during the *bidding phase*. Using the bids provided by PB and PC members, and based on the members' expertise, the chairs created reviewing and overseeing *assignments*. Specifically, the chairs assigned submissions for PC members to review and for PB members to oversee.

During the *PC reviewing/PB overseeing* phase, PC members reviewed papers assigned to them, while PB members were overseeing the process and moderating discussions. After all reviewers submitted their reviews, the PC Chairs sent the reviews to the authors, who were given a chance to submit a response during the *rebuttal* phase. In parallel to this phase, PB and PC members continued the *on-line discussion* about the submissions, which further continued after the responses were submitted by the authors. At the end of the on-line discussion, the chairs made one of three decision for each submission: *accepted by PC*, *rejected by PC*, or *undecided*. During the *PB reviewer assignment* phase, for each undecided submission the chairs assigned a PB reviewer who was to review the submission.

In the subsequent phase, *PB reviewing*, PB members reviewed the papers assigned to them, which led to the *extra rebuttals and discussion phase*. The main activity in this phase was a further discussion of the undecided submissions. In a few cases, when the additional PB review introduced new elements that changed the overall likely outcome for a paper, authors were given a chance to submit an additional rebuttal to address such new elements.

Also in this case, at the end of the online discussion, one of three decisions was made for each undecided submission: *accepted by PC+PB*, *rejected by PC+PB*, or *discuss at meeting*. For each submission in the latter category, the chairs assigned a PB *discussant*, whose task was to read the paper, its reviews, and the corresponding discussion, during the *PB reading* phase and in preparation for the PB meeting.

Finally, at the *PB meeting*, all the submissions that had not been already accepted or rejected earlier were discussed. During the discussion, for each paper, the following process was followed: first, a slide was shown to the program board with the ratings of each anonymous reviewer and their expertise. These ratings included the details of rating for each criteria of the structured review, as well as the overall decision recommendation. The PB overseer was then asked to summarize the paper and the feedback and recommendations of the reviewers, then to provide their personal assessment of the paper; then, the PB reviewer and the discussant were asked to provide their assessment; finally, the paper was discussed and a decision was reached. The final decision was reached by consensus whenever possible, with the chairs exceptionally breaking stalemates.

3 Evaluation Committee

Our review process involved an *evaluation committee* consisting of the two program co-chairs, a program board (PB), and a program committee (PC).

We sought to compose a **program board** with the following characteristics:

- Composed of senior members of the community with significant experience serving on the program committee or board of past ICSE conferences.
- For the board to cover the full range of necessary technical expertise.
- For the board to be inclusive in terms of gender and country of affiliation.

We sought to compose a **program committee** with the following characteristics:

- Composed of members of the community with long-term employment related to software engineering research or practice and with reviewing experience;
- For the committee to cover the full range of necessary technical expertise.
- For the committee to be inclusive in terms of career stage, gender, and country of affiliation.

To help assess expertise coverage, we used a list of 27 topics specifically aimed at partitioning the field of expertise (see Table 1).

We sent 44 invitations to join the **program board**, of which 8 were declined upon invitation. Three board member later requested to serve on the program committee instead of the board, resulting in a final composition of 33 board members from 15 different countries in all continents, and including 9 women and 24 men.

We sent 115 invitations to join the **program committee**. After accounting for declined invitations, post-acceptance drop-outs, and switches between the PC and the PB, the final composition of the PC included 93 members from 24 countries in all continents, and including 18 women and 75 men.

Table 1 shows the distribution of topics and their coverage by the program board and committee, respectively. The data is based on the self-reported indication of members. Two board members and 8 PC members did not enter any topics.

4 Submission Data

We collected data about the submissions from two sources:

1. From *EasyChair*, we extracted meta-data about the submissions such as country of affiliation of the authors and submission topics.
2. Through a *Survey*, we collected additional demographic information not available in EasyChair such as gender, age, and job status. The non-anonymous survey was sent to 1280 authors, 508 authors completed the survey (response rate 39.7%). The program chairs and data chair are grateful to all survey respondents for their contribution to this data collection effort.

4.1 Submitter Population

The data allowed us to compute the following statistics about the submitter population.

Gender identity (Survey). 79.2% male, 19.8% female.

Age (Survey). 5.5% were 18-28 years old, 48.9% were 25-24 years old, 26.9% were 35-44 years old, 11.8% were 45-54 years old, and 5% were 55 years old or older.

Status (Survey). 38.3% students (3.4% undergraduate, 34.9% graduate), 9.8% post-docs, 42.3% professors (14.7% assistant, 15.1% associate, 12.5% full), 3.2% academic researchers, 3.6% industrial researcher, and 2.0% other roles in industry.

PhDs (Survey). 57.9% did have a PhD, 33.3% did not have a PhD but were enrolled in a PhD program, and 8.7% did not have a PhD and were not enrolled in a PhD program. Submitters with PhDs completed their PhDs on average 9.57 years ago (median 7 years). Of the submitters enrolled in PhD programs, 6.4% expected to complete in 2016, 40.4% in 2017, 28.8% in 2018, 14.7% in 2019, and 9.7% in 2020 or later.

Previous ICSEs (Survey). 41.9% never attended ICSE in the past, 18.6% attended once, 13.2% attended twice, 14.6% attended three to five times, and 11.8% attended six or more times. In terms of submissions, 63.3% had previously submitted to the ICSE research track; 11.3% had previously submitted to other ICSE program elements (non-research tracks, workshop, co-located events) but not to the ICSE research track; 22.0% had previously submitted papers but never to any ICSE events; and for 3.4% the ICSE submission was the first-ever submission to any conference.

Table 1: Topic coverage for the program board (PB) and program committee (PC), including both total (-T) and normalized (-N) number of topics. The normalized metric is computed as follows: if a member has selected k topics, each topic adds $1/k$ to the count. The full name of the topic "Collaborative and human aspects of software engineering" includes the suffix ", including education".

Topic	PB-T	PB-N	PC-T	PC-N
Autonomic computing and (self-)adaptive systems	10	1.28	15	1.98
Collaborative and human aspects of software engineering	13	2.12	27	4.23
Components, middleware, services, and web applications	4	0.46	18	2.33
Configuration management and deployment	3	0.31	10	1.50
Dependability, safety, and reliability	11	1.63	21	2.77
Development tools and environments	8	1.22	33	4.40
Distributed, cloud, parallel, and concurrent software	3	0.34	16	2.05
Economics, processes, and workflow	2	0.15	7	1.04
Embedded and real-time software	6	0.73	7	1.00
End-user software engineering	5	0.84	16	2.49
Formal methods	10	1.38	18	2.81
Mining, big data, and recommendation systems	5	0.65	30	4.29
Mobile, ubiquitous, and pervasive software	2	0.15	25	3.58
Model-driven software engineering	10	1.32	16	2.04
Policy and ethics	1	0.07	3	0.53
Program analysis	11	1.81	43	6.99
Program comprehension and visualization	9	1.33	29	4.33
Programming languages	5	0.62	15	2.12
Requirements engineering	12	1.68	12	1.94
Reverse engineering	9	1.21	19	2.49
Search-based and knowledge-based software engineering	9	1.45	15	2.01
Security and privacy	6	0.76	20	2.75
Software evolution and maintenance	15	2.07	42	6.64
Software architecture and design	9	1.18	29	3.86
Software debugging and program repair	8	1.19	34	5.69
Software testing	17	3.20	37	6.22
Specification and verification	12	1.87	21	2.93

Country of current affiliation (EasyChair). The ten countries with the most respondents were: United States (28.8%), China (17.0%), Canada (7.3%), Germany (5.8%), Italy (5.1%), Brazil (3.7%), United Kingdom (3.8%), Singapore (2.6%), Japan (2.2%), and Australia (2.0%). The remaining respondents (21.6%) come from 35 different countries.

Country of undergraduate degree (Survey). The countries where the most respondents received their undergraduate degrees are: China (17.7%), United States (13.6%), Italy (10.5%), Germany (7.0%), Brazil (6.8%), India (6.1%), France (2.4%), and Canada (2.4%). Respondents from the United States received their undergraduate degrees in 22 different countries and respondents from Canada in 18 different countries.

Country of PhD (Survey). The countries where the most respondents received their PhD degrees are: United States (22.8%), Italy (12.1%), China (11.7%), Germany (5.7%) United Kingdom (5.3%), Canada (5.3%), Brazil (5.0%), Netherlands (3.6%), France (3.6%), Belgium (2.8%), and Japan (2.5%). Submitters from the United States received their PhD degrees in 15 different countries and submitters from Canada in 9 different countries.

4.2 Acceptance Rates by Demographics

We computed acceptance rates for subpopulations of the submitters using the demographic data collected through the survey. It is important to remember that the survey only had a response rate of 39.7% and therefore the statistics in this section are based on *incomplete author data*. To account for the incompleteness, we computed a 95% confidence interval for the acceptance rate.

The 94 papers with female co-authors have a lower acceptance rate (12.8% $\pm$ 6.0%) than the 293 papers with male co-authors (18.8% $\pm$ 2.2%). Note that the 95% confidence interval is wider for papers with female co-authors (6.0%) than for papers with male co-authors (2.2%). Therefore we cannot say with certainty that the acceptance rate for papers with female co-authors is lower. In fact, if only the 35 papers with complete author information are considered, papers with female co-authors have a higher acceptance rate of 35.7% vs 19.0% for papers with exclusively male co-authors; however, this difference is not statistically significant and based on a very small sample. Another possible explanation for the different acceptance rates is a possible response bias. Women who responded to the survey belonged less often to the age group '35-44 years old' (18.6% vs. 29.0%) and more often did not have a PhD and were not enrolled in a PhD program (15.3% vs. 7.2%).

As expected having a co-author who attended ICSE multiple times or who had papers previously accepted in the ICSE research track is correlated with higher acceptance rates (22.3% for two or more attendances and 23.5% for previous acceptances, respectively). For the full list of subpopulations with acceptance rates, see Table 4 in the appendix.

4.3 Policy to Cap Submissions

In the survey, 11.9% of respondents stated that they had been affected by the policy limiting the number of ICSE submissions to a maximum of three per author. Of the 42 authors in 2016 with four or more submissions, 4 did not submit, 14 submitted one paper, 10 submitted two papers, and 14 submitted three papers.

In 2017, 1082 authors submitted one paper (85.3%, compared to 81.4% in 2016 and 85.3% in 2015[2]). 136 authors submitted two papers (10.7%, compared to 12.5% in 2016), and 50 authors submitted three papers (4.0%, compared to 3.2% in 2016). The average number of submitted papers per authors dropped from 1.30 in 2016 to 1.19 in 2017. The average number of authors per paper dropped from 3.83 in 2016 to 3.62 in 2017.

5 Process Data

Following the call for submissions we received 415 complete submissions to the research track. In this section we detail the outcome of the evaluation process for these submissions as well as the reviewing effort involved.

5.1 Overall Outcomes

Table 2 summarizes the outcomes for the 415 submissions. Following a detailed inspection, we **desk rejected** 17 submissions that did not meet the submission requirements. In particular, we used the plagiarism detection software CrossCheck to automatically detect submissions with large overlap with other public documents, which led to one desk-rejection on the grounds of plagiarism. We also detected one submission by one author who had remained unaware of the policy to cap submissions at three per person and submitted four papers.

A total of 10 submissions were withdrawn by their authors. In two cases we received the request to withdraw the submission before the official author response period. After we sent the reviews to the authors as part of the response period, we received an additional 8 requests for withdrawal.

5.2 Review Load

In the **initial assignment**, each program committee member received 12 or 13 papers to review, and each program board member received 12 or 13 papers to oversee. As a result of requests for additional reviews from the program committee, the maximum load required from program committee members was increased to 14 for some members. During the PB reviewing phase, program board members were required to review 4 of 5 papers each.

[2]ICSE 2015 reported only the number of authors who submitted one, two to four, and five or more papers

Table 2: Final Outcome of the 415 Submissions to the Technical Research Track

Outcome	**Frequency**
Desk Rejected	**17**
Substance	5
Length	5
Scope	3
Format	2
Plagiarism	1
Submission cap	1
Reviewed	**398**
Withdrawn with Reviews	**10**
Before receiving reviews	2
After receiving reviews	8
Papers with Final Decisions	**388**
Rejected by the program committee	217
Rejected by the program committee and board	49
Accepted by the program committee	18
Accepted by the program committee and board	7
Rejected after discussion at the board meeting	54
Conditionally accepted after discussion at the board meeting	5
Accepted after discussion at the board meeting	38

5.3 Detailed Assessment Data

Table 3 summarizes the overall scores given by the reviewers to the 388 papers evaluated, in terms of ranges of scores. For example (second row), in the category of submissions that received an overall recommendation of least a -1 (weak reject) and at most a 2 (strong accept), 15 were accepted and 14 were rejected.

Table 3: Score Statistics for the 388 Papers with Final Decisions.

Range	**Nb. Accepted**	**Nb. Rejected**
$[1, 2]$	25	0
$[-1, 2]$	15	14
$[-2, 2]$	13	20
$[1, 1]$	2	0
$[-1, 1]$	12	49
$[-2, 1]$	1	54
$[-2, -1]$	0	183

6 Review Process Evaluation

6.1 Peer Review Evaluation

For the first time this year we conducted an anonymous peer-review evaluation exercise for all PC members. After the conclusion of the review process we sent all PC and PB members an email requesting them to supply evaluations on all the reviews for all the papers they had reviewed. For PB members this excluded the papers they had overseen but not reviewed. The respondents were requested to supply, for each of the papers they reviewed, a discrete score as follows:

- 3: EXCELLENT REVIEW: a detailed, insightful, and polished review with little or no room for improvement.
- 2: GOOD REVIEW: a useful review that covers some elements of the paper in details, but with some room for improvement.
- 1: WEAK REVIEW: a review that provides some potentially useful insights, but that is generally incomplete and/or shallow.
- 0: UNACCEPTABLE REVIEW: a review that does not meet the most basic standards for reviewing, and/or that is strongly biased or erroneous, and/or suffers from severe cohesion problems.

We received a total of 1422 valid review ratings from 11 PB members (33% response rate) and 39 PC members (42% response rate). The 93 PC members received on average 13.6 review ratings (min 7, max 25). When converted to an interval scale, the PC members received an average score of 2.12, which corresponds roughly to a "good review" rating. The minimum score for a reviewer was 0.89 and maximum 2.88. A total of 17 reviewers received at least one "unacceptable" rating, and 15 reviewers received only "good" or "excellent" ratings. At total of 63 reviewers received an average score of 2.0 ("good") or better.

6.2 Author Satisfaction

In addition to the peer review evaluation, we sent a survey to each author of each paper asking about their overall satisfaction with the review process and the quality of the reviews. We received a total of 205 responses (response rate of 13.6%), with 1435 ratings for reviews of 157 papers (which cover 39.4% of the 398 papers that were considered for review).

Overall satisfaction. On a scale from 1 to 10, where 1 is not satisfied and 10 is very satisfied, 80% of the authors of accepted papers were satisfied (a score of 6 or higher) and 56% of authors of rejected papers were satisfied. The difference in ratings between authors of accepted and rejected papers has been also observed in previous years.

Review quality For review quality, we reused the scale from the peer review evaluation (3: excellent review, 2: good review, 1: weak review, 0: unacceptable review). Authors of accepted papers were very satisfied with the reviews: 88% of the scores were good or excellent review. Authors of rejected papers were less satisfied, but still the majority of scores was positive: 54% of the scores were good or excellent review.

In addition, we asked the authors to score specific aspects of the review: *accuracy, constructiveness, fairness, thoroughness, usefulness*. When possible, we compare the results to data from previous ICSE conferences. However, it is important to note that the data collection varied over each year, and only subsets of authors participated in the ratings. In 2014 and 2015, the author surveys were post-notification (2014: 185 responses, assuming one response per paper, 37%; 2015: 182 responses, assuming one response per paper, 42%); in 2016 the author ratings were collected during the rebuttal (359 out of 513 responses, 70%). For 2017, we survey was post-notification and allowed each author of a paper to rate the reviews.

Accuracy. The survey participants scored 59% of the reviews as accurate (authors of accepted papers: 81%, rejected papers: 46%). Previous ICSE surveys did not ask about the accuracy of reviews. The closest is whether the reviews reflected sufficient knowledge by reviewers: 58% of authors agreed in 2014 that the reviewers' had sufficient expertise to evaluate their submission, 67% agreed in 2015, and 58% agreed in 2016.

Constructiveness. The survey participants scored 61% of the reviews as constructive (accepted papers: 80%, rejected papers: 50%). Compared to previous years, 64% of authors agreed in 2014 that the reviews were constructive, 64% agreed in 2015, and 57% agreed in 2016.

Fairness. The survey participants scored 62% of the reviews as fair (accepted papers: 84%, rejected papers: 49%). Previous ICSE surveys did not ask about the fairness of reviews.

Thoroughness. The survey participants scored 55% of the reviews as thorough (accepted papers: 75%, rejected papers: 43%). Compared to previous years, 66% of authors agreed in 2014 that the reviews were constructive, 69% agreed in 2015, and 66% agreed in 2016.

Usefulness. The survey participants scored 64% of the reviews as useful (accepted papers: 82%, rejected papers: 53%). Compared to previous years, 66% agreed in 2016 that the reviews were useful, no data is available for 2014 and 2015.

7 Reflection from the Program Chairs

With the benefit of hindsight, we can now comment on the main decisions we presented in the introduction and discuss their impact. Regretfully, we can only ever walk down one path through history, so the consequences of the alternative choices at our disposal will never be known and cannot be compared against. Nevertheless, a number of lessons emerged from our experience.

The Board Process Model. The two-tiered evaluation model is intended to help scale the review process while supporting a functional face-to-face meeting. Although it does help achieve these two goals, our experience is that the board model (as we implemented it) is very challenging to manage. First, there is the inherent complexity of the multiple stages and roles, which are visible in Figure 2 and in the corresponding discussion. Each stage requires involvement from the chairs, communication with the evaluation committee, and careful attention to innumerable details. Each stage introduces potential for confusion and questions from both authors and members of the evaluation committee. Second, it is very difficult, if not impossible, to avoid a certain amount of tension between the two tiers of the evaluation committee. We had experienced this tension first-hand while serving on the board of ICSE 2016. Despite clear awareness of the phenomenon and an explicit determination to foster a spirit of mutual understanding and cooperation between the two segments of the evaluation committee, we noticed evidence of friction during the on-line discussion, at the program board meeting, and through personal communication with the members of the evaluation committee. Finally, we witnessed a lack of shared understanding of their role by program board members. Part of this issue can be attributed to the relative novelty of the model in the ICSE community and to the rapid changes in its definition since it was introduced. We provided detailed instructions to both program board and program committee members and also created a Frequently Asked Questions (FAQ) page in which we summarized the questions we received from individual members of the evaluation committee, and yet the issue remained. Although it is not clear what the most satisfactory long-term solution will be for the evaluation process of ICSE, our opinion is that the current model does not provide value in proportion to the effort it requires.

Reviewer Anonymization. The impact of the decision to anonymize reviewers is difficult to assess. We have anecdotal evidence, in the form of personal communications, that it generated mixed feelings. Among committee members that expressed a negative opinion, the main issues raised were that *1)* reviewers are not held accountable for unprofessional behavior; *2)* conversely, there is less incentive for reviewers to excel since their contributions are not associated with them personally; and *3)* it is impractical and confusing to discuss papers with anonymous agents. Although this third limitation was made more acute in our case due to accidental user interface issues with EasyChair, the anonymity nevertheless hinders the development of a collegial spirit in the discussion of submissions. These issues, however, must be pitted against the very clear benefits of avoiding bias and side communications in the evaluation of reviews. Unfortunately, the benefit of *avoided bias* is not directly measurable. However, we observed many instances of the impact of reviewer anonymity during the discussion period, where completely unabashed and challenging questioning took place between reviewers who were either socially close or in asymmetrical status relations, which are situations that would normally preclude these types of interactions. One benefit of anonymization that is tangible is that it makes the peer-review evaluation of reviews immune to bias, as program committee and board members evaluated the quality of each other's reviews without knowing the identity of the author of

the review (see Section 6.1). Based on our experience with this and other conferences, we believe that anonymization of reviewers is worth doing again, as long as user interface support for anonymous discussions can be improved, and mechanisms are in place to ensure reviewers are more accountable for the quality of their work.

Cap on the Number of Submissions. Compared with previous years, for ICSE 2017 the workload for ICSE reviewers was considerably reduced. Although the precise cause for the lower number of submissions cannot be reduced to a single factor, the cap on the number of submissions provides a tangible safety valve to ensure that the number of submissions remains, at least roughly, in constant proportion to the size of the community served by ICSE. Our estimate, based on the data from ICSE 2015 and 2016, indicated that cap of 3 submissions per authors could have saved around 10% of the reviewing effort (excluding ancillary discussion and coordination effort). Our actual numbers seem to confirm this estimate, with the aforementioned caveat that more than one factor could have played a role in this reduction. Considering all the measures we took to bring the burden on the evaluation committee to what we felt was a more reasonable workload, we were satisfied to have reached a record low number of review assignments per committee member. As we discussed in Section 5.2, PC members (resp., PB members) only had to review (resp., oversee) between 12 and 14 papers, and PB members only had to review between 4 and 5 papers.

Structured Reviews. The use of structured reviews required reviewers to explicitly consider the impact of distinct dimensions of evaluation when reviewing a paper, and to articulate their arguments along these dimensions. The use of the explicit evaluation criteria helped us focus the evaluation of submissions both during the on-line discussion and at the program board meeting. Additionally, the use of structured reviews is very well supported by EasyChair. Overall we recommend the use of structured reviews as part of the ICSE review process.

The EasyChair Conference Management System. EasyChair was able to handle the workload of the reviewing process without any scalability problem. The issues we experienced were of two main types. First, we faced issues with the user interface of the system, some of which we mentioned earlier in this section. We were aware of these issues from the beginning, so we planned workarounds and in some cases requested and obtained patches. The second type of issues was related to the use of the new plug-in that the EasyChair team developed to support our reviewing model. The plug-in had some defects and missing functionality. Before the submission deadline, we conducted a complete simulation of the process using the demonstration feature of EasyChair and were able to identify both malfunctions and erroneously implemented features (due to misunderstandings in the requirements collection phase). The EasyChair team managed to fix all the critical issues before we opened the submission site. Although the final system still had some weak-

nesses, we were able to successfully use the system to complete the evaluation process. Overall we recommend continued use of EasyChair.

Achieving High Standards of Reviewing Quality. Although the majority of the reviews that we received were of high quality (see Section 6.1), ensuring the timely delivery of reviews and *overall* high standards of reviewing was one of the more arduous tasks we faced. Despite proactive intervention by board members and program chairs, including in some cases personal outreach, on the order of 10% of reviewers did not fulfill their commitment with sufficient professionalism by either submitting their reviews (exceedingly) late or by consistently providing unacceptable reviews. We felt that, given our careful selection of program committee members, the total number of reviewers who failed to uphold their commitment was problematic because substandard reviewing behavior directly harms both the conference and the authors. Regretfully, one of the only tools as our disposal for mitigating missing or unacceptable reviews was to further impose on the remainder of the program committee, a solution that raises the question of fairness. Recognizing that personal circumstances can change and that sub-par reviewing is inevitable, our recommendation is to adopt explicit measures for addressing, as early as possible, situations where a reviewers is unable or unwilling to fulfill their commitment.

A Process Timeline

This is the timeline for the different phases of the ICSE 2017 reviewing process, as depicted in Figure 2.

Papers Submission Deadline: August 26, 2016
Desk Rejects: August 27–28, 2016
Bidding: August 29–September 2, 2016
Assignments: September 3–7, 2016
PC reviewing + PB Overseeing: September 8–October 21, 2016
(with the first half of the reviews due on October 1, 2016)
Rebuttals: October 22–26, 2016
Online Discussion: October 22–November 10, 2016
PB Reviewer Assignment: November 11–15, 2016
PB Reviewing: November 16–December 4, 2016
Extra Rebuttals and Discussions:

- Extra Discussions: November 11–December 7, 2016
- Extra Rebuttals: December 7, 2016

PB Reading: December 5–7, 2016
PB Meeting: December 8–9, 2016
Authors Notifications: December 12, 2016

B Additional Data

Table 4: Acceptance rate (AR) by demographic with a 95% confidence interval. N is the number of submitted papers that could be linked to a demographic based on the survey response (response rate 39.7%). The term "ICSE event" includes ICSE tracks, workshops, and co-located events.

Demographic	N	AR
Female	94	12.8% ± 6.0%
Male	293	18.8% ± 2.2%
18-24 years old	34	8.8% ± 9.5%
25-34 years old	204	19.6% ± 3.8%
35-44 years old	147	19.7% ± 5.1%
45-54 years old	68	20.6% ± 8.9%
55 years old or older	35	17.1% ± 12.4%
Graduate student	158	17.7% ± 4.6%
Post-doc	53	22.6% ± 10.7%
Assistant Professor	87	25.3% ± 8.2%
Associate Professor	85	11.8% ± 6.1%
Full Professor	88	21.6% ± 7.7%
No PhD and not enrolled in PhD program	40	10.0% ± 9.1%
No PhD but enrolled in PhD program	155	18.7% ± 4.8%
PhD	259	19.3% ± 2.8%
Never attended	163	13.5% ± 4.0%
Attended 1 time	91	19.8% ± 7.3%
Attended 2 times	75	25.3% ± 9.0%
Attended 3-5 times	94	22.3% ± 7.4%
Attended 6+ times	79	24.1% ± 8.6%
Previous submission experience.		
Never submitted to any ICSE event	95	10.5% ± 5.4%
Submitted to ICSE events but never to the Research track	55	16.4% ± 9.3%
Submitted to ICSE Research track but had no papers accepted	126	15.9% ± 5.3%
Submitted to ICSE Research track and had papers accepted	179	23.5% ± 4.6%
Not affected by 3-paper policy	294	17.0% ± 2.1%
Affected by 3-paper policy	74	23.0% ± 8.8%

Table 5: The number of submitted papers (N), accepted papers (A), program committee and board members (PCB), and acceptance rate (AR) by topics. The full name of the topic "Collaborative and human aspects of software engineering" includes the suffix ", including education".

Topic	N	A	AR	PCB
Autonomic computing and (self-)adaptive systems	14	1	7.1%	25
Collaborative and human aspects of software engineering	58	10	17.2%	40
Components, middleware, services, and web applications	22	4	18.2%	22
Configuration management and deployment	11	2	18.2%	13
Dependability, safety, and reliability	24	3	12.5%	32
Development tools and environments	59	7	11.9%	41
Distributed, cloud, parallel, and concurrent software	12	2	16.7%	19
Economics, processes, and workflow	17	3	17.6%	9
Embedded and real-time software	1	0	0.0%	13
End-user software engineering	18	0	0.0%	21
Formal methods	25	2	8.0%	28
Mining, big data, and recommendation systems	74	9	12.2%	35
Mobile, ubiquitous, and pervasive software	40	6	15.0%	27
Model-driven software engineering	19	2	10.5%	26
Policy and ethics	8	0	0.0%	4
Program analysis	84	18	21.4%	54
Program comprehension and visualization	31	5	16.1%	38
Programming languages	28	8	28.6%	20
Requirements engineering	31	2	6.5%	24
Reverse engineering	15	2	13.3%	28
Search-based and knowledge-based software engineering	31	3	9.7%	24
Security and privacy	43	9	20.9%	26
Software evolution and maintenance	111	15	13.5%	57
Software architecture and design	34	3	8.8%	38
Software debugging and program repair	48	12	25.0%	42
Software testing	75	16	21.3%	54
Specification and verification	37	3	8.1%	33

Table 6: The number of authors (AU), co-authored papers (N), accepted papers (A), and acceptance rate (AR) by country. Only countries with at least 10 authors are shown for privacy reasons.

Country	AU	N	A	AR
Australia	25	10	1	10.0%
Austria	13	8		0.0%
Brazil	48	15	2	13.3%
Canada	93	43	4	9.3%
China	215	69	10	14.5%
France	17	6		0.0%
Germany	73	29	4	13.8%
Hong Kong	16	8	1	12.5%
India	22	11		0.0%
Israel	19	9	2	22.2%
Italy	65	26	5	19.2%
Japan	28	11	1	9.1%
Luxembourg	13	5	3	60.0%
Netherlands	24	15	3	20.0%
Portugal	22	7	2	28.6%
Singapore	34	16	4	25.0%
Spain	14	6		0.0%
Sweden	19	10	2	20.0%
Switzerland	20	12	5	41.7%
United Kingdom	47	27	6	22.2%
United States	367	150	38	25.3%

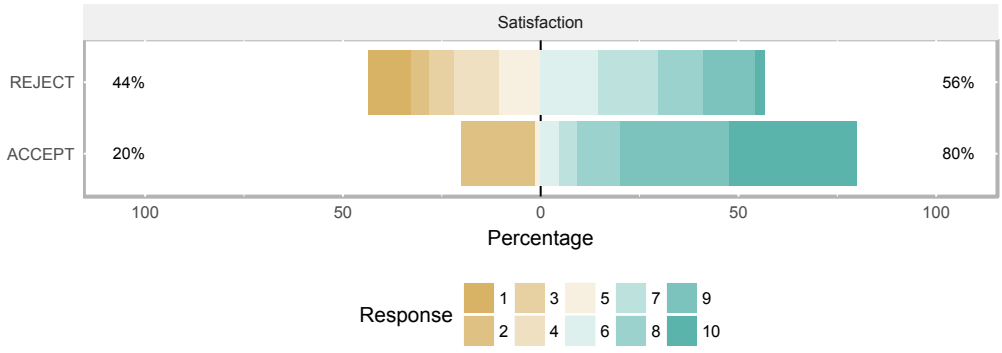

Figure 3: Overall satisfaction with the review process. (*"Please rate your overall satisfaction with the ICSE 2017 review process. The scale is from 1 to 10 where 1 is not satisfied and 10 is very satisfied."*)

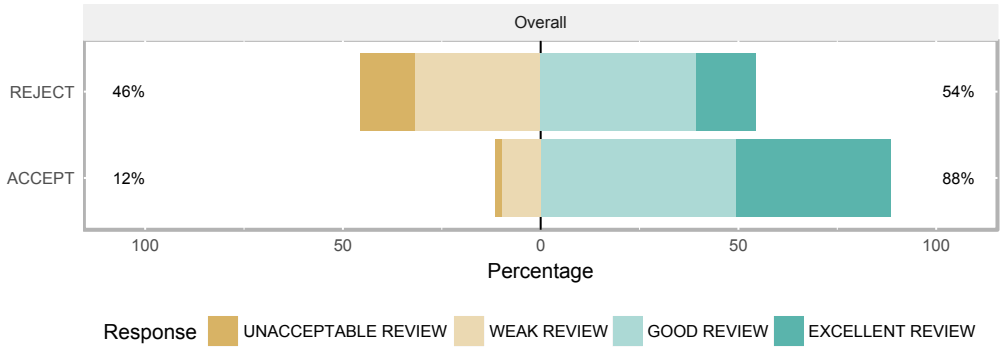

Figure 4: Overall rating of the reviews. (*"Please rate the review by Reviewer n."*)

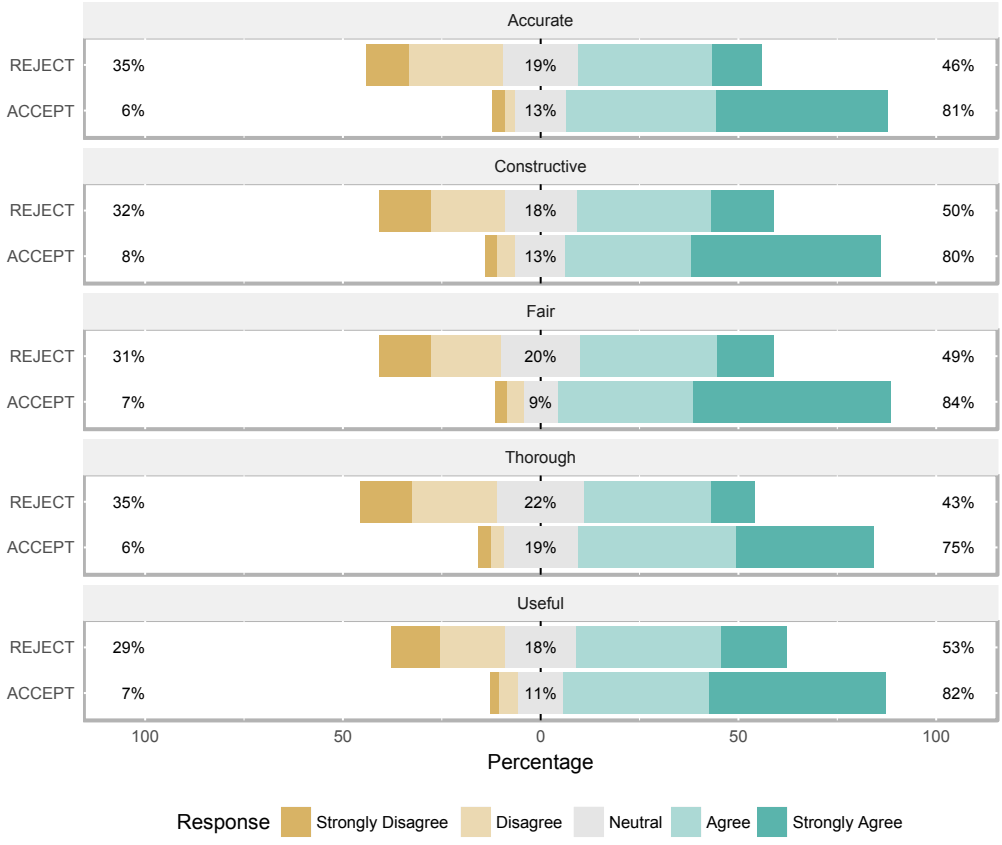

Figure 5: Ratings for specific aspects of the reviews. (*"Please rate your agreement with the following statements about the review by Reviewer n. The review was..."*)