Report on the Technical Track of ICSE 2015

Gerardo Canfora and Sebastian Elbaum
Program Co-Chairs

April 16, 2015

Executive Summary

The goal of this report is to inform the community about how we, the program chairs, organized and managed the technical track of ICSE 2015. The technical track, the most prestigious of our flagship conference, is continuously evolving in an effort to make better decisions, give better feedback to the authors, and more wisely involve the community resources. This year we adjusted the reviewing process to use a model consisting of a traditional Program Committee (PC) and a Reviewing Committee (RC) to get more reviewing expertise and balance the reviewing load, provided early rejection notices, made the process more transparent to the community, monitored more closely the quality of the reviews, used online discussions and summaries to improve the communication between reviewers and with authors, and assigned as many reviewers per submission as possible to make better decisions.

The report provides details on how the technical track was organized (Section 1) and performed based on data collected through the conference management tool (Section 2), a survey of reviewers and authors (Section 3), and our personal experiences (Section 4).

The Technical Track at a glance:

- 452 submissions by more than 1250 authors
- Reviewing Committees
  - 48 PC + 33 RC members from 25 countries
  - 23% new to ICSE technical track reviewing committees
- Reviewing Process
  - $\sim$ 100 reviewing days
  - 1137 reviews + summaries + online discussions ($\sim$100 ICSE submissions)
  - +15% of reviews were revised for anti-patterns found by program chairs
- Reviewing Expertise and Quality
  - Reviewing expertise was higher than previous ICSEs (as reported by authors and reviewers)
  - Authors assessment show slight improvements over last ICSE in terms of reviewing quality, but about 20% of authors still disliked their reviews
  - Authors of accepted papers had a much more favorable view of their reviews
  - 87% of the PC/RC agreed that the final decisions were balanced and justified
- Reviewing Load
  - Max 18 reviewed submissions for PC members and 9 for RC members
  - 90% agreed that the load and schedule was manageable
- Program includes 84 papers (18.5% acceptance) and 6 ACM-Sigsoft D. Awards

Brief Reflections from the Chairs (expanded in Section 4).

- Refine: Process. It received improved grades for reviewing expertise, slightly improved for quality, and controlled the reviewers' workload as planned. But there is still a room for improving the quality and the delivery of the feedback.
- Drop: Panic about scalability of reviewing process.
- Keep: RC and take on new committee members (with proper due-diligence).
- Keep: Early rejection notification.
- Explore: Alternative to bidding process (expensive, very inconsistent).
- Keep: Quality control on reviews (expensive but very effective).
- Keep: Assign as many reviewers as required and available to make a good decision.
- Refine: Use of summaries and categories.
- Explore: Encourage for artifacts/tools to be made available.
- Explore: Alternative conference management systems.
- Explore: Bounding number of submissions.
- Explore: Double-blind reviews.
- Keep: Measuring and communicating how the technical track operates and performs.

We would like to thank our strong team of reviewers, student volunteers, conference system administrator, Antonella Bertolino and her CNR team, and all the authors for contributing to the technical track of ICSE 2015.

Contents

1 Reviewing Process and Committees

The reviewing model for ICSE 2015 is depicted in Figure 1. It involved two program co-chairs and two committees, the Program Committee (PC) and the Reviewing Committee (RC) working in concert. Members of both committees contributed reviews and participated in the online discussions, but the PC members had a heavier reviewing load (17 vs. 9 submissions) and made the final decisions at the physical meeting. The PC comprised 48 members and the RC had 33 members. Altogether, **PC and RC members came from 25 countries, 21% were female, and 23% had not been in an ICSE reviewing committee before.**

After the submission deadline, the process started with a cursory submission review by the chairs for format and scope compliance. This was followed by a bidding process where the reviewers marked the submissions according to their level of interest and expertise. The chairs then performed an initial manual assignment of submissions to reviewers considering expertise, interest, and load distribution. With this assignment in place, PC and RC members completed their first reviewing phase, providing a total of two reviews per paper (every submission received at least one review from a PC member in this phase). Papers with at least one supportive review or without enough expertise advanced to the second phase, the rest of the submissions were rejected and their authors (except for submissions including PC or RC members) received an early rejection notification. During the second phase, PC members were assigned to contribute a third review per qualifying submission.

For each phase, reviewers were given specific instructions and access to different parts of a conference management system to conduct the reviewing process. The system provided reviewers with an evaluation form that included sections for a concise description of the submitted work, a list of strengths and weaknesses, and a detailed assessment. In addition, each reviewer was asked to score each submission with one of four marks: *A - I will champion this paper (Advocate/Accept); B - I can accept this paper, but I will not champion it (accept, but could reject); C - This paper should be rejected, though I will not fight strongly against it (reject, but could accept); D - Serious problems. I will argue to reject this paper (Detractor).* Reviewers were also asked to classify their expertise in the submission topic area according to the following criterion: *X - I am an expert; Y - I am knowledgeable in the area, though not an expert; Z - I am not an expert. My evaluation is that of an informed outsider.*

Throughout the process reviewers were reminded about common reviewing anti-patterns to avoid and the chairs monitored the quality of the reviews (more on this later).

The second reviewing phase was followed by an online discussion aimed at clarifying reviewers positions, sensitizing reviewers to strengths/weaknesses they may have not considered, and making sure RC members get a chance to clarify/emphasize their perspective in view of the discussion at the PC meeting. For each submission, a PC member lead the online discussion and prepared a summary to later guide the discussion at the PC meeting, and to communicate the essence of the discussion to the authors. A fourth reviewer was

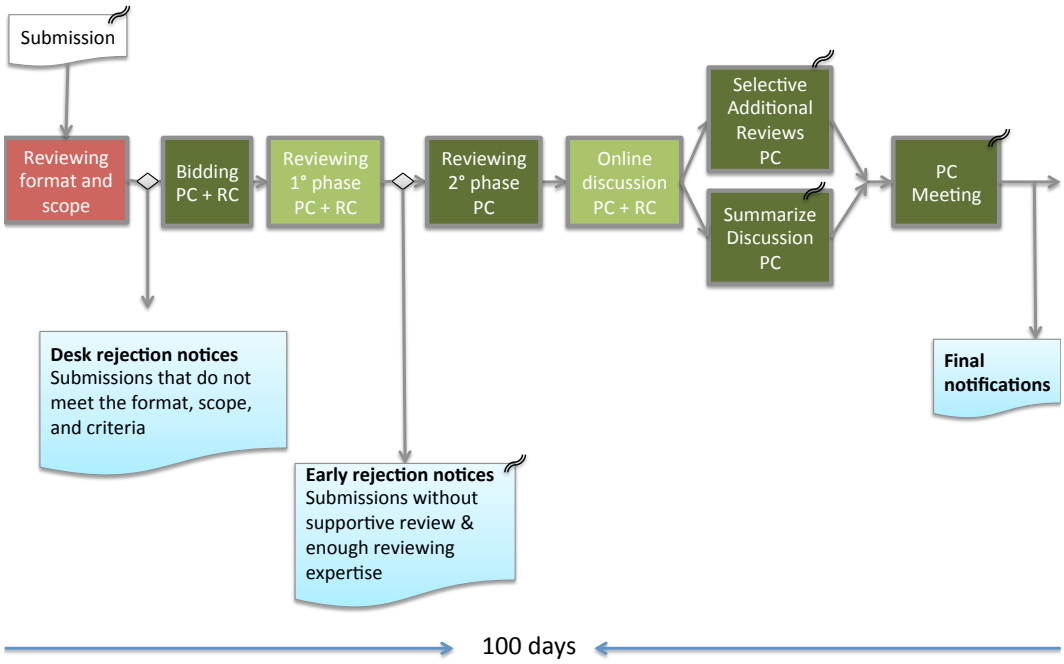

Figure 1: Reviewing Process and Roles of PC/RC Members

added for submissions with diverging perspectives or particular expertise needs. Finally, a physical meeting was held in Pisa on December 12 and 13, 2014, with the PC members and chairs to decide the final list of accepted papers. Authors were notified of the outcome within 100 days of the submission deadline.

This process and some of the most frequently asked questions about it were documented and made available to the community almost a year in advanced at http://2015.icse-conferences.org/call-dates/call-for-contributions/technical-research-cfp?id=18

The timeline of activities from committee selection to the conference is provided in the appendix.

2 Submissions Data

ICSE 2015 received 452 submissions, with an acceptance rate of 18.5%. Table 1 puts that information in the context for the previous five ICSEs. The number of submissions to ICSE averages slightly over 450 in the last five years.

Of the 452 submissions, 20 were desk-rejected or withdrawn before the reviewing process started, and 2 more submissions were withdrawn during the first phase of the reviewing

Table 1: Submissions and Acceptance Rates at last five ICSEs

Year	Submitted	Accepted	% Accepted
2011	441	62	14
2012	408	87	21
2013	461	85	18
2014	495	99	20
2015	452	84	18.5

process. At the end of first phase, 189 papers were rejected based on the two available reviews. By default, all papers scored "DD" were rejected after an examination by the chairs for enough expertise. Papers with scores of "CC" and "CD" were promoted to the second stage of the review process when (i) one of the first-round reviewers asked for an additional review; (ii) the reviewers declared low expertise; (iii) the scores were not convincingly supported by the arguments in the reviews or the reviews did not aligned as per the chairs judgements. 17 additional papers were rejected at the end of the second stage of the reviewing process, before the online discussion started, as the reviews were consistently negative with all scores lower than "B".

Throughout the first and second reviewing phases, the **chairs monitored every review, suggesting changes to over 200 of them over the reviewing process. Most of these recommendations were adopted, and this positively affected the reviewing quality of many reviewers later on, particularly the junior ones.** As an example, here are some of the frequent reviewing anti-patterns (collected by the ICSE community) that we detected and their corresponding fixes:

- "it's been said before" to "It's not clear why this result improves upon *specific citations*"
- "I don't think anyone will use this" to "Unlikely to find practical application because *specific reason*"
- "Paper is unreadable" to "Here are some *examples* of writing choices that make the paper hard to read"

During the second phase of the reviewing process, and in a few cases after the online discussion, **a fourth reviewer was assigned to 35 papers that could benefit from more expertise or presented very diverging views**, specially when the RC member was one of the divergent views. At the end of the second phase of reviewing, a PC member was appointed as a discussion leader for each submission requiring a discussion. A total of 224 papers were discussed online; most PC members lead the discussion between 4 and 6 papers.

Table 2: Decisions per Reviewing Phase

Phase	Submissions
Desk reject/withdrawn	22
Rejected after first phase	189
Rejected after second phase (no on-line discussion)	17
Rejected after on-line discussion	78
Pre-Accepted after on-line discussion (subset discussed at the PC meeting)	31
Rejected at PC meeting	62
Accepted at PC meeting	53
Total	452

The role of the discussion leader was twofold: to initiate and stimulate the discussion and to prepare a discussion summary for each submission, capturing the reviewers' key arguments and points of view. All but one PC/RC members participated in the online discussion, and most participated very actively and helped shape the discussion summaries, together with the discussion leader. This worked fairly well to capture the RC members perspective and the underlying issues worth discussing at the PC meeting. The discussion process resulted in over 1200 comments. We, as chairs, monitored all the discussions and in many cases initiated or stimulated them by posing specific questions or requests to the reviewers. An additional 78 papers were rejected after the online discussion based on both the reviews and the arguments emerged in the discussion. **31 papers were pre-accepted before the PC meeting. These papers had uniformly positive reviews and online discussions. 9 of these papers (plus 3 already rejected) were briefly discussed at the PC meeting to help calibrate the review process, and reviewers were asked if they wanted to discuss any of the other pre-accepted submissions at the PC meeting.** Pre-accepting submissions with strong and consistent reviews saved precious time for papers that needed more discussion at the PC meeting.

The PC meeting entailed the discussion of 115 submissions; out of these, 53 papers were accepted and 62 were rejected. Discussion of submissions at the PC meeting was insightful and engaging, with most reviewers knowing their assigned submissions very well. This was one of the key reasons we favor this model for others where the decisions for acceptance is given to the equivalent of associate editors. The average number of papers discussed by PC members at the physical PC meeting was above 6.

Table 2 shows a summary of the decisions along the entire reviewing process. Overall, out of 452 submissions, 84 papers were accepted for presentation at ICSE 2015, for an acceptance rate of 18.5%. The average load per PC member was 17 papers to review, with a maximum load of 18 papers; for RC members the maximum load was 9 papers.

2.1 Reviewing Scores and Decisions

In this section we discuss how accepted and rejected papers are distributed with respect to the scores assigned by the reviewers. It is important to stress that scores were not the sole element on which decisions were made; the chairs and other committee members checked, discussed, and assessed the presence of convincing arguments in the reviews to support the scores. Table 3 shows the distribution of acceptance/rejection decisions versus scores. Specifically, the first three columns on the left of the table show the number of accepted and rejected papers for seven classes of score ranges, while the three columns on the right provide details on the number of accepted and rejected papers for each individual score combination.

The table confirms the intuition that scores are a good indicator for the final decision when they are consistently and uniformly good or bad; all submissions but one in the A-B range were accepted and all papers in the B-D and C-D ranges were rejected. **Submissions in the other classes of scores needed careful analysis and discussion; pre-accepting and rejecting submissions in the top and bottom classes was effective at saving time at the PC meeting to discuss submissions that needed further discussion.** Of course, this requires that all reviews be checked for consistency and coherence of scores and actual arguments provided to support the scores; this effort to revise the reviews was made by the PC Chairs. The distribution ended up being similar to that of ICSE 2014 (e.g., 7% and 8% of submissions in the AB range, 47% and 51% submissions in the CD range for ICSE 2014 and ICSE 2015, respectively).

2.2 Self-reported Reviewing Expertise

This section discusses the degree of reviewing expertise as assessed by the reviewers' themselves. In general, a higher level of expertise is associated with a higher level of confidence in the judgment provided by the reviewer. Although we strived to maximize the expertise available for every submission, this was not an easy resource allocation problem. First, the sources of information like the bidding is noisy and the list of publications by reviewers may not accurately characterize reviewers' current expertise. Second, we had to balance reviewers' load. Third, even an expert in the topic area may not feel completely confident about all areas of a review (for example, the paper uses a specific statistical technique for data analysis he or she is not familiar with). Fourth, the best potential reviewers sometimes have a conflict of interest. That is why the **manual assignment of submissions to reviewers required multiple iterations for each phase (∼45 hours for the first phase, and ∼30 hours for the second).** One thing that we did different this year is to **tap on the expertise of the committees to help us identify complementary reviewers for the second phase**, in a sense crowd-sourcing the potential review assignments. We received recommendations for approximately 30% of the submissions, most of which were useful for deciding reviewers for the second phase. During the review process,

Table 3: Reviewers Scores and Decisions

Score Range	Accept	Reject	Scores	Accept	Reject
A-B	32	1	AAA	1	0
			AABB	1	0
			AAB	12	0
			ABBB	1	0
			ABB	17	1
A-C	27	16	AABC	2	0
			AAC	7	0
			ABBC	1	0
			ABC	10	5
			ABCC	1	5
			ACC	5	4
			ACCC	1	2
A-D	7	8	AABD	1	0
			AAD	1	0
			AACD	2	0
			ABBD	1	1
			ABD	2	1
			ABCD	0	3
			ABDD	0	1
			ACCD	0	1
			ADD	0	1
B-B	7	2	BBB	7	2
B-C	11	58	BBBC	0	2
			BBC	11	16
			BBCC	0	2
			BCC	0	38
B-D	0	41	BBD	0	6
			BBCD	0	4
			BCD	0	23
			BCCD	0	2
			BDD	0	6
C-D	0	220	CC	0	47
			CCC	0	13
			CD	0	88
			CCD	0	15
			DD	0	54
			CDD	0	1
		10	CCDD	0	1
			CDDD	0	1
Total	84	346		84	346

Table 4: Expertise Reported by Reviewers

Expertise	Reviews
X	600
Y	495
Z	42
Total	1137

reviewers completed 1137 review reports. Together with the online discussion and summaries, they wrote more than 1100 pages, the equivalent of about 100 ICSE submissions. This is a lot of qualified feedback going back to the authors to improve their work.

In terms of reviewer expertise, Table 4 shows the number of reviews by expertise; only a small fraction of reviewers considered themselves a "Z" when reviewing a submission (3.7% of the 1137 review reports produced), which denotes a very good overall level of confidence of reviewers about their judgments.

Table 5 summarizes the distribution of expertise levels across the papers. The table highlights a very high percentage of papers with at least one "X" (83.9% of the 430 papers reviewed) and a minimum of at least two "Y" (97.7%). A large share of papers (45.6%) received at least two "X"s. Finally, a low percentage of papers has at least one Z (7.2%) and no paper was decided with "Z" only expertise.

If getting more expertise on a submission helps to make better decisions, these figures represent an improvement over ICSE 2014, which already performed very well. For ICSE 2014, the percentage of papers with one "X" was lower (77%) and lower was also the percentage of papers with at least two "Y" (96%). In addition, the percentage of papers with at least a "Z" was higher in 2014 (18%), when two papers were decided with "Z" only expertise.

We also observe that there were 36 submissions rejected with 2 reviews and no "X". Ideally we would have preferred to reject submissions early only if there was at least an "X" on it. This was not the case for at least three reasons. First, some submissions were on what we considered the boundaries of the scope of the conference, so experts in those areas were difficult to find. Indeed, for 7 of those 36 submissions at least one of the reviewers was doubtful about the submission fitting the scope of ICSE (e.g., too much focus on the hardware, it seems more like an IS type submission). Second, in some cases we felt reviewers could have been an "X" but may have been conservative in their assessment because they were not expert in all the topics covered, approaches applied, or tools used by the authors. Indeed, we had four cases out of the 36 that explicitly stated something along the lines of "I am not expert in X, however, I feel quite confident of my review ... ". Third, in a few cases the most expert reviewers had a conflict of interest with the authors. In all these cases, however, the reviewers showed enough expertise for us chairs to feel confident enough to make a reject decision at that point.

Table 5: Reviewing Expertise Distribution Across Submissions

Expertise Distribution	Submissions
XXXX	3
XXXY	5
XXXZ	1
XXX	31
XXYY	14
XXYZ	1
XXY	81
XXZ	3
XX	57
XYYY	9
XYYZ	1
XYY	52
XYZ	6
XY	88
XZZ	1
XZ	8
YYYY	1
YYYZ	1
YYY	22
YYZ	8
YY	27
YZZ	1
YZ	9
Total	430

We conjectured that several factors contributed to improving self-reported reviewing expertise over the previous years:

- Compared to conferences prior-ICSE2014, the review process added a Reviewing Committee, in addition to the traditional PC, which provided a pool of broader and complementary expertise.
- Papers were manually assigned to reviewers through several iterations of re-assignment by the program chairs. Bidding data, which was noisy, was used only marginally and reviewers were selected primarily based on their publication profile.
- As described, PC and RC members were involved in the selection of potential reviewers for the second phase.
- A fourth reviewer was introduced when needed and available. The idea was to get as many reviews as needed to make informed decisions. This principle has long been implicitly used for the 2-phase reviews, adding a third reviewer when needed; this year the idea was pushed a step further, assigning a fourth reviewer to 35 papers that could benefit from more or diverse expertise.

To stress that bidding data can be very noisy, Table 6 shows how the "Z" reviews were distributed across the bids; more than 73% of the review reports marked as "Z" were for papers on which the reviewer had expressed a high bid.

Table 6: Low Expertise and Bids

	High Bid	Low Bid	No Bid	Total
Z reviews	31	8	3	42
	73.81%	19.05%	7.14%	100.00%

2.3 Topics and Categories

The topic list was revised to reduce overlap between the topics and the categories (e.g., the all encompassing "empirical" topic was removed as it is now a category). Table 7 shows the number of papers submitted and accepted for each of the topics listed in the call for papers. ICSE 2015 received and accepted many papers in perennial strong ICSE areas of software testing, analysis, maintenance and evolution, and tools and environments. The recent growth in mining software repository research is also evident in the number of submissions in that area. Also the growing concern for security and privacy of software-intensive systems and infrastructures is reflected in the increased number of related submissions. On the contrary, despite the general trend towards more environmentally conscious technologies and systems, the smallest number of submissions was received for green and sustainable technologies.

In terms of acceptance rate, Table 7 shows a great variability among topics, with the minimum (0%) for component-based software engineering, embedded/cyber-physical systems, and model-driven engineering, and the maximum (greater than 30%) for program analysis, testing and cooperative, distributed, and collaborative software engineering. **Out of 40 topics, 16 exhibit an acceptance rate higher than the overall acceptance rate of 18.5%.**

ICSE 2015 continued the used of categories, first introduced in 2014, where each category delineates a particular type of papers: analytical, empirical, methodological, perspectives and technological. Topics and categories are two complementary ways for authors to classify their submissions, where Topics are some of the areas covered by the conference and are helpful to match submissions to reviewers, and categories are meant to help position the intended contribution of each submission, as well as set the expectations on how this contribution should be evaluated by the reviewers. Table 8 reports the distribution of submissions among the combination of categories.[1] Most submissions have one (56.7%) or two categories (42.7%).

Table 9 shows the number of papers that belong to each category and the fraction of accepted papers.[2] The distribution across categories for ICSE 2015 is very similar to that of ICSE 2014. The most popular categories are Technological (49% of the ICSE 2015 submissions listed this category; the percentage was 43% in 2014) Empirical (40%; 37%) Methodological (25%; 24%); Analytical (23% for both years) and Perspectives (7%; 6%). **The number of Perspectives submissions remain low compared to the rest, which points to the need to devise other actions if we want to see more of those submissions.** In terms of distributions of accepted papers across categories, the categories Analytical and Empirical have higher acceptance than the overall acceptance rate.

2.4 Authors

Over 1250 authors submitted their work to the technical track of ICSE 2015. Figure 2 shows the number of submissions per author across all authors (figure on the left side) and just for the authors with 5 or more submissions (figure on the right side). Fourteen authors participated in five or more submissions, 184 authors participated in two or more submissions, and the vast majority of authors participated in just one submission. Note that in these figures a submission with multiple authors is counted under each author because we are trying to characterize the contribution per author. Therefore, these numbers should not be aggregated across authors because of the multi-authorship possibility.

[1] The total number of submissions considered in this table is 448, not 452, as 4 withdrawn papers did not present category information.

[2] The sum over the columns "Submissions" and "Accepted" is larger than the actual number of papers submitted and accepted, respectively, as several papers have more than one category.

Table 7: Distribution Across Topics

Topic	Submissions	Accepted	% Accepted
1: Agile software development	16	3	18.75%
2: Autonomic and (self-)adaptive systems	23	3	13.04%
3: Cloud and service-oriented computing	23	1	4.35%
4: Component-based software engineering	18	0	0.00%
5: Configuration management and deployment	14	2	14.29%
6: Cooperative, distributed, and collaborative software eng.	23	7	30.43%
7: Debugging, fault localization, and repair	48	12	25.00%
8: Dependability, safety, and reliability	21	5	23.81%
9: Embedded and cyber physical systems	5	0	0.00%
10: End-user software engineering	16	1	6.25%
11: Formal methods, verification, and synthesis	49	12	24.49%
12: Green and sustainable technologies	4	1	25.00%
13: Human factors / social aspects of software engineering	44	11	25.00%
14: Human-computer interaction	20	4	20.00%
15: Methodologies and measures for empirical software eng.	36	9	25.00%
16: Middleware, frameworks, and APIs	18	1	5.56%
17: Mining software engineering repositories	71	11	15.49%
18: Mobile applications	37	6	16.22%
19: Model-driven engineering	22	0	0.00%
20: Parallel, distributed, and concurrent systems	22	4	18.18%
21: Performance	15	2	13.33%
22: Program analysis	69	22	31.88%
23: Program comprehension and visualization	25	5	20.00%
24: Programming, specification, and modeling languages	22	4	18.18%
25: Recommendation systems	16	1	6.25%
26: Requirements engineering	19	1	5.26%
27: Reverse engineering	19	3	15.79%
28: Search-based software engineering	17	3	17.65%
29: Security, privacy and trust	26	6	23.08%
30: Software architecture	22	2	9.09%
31: Software economics, management, and metrics	20	2	10.00%
32: Software evolution and maintenance	85	16	18.82%
33: Software modeling and design	26	2	7.69%
34: Software processes and process improvement	21	3	14.29%
35: Software product lines	14	4	28.57%
36: Software reuse	19	1	5.26%
37: Software testing	65	20	30.77%
38: Tools and environments	58	13	22.41%
39: Traceability	11	2	18.18%
40: Ubiquitous/web/pervasive software systems	14	2	14.29%

Table 8: Distribution Across Categories

Categories	Submissions
Analytical	28
Empirical	84
Methodological	28
Perspectives	12
Technological	102
TOTAL submissions with 1 category	254
Analytical & Empirical	19
Analytical & Methodological	16
Analytical & Perspectives	3
Analytical & Technological	36
Empirical & Methodological	22
Empirical & Perspectives	7
Empirical & Technological	40
Methodological & Perspectives	8
Methodological & Technological	35
Perspectives & Technological	2
TOTAL submissions with 2 categories	188
Empirical & Methodological & Technological	3
Analytical & Empirical & Technological	3
TOTAL submissions with 3 categories	6

Table 9: Acceptance Across Categories

Categories	Submissions	Accepted	% Accepted
Analytical	105	22	21%
Empirical	178	40	22%
Methodological	112	14	13%
Perspectives	32	4	13%
Technological	221	43	19%

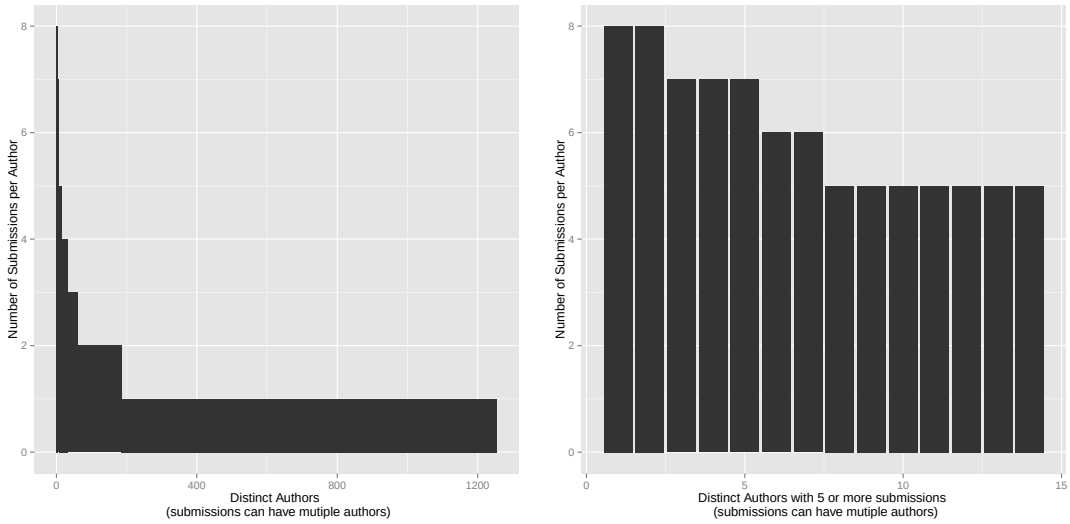

Figure 2: Submissions per Author

Figure 3 provides a complementary view, showing the accumulated number of distinct reviewed submissions (totaling 430) contributed by the authors. It is worth noting that **1% of authors participated in 19% of the submissions.** The large number of submissions for some authors may be a reflection of an undesired and costly shotgun submission pattern, but it also seems to be associated with authors that carry many active collaborations. We have no enough data to tease this out further. However, we stress that if we had bounded the number of submissions per author to a maximum of three submissions, we would have 34 less submissions (8% of reviewed submissions which consumed approximately 85 reviews). If we select those 34 submissions at random from the authors with more than three submissions, then approximately seven currently accepted papers would not have been submitted. Clearly, we expect that number to be smaller as authors selection is unlikely to be random, but rather include their strongest work for this conference.

As per the ICSE tradition, PC and RC members were allowed to submit papers and several of them (namely, 30 PC and 15 RC members) actually submitted their work. Table 10 shows the number of papers that listed at least one author from either the PC or the RC, and the fraction of accepted papers. **The acceptance rate for PC and RC submissions is higher than the overall one.** One possible explanation for the higher acceptance rate is that entering the ICSE reviewing committees recognizes a high level of research expertise, which, in turn, often entails a higher rate of success when submitting papers. Of course, another possible explanation is that reviewers of PC and RC submissions are somehow biased in their judgments. Unfortunately, not enough data is available to corroborate either conjecture.

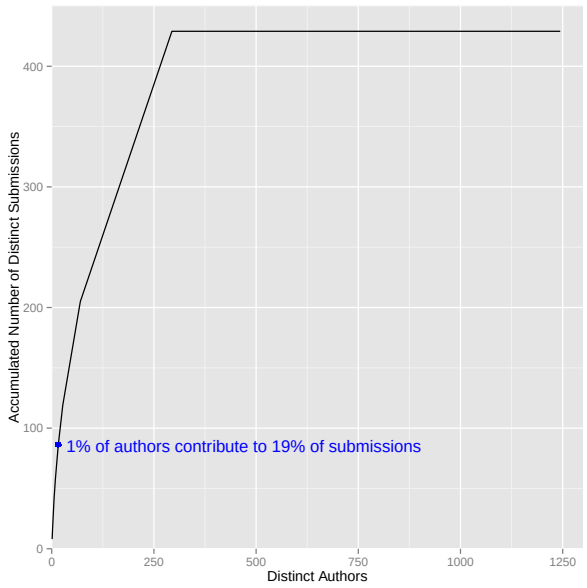

Figure 3: Accumulated Distinct Submissions

Table 10: Submissions from the PC/RC Committees

	Submissions	Accepted	% Accepted
PC	66	19	28.79%
RC	23	5	21.74%

Table 11: Surveys' questions

For PC/RC	
1	I mostly reviewed papers within my area(s) of expertise.
2	The reviewing load was manageable. (new)
3	The allocated time was sufficient to complete my reviews. (new)
4	Reviews by other reviewers were (mostly) thorough.
5	Reviews by other reviewers were (mostly) constructive.
6	Most online discussions were insightful and useful.
7	Decision summaries were (mostly) helpful. (new)
8	Final decisions of acceptance/rejection were balanced and justified.
9	Submission categories (empirical, analytical, technological, methodological, perspectives) were useful in setting expectations for evaluation. (new)
For Authors	
1	Was your paper accepted?
2	Reviews were thorough.
3	Reviews were constructive.
4	Reviews reflected sufficient level of knowledge on the part of the reviewers.
5	Reviews provided valuable feedback. (new)

3 Survey of PC/RC Members and Authors

As part of the technical track self-assessment we conducted a survey of PC and RC members and authors to characterize their perceptions about the reviewing process. This is in line with the healthy practice started by the ICSE 2014 chairs to get a better grasp of how the reviewing process is working.

The survey for PC and RC members had nine questions and a free text space for suggestions and comments, the one for authors had five questions and a free text space for suggestions and comments. The questions appear in Table 11 (we have marked as "new" the questions that were not present in the 2014 survey). The surveys' answers were measured in a 7-point likert scale: 1) Strongly disagree, 2) Disagree, 3) Somewhat disagree, 4) Neutral, 5) Somewhat agree, 6) Agree, 7) Strongly agree (matching what was done last year to enable their comparison). All authors and PC and RC members were invited to fill out the surveys. The survey did not collect any identification or personal information, and responses were anonymous. We requested for authors to coordinate with coauthors to fill one survey per submission in order to get one record per submission. We received 51 responses from PC and RC members (63% response rate) and 182 authors' responses (assuming one response per paper that renders a 42% response rate).

3.1 Reviewers Expertise

The reviewing process is meant to match each submission with a set of experts in the corresponding area that are able to assess the contribution and provide feedback. We asked PC/RC members whether the submissions they reviewed were within their area of expertise, and authors about the expertise reflected by the reviews they received.

The PC/RC responses are depicted in Figure 4. **From the PC/RC members' perspective, 94% strongly agreed, agreed, or somewhat-agreed that the submissions reviewed were within their areas of expertise**, which is consistent with the expertise reported through their reviews. This is an improvement over what was reported by the PC members in ICSE 2014 (88% agree to some extent), although the role of the PC was then somewhat different.

Authors' responses about reviewers' expertise are depicted in Figure 5. **From the authors' perspective, 67% agreed to varying degrees that the reviewers' had sufficient expertise** to evaluate their submission, 9% were neutral, and 24% disagreed to varying degrees. We note, however, very distinct response patterns among authors depending on whether the submission was accepted or not. The response patterns for accepted papers aligned with those by the PC/RC responses.

The corresponding responses from ICSE 2014 showed lower levels of agreement, with 58% of the authors agreeing, to some extent, that the reviewers were knowledgeable of the target area. We attribute this improvement in part to the availability of additional strong reviewers through the RC pool, the adjustments made to the PC and RC to match the submission topics, the careful and largely manual process we used to match reviewers to submissions, and the assignment of additional reviewers to submissions that required more expertise.

3.2 Quality of the reviews

Regarding the quality of the reviews, we asked PC/RC members whether the reviews performed by their peers were constructive and thorough. The responses are summarized in Figures 6 and 7. The response patterns for both questions are very similar, with **90% and 88% of the responses agreeing, to different extent, that the reviews conducted by their peers were constructive and thorough.** This shows an improvement over the responses from the PC from ICSE 2014, which reported 83% and 79% agreement respectively, although again the roles played by the PC/RC reviewers are somewhat different.

Authors' responses to the same questions revealed more variability, as shown in Figures 9 and 8. In terms of the reviews being **constructive, 64% of responses indicated agreement**, 17% were neutral, and 19% disagreed. In terms of being **thorough, 69% agreed**, 12% were neutral, and 19% disagreed. Again, we find very clear differences between authors of accepted and rejected submissions (agreeable responses for accepted papers were over 90%). Last, we asked authors about the value of the feedback they received

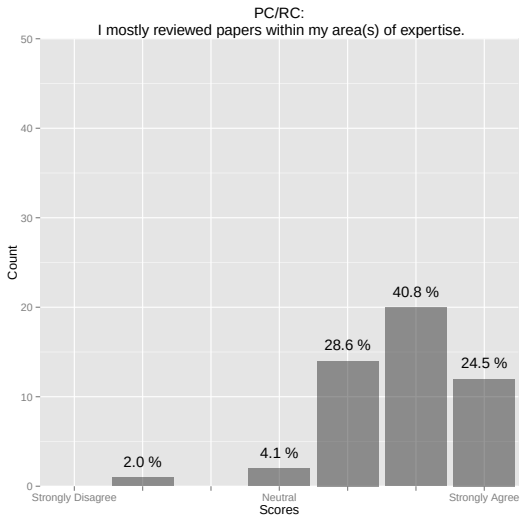

Figure 4: PC/RC Perspective of Reviewing Expertise

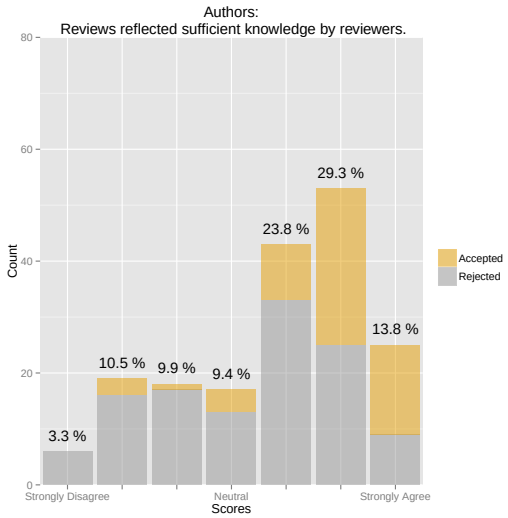

Figure 5: Authors Perspective of Reviewers' Expertise

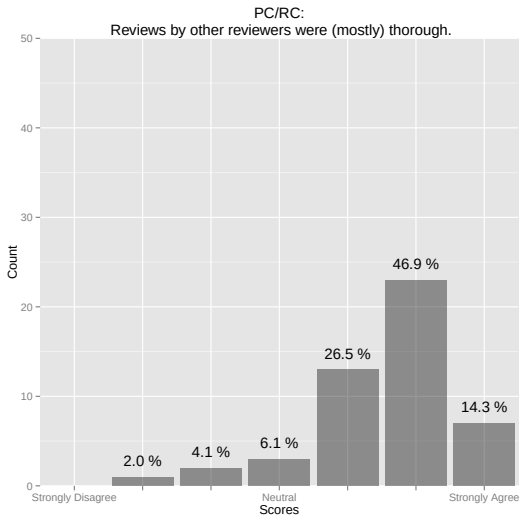

Figure 6: PC/RC Perspective on the Reviews being Thorough

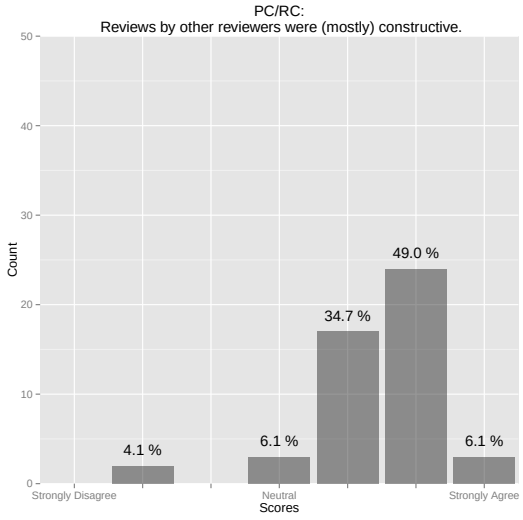

Figure 7: PC/RC Perspective on the Reviews being Constructive

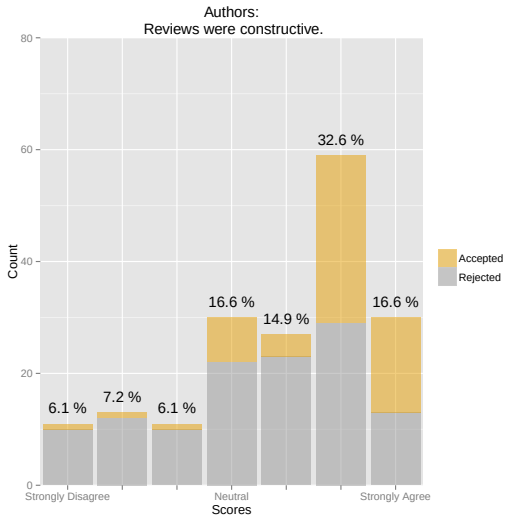

Figure 8: Authors Perspective on the Reviews being Constructive

in their reviews. As shown in Figure 10, 66% agreed to different degrees that the reviews provided valuable feedback, 13% were neutral, and the rest disagreed.

The range in the responses of authors of accepted and rejected papers about the quality of the reviews is reflected in the comments as well:

[Author of accepted paper] *These were, by far, the best batch of reviews I have ever received in my four years as a graduate student. Uniformly, they were extremely clear and detailed. Criticisms were constructive in nature, and each was backed with an objective line of reasoning rather than a strong opinion. Our camera-ready version is a much better paper than what we submitted, largely because of the helpfulness of these reviews.*

[Author of accepted paper] *We got 9 pages of reviews for a 10 pages paper, so yes, it was very thorough. It was also very useful, and it provided us with strong references and examples that enabled us to build a much better paper.*

[Author of rejected paper] *A couple of the reviewers for one paper unfortunately didn't "get" the result / contribution, which was a pity as I think if they had we would have got strong support for the paper. But overall despite the disappointment of not getting our papers accepted, the feedback was valuable in helping us improve the research and the presentation - just like peer review should be :-)*

[Author of rejected paper] *The feedback completely missed the mark with the reviewers placing more than necessary emphasis on experimental results and completely discarding the novelty of the problem and solution. If it were another conference (e.g., FSE, OOPSLA, etc.), the paper would have been accepted.*

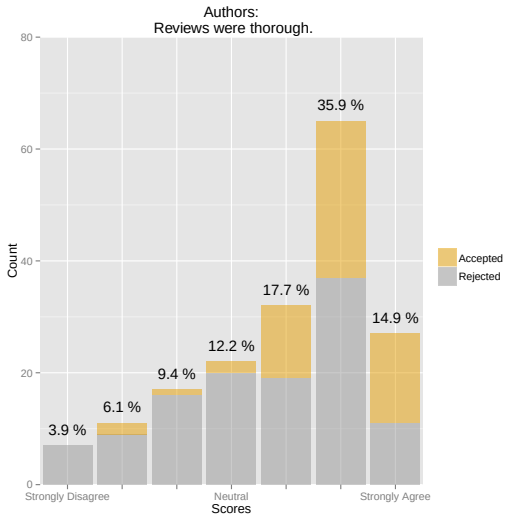

Figure 9: Authors Perspective on the Reviews being Thorough

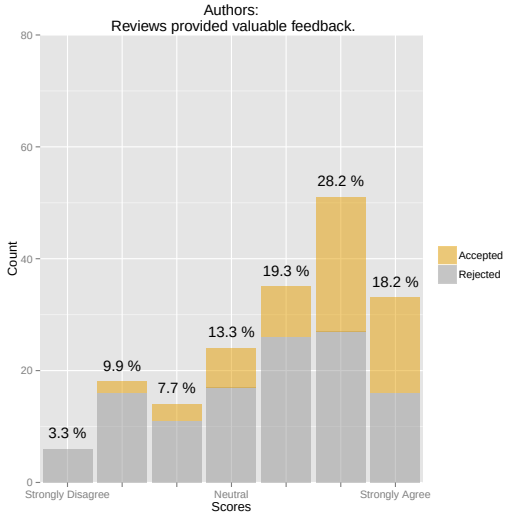

Figure 10: Authors Perspective on the Value of the Feedback Received

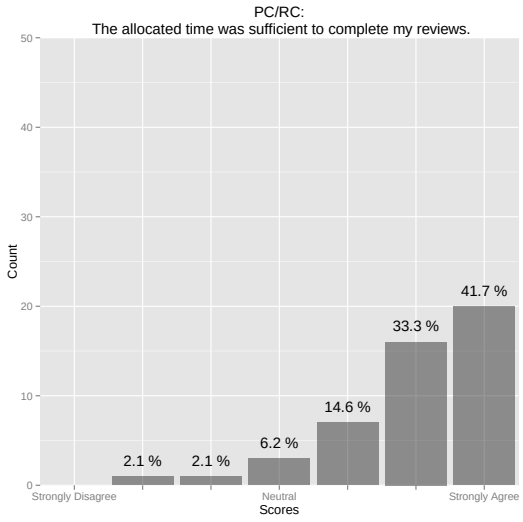

Figure 11: PC/RC perspective on the Allocated Time

Overall, the responses revealed strikingly similar patterns to the ones found in ICSE2014. 64% and 66% of authors agreed to different extents that the reviews they received were constructive and thorough, with drastic differences between authors of accepted and rejected submissions, and marked differences between the perception of the reviewers and the authors.

3.3 PC/RC Workload

We asked the PC/RC members whether they had enough time to complete their reviews and whether they thought the workload was manageable. The results are depicted in Figures 11 and 12. This really matters as load and scheduling conflicts can directly affect the quality of the reviews and cause for some strong members of our community to shy away from participating in future program committees.

90% of the PC/RC agreed, to different degrees, that the allocated time was sufficient to complete the reviews and that the load was manageable (42% and 31% strongly agreed, 33% and 31% agreed, and 15% and 29% somewhat agreed). This data was not collected for ICSE 2014 to enable a comparison.

3.4 Discussions, Summaries, and Categories

We also asked PC/RC members about three specific parts of the reviewing process that we felt we are still learning how to best use and manage: online discussions, summaries, and categories.

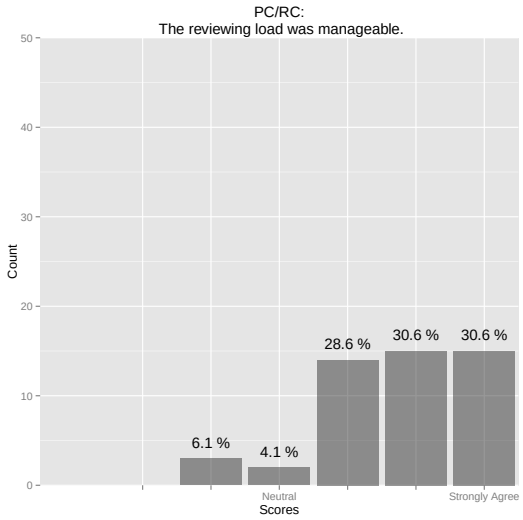

Figure 12: PC/RC perspective on the Allocated Load

Online Discussions. Recall that we instructed PC/RC members to use the online discussion to clarify reviewers positions, sensitize reviewers to strengths/weaknesses they may have not considered, and make sure RC members get a chance to clarify/emphasize their perspective in view of the discussion at PC meeting. As shown in Figure 13, 82% of the respondents agreed, to different degrees, that most online discussions were insightful and useful. For us, the program chairs, the online discussions were extremely helpful to get an early grasp in divergent perspectives and ask for clarifications and extra reviewers if needed before the PC meeting.

Summaries. Summaries prepared by the PC member leading the discussion and refined with the rest of the assigned reviewers was meant to streamline the discussion at the PC meeting, and help the authors gain insights into discussions that may not make it into the reviews but were part of the decision process. As shown in Figure 14, 77% of the PC/RC members agreed to different extent that the summaries were helpful.

For the program chairs, the summaries were very helpful to identify the key issues among many present in the reviews. At the same time, we recognize that there was much variability in how the summaries were written and more guidance and checks need to be put into place for them to be consistently useful as feedback for the authors. A comment from the PC/RC members reflects this well: *"Some of the summaries were well done. Others didn't get to the point of what was wrong with the paper or what could be improved."*

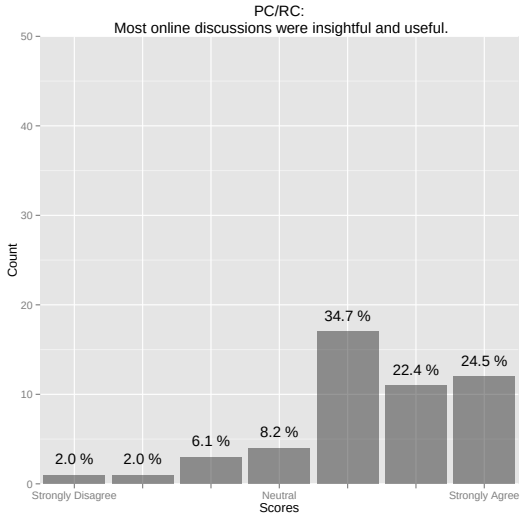

Figure 13: PC/RC perspective on the Online-Discussions

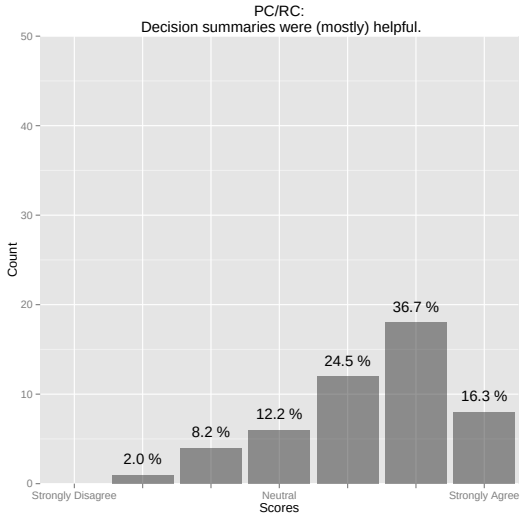

Figure 14: PC/RC perspective on the Decision Summaries

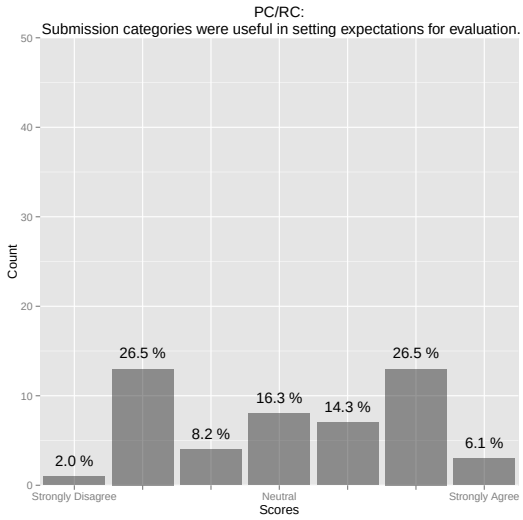

Figure 15: PC/RC perspective on the Usefulness of Categories

Categories. Categories were introduced as a mechanism to help position the intended contribution of each submission and set the expectations on how this contribution should be evaluated by the reviewers. We asked the PC/RC members whether they found categories useful in setting the expectations for evaluation. The responses, summarized in Figure 15, were quite mix, with 47% agreeing to different degrees about their usefulness, 16% being neutral, and 37% disagreeing. For us, incorporating categories into the process was a very minor effort, but it was unclear whether the use of categories significantly benefited the discussions.

3.5 Final Decisions

Making sound and fair decisions on the technical submissions for a conference like ICSE is a complex and difficult task that involves coalescing a large number of experts with diverse backgrounds, and sometimes diverging perspectives. We asked the **PC/RC members whether the final decisions we made regarding acceptance and rejection were balanced and justified. As shown in Figure 16, 88% of the respondents agree to different degree that they were.**

Throughout the process we found that submissions that were on the fringes of the core competences of the PC/RC members were among the hardest to judge. Even identifying those fringes was hard in some cases because they are continuous and shifting. We also found that diverging views often centered on whether the balance between technical novelty and evaluation was sufficient. Overall, we tried to be inclusive of submissions discussed at

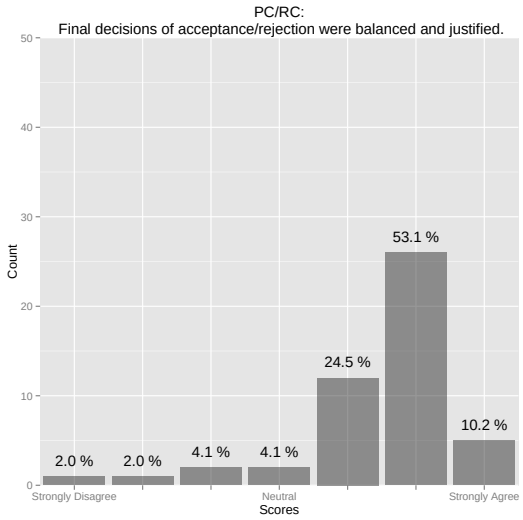

Figure 16: PC/RC Perspective on Balance and Justification of Decisions

the PC meeting that had strong support by at least a reviewer and no technical flaws or what we considered major objections were identified by the other reviewers. We conjecture that this strategy may have contributed to the percentage of respondents that did not feel the decisions were always balanced and justified.

3.6 Responses to comments

In this section we proceed to respond to some common threads that emerged from the comments gathered in the survey (others were already answered through the conference FAQ or in Section 4).

Comment thread: Submissions with two reviews may be rejected prematurely.
Response: It is possible. To reduce that risk, we checked all submitted reviews and pay particular attention those that were going to be rejected. If we believed that we did not have enough expertise or the reviewers had doubts, or the reviews were not aligned, we requested a third reviewer and let the submission move to the second round.

Comment thread: It is becoming really hard to accept a visionary paper.
Response: There is a tendency to give increasing importance to the evaluation of the contribution and this may make it harder for papers lacking that element not to be accepted. Categories were introduced with the aim of helping with that problem, by tuning reviewers' expectations to the nature of a paper's contribution as intended by the authors. Unfortu-

nately, we have no data to understand if categories were effective in their aim, although it seemed to benefit at least one submission in the PC meeting discussion.

Comment thread: Some reviews were wrong / subjective / biased / non-specific / ...
Response: We apologize for that. We recognize that the reviewing process is noisy and as discussed in the report, we have strived to reduce that noise and to generate valuable reviews. More specifically, we did check all the reviews and more than 200 cases we asked reviewers to improve their reports, and in several cases we went through multiple iterations of improvement. However, not all reviewers were equally responsive and in some cases we were not able to insist enough on the required changes as reviews arrived at the very last minute. In other cases, we and the co-reviewers may not have detected or be knowledgeable enough to detect a problematic review. As chairs we also strived to assign as many reviewers as needed (for 35 submissions we had 4 reviewers) to make a more informed decision, and we think that is one of the keys to improve the overall reviewing process. Another improvements may include setting intermediate deadlines for reviews to be submitted sooner, hence adding more slack time for review checking, and providing more incentives for reviewers (e.g., Distinguished Reviewer Award) could be useful as well.

Comment thread: Quality of summaries varied.
Response: Summaries are a relatively new tool to the ICSE reviewing process and, as such, we need to learn how to best write them and use them. In most cases we found that summaries were able to capture the strengths and weaknesses of a submission, and the points of disagreement among reviewers. However, many of them were not revised carefully based on the PC meeting discussion to explain the reasons for the final decision. More precise instructions and the maturation of this practice over time will most likely improve the quality and usefulness of summaries to committee members and authors.

Comment thread: On having a physical PC meeting.
Response: Several comments in the survey are concerned with the value of having a physical PC meeting. These comments reflect a full spectrum of opinions, from those seeing little value in it to those that think it is critical. Informal comments we have received outside of the survey seem to be overwhelmingly in favor of hosting a physical PC meeting for ICSE. Curiously, many of these comments came from the newer and most junior PC members that really seemed to value and learned from the experience. From the chairs perspective, we do see the extra value of the PC meeting both to make more conscious decisions and to forge the sense of belonging to the ICSE community.

Comment thread: on Running the PC Meeting.
Response: As mentioned in the comments, we prepared extensively for the meeting, reading every review, with at least one of the chairs being familiar with every submissions under

discussion, and with questions ready to get more quickly to the point on difficult to decide submissions. In the end, as captured by the survey, the PC/RC believe the decisions were mostly fair and balanced. Still, we made a couple of mistakes in the meeting organization in terms of the space (small to easily hear everyone but harder to get in and out of the room to handle conflicts) and the seats distribution (a hybrid between square and auditorium, which fitted in the smaller space, but did not let everyone see everyones face). We also did not manage the time as intended, and finished an hour later on the second day. Overall, and in spite of having 2-3 reviewers that were overly negative and in one case inappropriate, the discussion of submissions at the PC meeting was insightful and engaging, with most reviewers knowing their assigned submissions very well. Indeed, most discussion leaders did a great job of capturing the RC members perspective and the underlying issues worth discussing at the PC meeting. Only in a very few cases we had to intervene at the PC meeting discussion to state the RC viewpoints.

4 Reflections from the Chairs

In this section we take the opportunity to reflect on several important aspects of the conference organization that need to be carefully considered in the future.

Balanced process. We believe that, given the number of submissions and its rate of growth, the reviewing model we used this year struck a good balance between maintaining a manageable load for reviewers (the average load per PC member was 17 papers, maximum 18; for RC members the maximum load was 9 papers; 90% of PC/RC members agreed that the load was manageable) and having the reviewers, who know the submission most intimately, at the PC meeting to discuss and make the final decisions (about 2/3 of the papers discussed at the PC meeting had 3 reviewers present, and 88% of the PC/RC agreed that the final decisions were balanced and justified).
Note that simply adding more RC members to contribute reviews in the first phase would make the process scale further, shifting the PC members to review more papers in the second or third phase (4th reviewer assignment), and still maintaining the size of the PC viable for a physical meeting. We strongly feel that having reviewers at the meeting resulted in meaningful, thorough and sometimes passionate discussions and, on the long term, it is a means to shape the future of the community and to help the growth of new generations of competent and committed reviewers.

Increased reviewing expertise. Including a Reviewing Committee (RC), in addition to alleviating the load on PC members and helping us reach a broader section of the community, was also useful to inject into the process fresh energy and additional expertise. This, together with the careful and largely manual process we used to match reviewers to submissions, and the assignment of additional reviewers to submissions that required more expertise, resulted in an increase of expertise on reviews over ICSE 2014, as captured by the reviewers marked expertise (e.g., 7% more of submissions with an "X", 11% less with a "Z") and the reviewers and authors survey (e.g., 9% more of authors agreed that the reviewers had sufficient expertise).
We strongly encourage chairs to continue applying the principle of assigning as many reviewers as needed and available (2, 3, or 4 in our case) to make the best possible decision on each submission. This helps to best use the community reviewing resources. Furthermore, using the same principle, it is cost-effective to spend more time on submissions that require it (an example of this is to spend more time at the PC meeting on undecided submissions instead of those that would be surely accepted based on the reviews and online discussion).

No reason to panic about scalability of process. Given the somewhat increasing numbers of submissions in the last few years (441 in 2011, 408 in 2012, 461 in 2013, 495 in 2014), the scalability of the reviewing process has been enough of a concern to serve as one of the drivers for experimenting with different reviewing models. However, when considering the 452 submissions for this ICSE, the growth may be linear but with a very small coefficient. This clearly requires close monitoring in the future but no dire measures.

Do due-diligence and take risks on new PC/RC members. 23% of the PC/RC members were new. Not all of them performed to our expectations, but the great majority did, and several of them were among the best reviewers we had. Furthermore, they all improved as the process went on and they fully realized the impact of their performance. We also noticed that our most Junior PC and RC members got better throughout the reviewing and on-line discussion (and, for PC members, discussion at the physical meeting). This is, in our opinion, a key way to educate future leaders in the ICSE community and, as such, it has a great value for the community.

Revise bidding process. The bidding process is expensive for many reviewers that take the time to carefully choose their preferred submissions, but it is not a reliable source of information to support the assigning process of submissions to reviewers. Chairs may want to explore alternative mechanisms to support the bidding, perhaps through some form of recommendation system based on the reviewers' publication records, and by asking authors to indicate preferred/non-preferred reviewers among the ones without a conflict of interest.

Understanding and improving misalignments. One of the main misalignments we are concerned about is between the expectations of the authors of rejected submissions and the reviews and outcome they receive. Approximately 20% of the authors surveyed did not agree that the reviews showed expertise, were constructive, were thorough, and provided valuable feedback. But only 3% of those responses came from authors of accepted submissions. We do have enough points of reference to compare against these numbers. We also recognize that there will be authors that will be negative in the face of a rejection whether justified or not. But for other cases, having a channel of communication between authors and reviewers may help to set expectations and to mitigate misunderstandings that lead to overly-negative perceptions. Perhaps variants of the rebuttal stage and the mentoring program used in past ICSEs might help.

Consider double-blind reviews. This is being tried and adopted by other strong venues (e.g., PLDI, CHI, POPL, OSDI). We find reasons to support it such as mitigating the potential for bias or perception of bias (this latest one is confirmed by some of the authors' comments in the survey). We also realize that it may lead to some other types of decision errors (e.g., rejecting based on previously unpublished and not cited work by the same authors) and that it will require overcoming a learning curve (for reviewers, authors, and conference management systems). Several sources now exists that offer more compelling arguments for it, some from authors of ICSE papers [3], that are worth exploring in more detail. Now, we did not collect any specific data to support any position. However, we note that the acceptance rate for submissions coming from PC and RC members are higher than the general one, which may support at least the perception of bias. A double blind review could help understand if such influence actually exists.

[3] http://www.cs.utexas.edu/users/mckinley/notes/blind.html, http://www.cs.umd.edu/~mwh/dbr-faq.html

Early rejection notification. We only received positive feedback about this change, which enabled authors of 189 submissions to received their reviews almost 5 weeks earlier, and in some cases rework them for other ICSE tracks. In the future we recommend for chairs to coordinate with the other tracks so that this early notification enables the rework and resubmission to other fitting tracks as much as possible (particularly the industry track).

Encourage artifact/toolset sharing in CFP. Throughout the reviewing process we noticed that many reviewers wished and often requested for authors to make their empirical artifacts and tools available. In some cases reviewers even argued that having access to certain artifact or tool may have affected their decision. Although inclusion of artifacts and toolsets may introduce some logistic challenges and add another requirements just for some types of submissions, it may be worth it to at least encourage authors in the CFP to share their tools and artifacts.

Once something is published/announced do not change it. We accepted a change (what it seemed like a minor improvement in fact) in the publication format template before the submission deadline. We published the change in the conference and submission websites, and announced it through various channels more than a month before the deadline. Still, this late change generated some inconsistencies in some of the submissions format and added unnecessary confusion among authors. Since nothing can guarantee that the authors will re-check the instructions once they are posted, changes to posted information should be minimized.

Consider alternatives for conference management system. A reviewing process of ICSE scale relies extensively on a conference management system. In our case, we employed CyberChair. This system has been used by ICSE for many years, which on one hand means that it has evolved with ICSE to support its processes, and on the other hand it signifies that its underlying technology and interface is showing its age. The system seems also to rely heavily on its developer/administrator, which on one hand means that new features and anomalies can be addressed quickly, while on the other hand some tasks remain manual and often requests are delivered at the last minute. In the end, we were satisfied with CyberChair but would encourage to look at other viable alternatives.

Build slack in the review process. Having additional days to monitor and enforce changes in the reviews and summaries would have made our effort more effective. This is important not just during the reviewing phases, but also before the notifications are sent. One way to introduce slack without extending the reviewing process would be by implementing intermediate reviewing deadlines so that reviewers submit their work incrementally instead of in a single batch at the due date.

Bounding the number of submissions per author. It may be worth exploring whether defining a maximum number of submissions per author would help to curb the abusive shotgun approach to submissions and encourage authors with multiple collaborations to submit just their best work. As exemplified in our analysis, enforcing a limit of three submissions per author would reduce the number of submissions by approximately 8% and we conjecture that the program will not suffer as authors would self-select their best work for ICSE.

Keep assessing and communicating how the technical track operates and performs. We strongly encourage future program chairs to share their experiences through reports like this and through other mediums like the FAQ site. This healthy practice started by our predecessors provides often missing organizational memory and some assessment of the directions we explore. It also helps to make the whole process more transparent for the community. Using similar questions, processes, and measures, will also enable us to perform some more meaningful comparisons across the years. Still, plenty of opportunities remain to explore further questions about reviewing quality, reviewers strengths, and the cost-effectiveness of different parts of the process. More specifically, having some links between the surveys and submissions could help tease out many of the remaining questions.

A Process Timeline

Table 12: Timeline

Year	Month - Day (approx.)	Activity
2013	April - Aug	Define reviewing process
	June - Sept	Prepare candidate list of PC + RC
	July 15 -	Contact and setup cyberchair
	Aug - Sept	Define reviewing timeline
	Oct 20 - Nov 20	Invite PC + RC
	Dec	Update web site with committees
2014	Jan 1 - 30	Simulate process load based on previous ICSE data
	Feb 1 - Feb 15	Revise topic list
	Feb 1 - Feb 15	Invite other RC members to complete expertise
	Feb 1 -	Draft CFP
	March 1 -	Solicit feedback on CFP
	March 15 -	Post CFP
	April 30 -	Request PC/RC input expertise
	June 5	PC+RC meeting at ICSE 2014
	July 5 -	Test / Enable paper submission site
	5-Sep	Submission Deadline
	Sept 6 - 7	Format and scope check
	Sep 8 - Sept 11	PC/RC Bidding
	Sept 16 - Oct 25	PC/RC First round of reviews
	Oct 30 - Nov 16	PC Second round of reviews
	15-Nov	Early notifications to authors
	Nov 18 - 30	PC/RC Online discussion
	Dec 2 - 8	PC Selective additional reviews
	Dec 11 -	Pre-conference meeting (maybe)
	Dec 12 - 13	PC meeting
	Dec 20 -	Notifications to authors
2015	Jan 15 - Feb 10	Survey authors, PC, RC
	Jan 20 - 30	Distinguished paper selection
	Feb 15 -	Camera ready copy
	Jan - Feb	Program planning
	1-Apr	Request session chairs
	1-May	Send presentation information to authors
	May 16 - 24	ICSE 2015 @ Florence

B Survey Comments from PC and RC Members (verbatim)

I think that the get-together PC meeting is redundant, not only for ICSE but also for other conferences. Double-blind submission process should be enforced.

I was initially worried about having only two reviewers out of three at the PC meeting, but the PC chairs managed this issue so well that I never felt the need of physically having the third reviewer at the PC meeting to take a better decision.

One thing I noticed was that in the ICSE'15 review process and PC meeting – much like in other conferences and PCs before – there were again one or two reviewers that were just plain negative. This is not very helpful (basically, you could reject every paper) and clearly against the exceptions of the PC chairs to look for reasons of acceptance. Fortunately, the PC was able to handle this issue.

Some negatives: Room was overcrowded. Some papers had only 2 reviewers in the room. Some debatable papers were accepted, some even without discussion. Time management could have been better.

Two major comments: 1. It is becoming really hard to accept a visionary paper since reviewers are typically expecting a table that shows data. In some sense this is against the nature of a conference that should be also the place where we can discuss new ideas. On the other side it is quite easy to accept a paper presenting an incremental work, even no so exciting, if the work is extensively validated, even reusing the validation setting of previously published papers. I feel that this should change. 2. The implementation of the two phases of reviews is by construction unfair. In fact the order in which grades are received matter. I explain why. Imagine to receive two C at the first phase, then most likely the paper will be rejected (without chance for a third reviewer that might give B or even A). Having a B and a C at the first phase might lead the paper to the second phase. Then if the paper gets another C, then the paper has still some possibility to be discussed and also accepted. We cannot rely too much on a C or B since often during the discussion reviewers say, "yes, I have given C but means B", and vice versa.

The advanced preparation was very effective. I question the value of a physical meeting of the full PC.

I think you did a very good job at checking the reviews while they were coming and at timely require updates and clarifications where needed.

Some unacceptably vacuous and subjective reviews were accepted and in some cases had an impact on the entire paper evaluation process including the final decision.

One problem I see with the reviewing process in software engineering conferences with a wide scope (like ICSE and FSE) is that the expectations of reviewers in different areas of software engineering research are very different. This means that papers in a certain area maybe at a disadvantage just because of being in that area. I do not think the submission categories helped with this issue. Here are some radical approaches to address this problem: 1) add a 4th out-of-area reviewer for papers that pass the first phase, or 2) divide the whole conference to tracks that have their own sub PCs (like WWW). I am not sure either of these would work. It is possible that this is an unsolvable issue that we have to live with.

I am not sure the categories helped, in most cases it was not addressed in the reviews or discussions

Some reviewers were not interested in participating meaningfully in the online discussion, preferring to hash things out face-to-face. That gives an undue advantage to native English speakers, especially those who are skilled in debating. They can browbeat others into accepting their opinions, even when most of what they're saying is factually incorrect.

Some reviewers appeared to reject an entire paper when triggered by a single tiny thing that they felt was wrong. I can acknowledge that sometimes a single flaw can ruin an entire paper, but it's important to proceed to review each section of the paper as if the rest of the paper were correct. Otherwise, if the reviewer is shown that the flaw they identified is not really there, they have contributed nothing of value to the review that could help the authors understand the evaluation of all of the rest of the parts of their paper.

The review process from beginning to end was well managed by the chairs. The time in the PC meeting could have been managed a little better but when you are herding cats, you did the best you could.

The second reviewing period was too short. It was not unreasonable to review the planned number of papers in that period, but if other things come up there is little time to allocate. Better to reallocate some time from the online discussion period, or gap time before the PC meeting, to this.

ICSE should seriously consider light double-blind review, author response periods, and having the ERC review all PC papers to avoid an awkward discussion of PC papers in the PC meeting. Other CS conferences have adopted these practices with a great deal of success. My observation is that every conference who tries them, never goes back. They are just a better way to do things (and BTW, I think that's true of ICSE's discussion summaries and online discussions too, so great job there!)

The online discussion was helpful but not sufficient by itself. The in-person meeting was critical to the ultimate quality of the decisions.

Response represents my experience as a reviewer; my experience as an author, however, may differ :)

The chairs did a great job in communicating. This was excellent. Also - it was obvious that they spent a lot of time going through individual reviews and providing feedback. This was all very well done.

The one suggestion I would make is not to ignore reviewer bids. While I felt competent on all the papers I reviewed, I saw a handful of *other* papers that would have been exactly in my area of expertise. I had bid on them. I also saw that some of those papers lacked a true expert reviewer. So in the end - I don't think it was a good idea to ignore reviewers' bids. It is the difference between expertise and competence. I expect that in some areas there are so many experts that it isn't an issue - but in other more outlaying areas - this becomes an issue.

The process was quite smooth and worked very well.

For the most part reviewers did a great job of commenting on the papers. A large amount of very good discussion went on. Also, while I did get papers in my area, some reviewers were obviously not in the area and had little understanding of the literature.

Some of the summaries were well done. Others didn't get to the point of what was wrong with the paper or what could be improved.

Still a bit too much "feeling" or "opinion" going on rather than scientific fact in the reviews.

A bit too much of "I think XX did this previously" when the reviewers actually don't know what XX did and it is only slightly related.

C Survey Comments from Authors (verbatim)

The feedback completely missed the mark with the reviewers placing more than necessary emphasis on experimental results and completely discarding the novelty of the problem and solution. If it were another conference (e.g., FSE, OOPSLA, etc), the paper would have been accepted.

Great work!

One review in particular provided multiple ideas about how extending the paper.

It would be great if the conference has rebuttal phase.

I have to admit I was a bit disappointment after reading one of the reviews. There had 2 reviews: a constructive one and another one, which I considered to be a bit unprofessional (even mean I would say). Maybe our work was not good at all (maybe quite bad) but we still invested a lot of time in it ... Anyways, at least the second reviewer had really nice indications. We will try to be better prepared next time :P Thanks a lot, all the best!

Seems that as always in the history of ICSE, reviewers are conservative and prefer variations of old ideas over radically new stuff. If a single reviewer is slightly out of his comfort zone that will be the end of the paper. Unfortunately, this tendency is now even carrying over to the new ideas track. NIER is becoming conservative as well.

One of the two reviewers asked us to cite one particular single-author paper, and wrote that the outcome of evaluation would have been different had we included this particular citation.

It appears to us that one of the reviewers was biased against the paper because it exposed significant errors in their previous work. As a result we question the integrity of the reviewing process and the accept/reject decision.

ICSE NIER paper here.

The PC played his role very well, even going beyond what is expected of him, and provided a good review. This was excellent!

The quality of the other ('normal') reviews was disastrous and showed that the reviewers did not take time to read the (entire 4 page) paper. IMHO this is not acceptable for ICSE.

While I am, of course, happy that our paper was accepted, some of the reviewer comments indicated that some of the reviewers did not carefully read the entire paper. This is not a serious problem for me, given that the paper was accepted, but it is troubling for rejected papers, where the reviewer may well have objected to something and that objection was addressed in the unread portion of the paper. Sadly, I have no idea how you can systematically prevent such problems, other than reminding reviewers to be thorough and fair.

I found the numerical answers difficult as the quality of the reviews differed quite strongly. The length of the reviews was appropriate overall.

Some of the reviewers did not understand some of the empirical methods we employed, especially the qualitative analysis.

The reviews with insufficient knowledge won and led to the paper being rejected. I think this is unfortunate for a scientific *conference* where I rather would like to see more and more diverse ideas.

Where there were some interesting points made by reviewers, the reviews gave this impression that the

reviewers did not try much to understand the paper and its novelty. There were no concrete remarks to improve the paper, and the reviews remained at a higher level, e.g. the contribution of the paper is not clear.

We got 9 pages of reviews for a 10 pages paper, so yes, it was very thorough. It was also very useful, and it provided us with strong references and examples that enabled us to build a much better paper.

Reviews were important to conduct research.

I normally am appreciative of the reviews, even when my paper is rejected. These reviews missed the mark, though, imposing a personal view of relevance of own (reviewer) research that, sad but true, is just not.

Reviewers were obviously from academia and were looking for a neat little reductionist hypothesis and 'experiment'. They were not intellectually/experientially equipped to review an industry CASE STUDY paper which had a lot of insights from the field where controlled experimentation is not possible. They were not able to understand this difference, so their comments were not helpful.

The only issue pointed out was a misunderstanding of our evaluation which reflected a misreading of our work - a rebuttal would have quickly addressed that. You guys should consider it and not allow the reviewers to get away with saying things that are unchallenged,. It is to encourage a rebuttal that I filled this report I was quite satisfied – no further comments.

The reviewers identified problems, generally, but did not specifically give examples where they encountered these problems. This makes it difficult to improve the paper.

One reviewer asked for a usability study, which is surprising. I believe some SE researchers use this shallow requirement as a way justify their inability to comprehend the technical details of the approach. Usability studies appear to be experiments, but they are highly confounded. We need to educate our community that usability studies evaluate the quality of training materials and not the theory underlying the tool.

Our paper received two reviews. The first reviewer was incredibly biased and ignorant. They simply did not take the time to read and understand the work and rejected the paper because they thought it was the same as one of our previous ICSE papers, which is absolutely not true. They are in a way accusing us of plagiarism (selling the same idea twice), which is a pretty serious accusation. I wish the chairs had taken this review report a bit more seriously and followed up on it. The reviewer says "I am arguing that the problem that the authors address in this paper can be solved and it is already addressed in their last year's ICSE publication". Of course there was no way for us to argue that they are mistaken since there is no rebuttal phase in ICSE. Also the comments from this reviewer were quite patronizing, which is absolutely not necessary when writing professional reviews at this level. The second reviewer was much more constructive and thoughtful, for which we are thankful.

Overall, as an author, I believe it is better to receive 3 reviews than 2 because of such problems that can emerge: a reviewer does not understand the work and the paper is killed already. On the positive side, sending out early notifications is very helpful since it allows authors to improve the paper and target the next conference.

Just for the record, we submitted three papers and one got accepted. This feedback is for one of the papers that was rejected. It is not the rejection that bothers me with this particular paper, it is the tone, unfunded accusations, and the erroneous nature of the review report. I'll be on the ICSE PC next year myself, and hope to do a better job than this particular reviewer. Thanks for all the hard work!

Maybe the reviews can be formatted to look neater.

If the reviewers can give scores, that could be better and clear.

Most reviewer's comments were quite minor, though the paper was still not accepted. I had one reviewer who voiced an issue regarding the hypothesis tests in the paper. Their comment and suggestion for "correction" was fundamentally incorrect, given the explanation of one-sided hypothesis testing in any first-year undergraduate statistics textbook. I was very disappointed to see the quality of this review.

Considering the impact factor of this conference, the review process should really be double blind.

I think that the first round of reviews was more focused on rejecting a large number of paper than on selecting high-quality work, which could eventually be improved before the camera ready. But at the same time our paper was rejected, so I cannot say that I'm not biased. Nevertheless, the reviews were high-quality and provided valuable feedback. Best regards, C**** T****

The reviews were detailed and helpful.

A couple of the reviewers for one paper unfortunately didn't "get" the result / contribution, which was a pity as I think if they had we would have got strong support for the paper. But overall despite the disappointment of not getting our papers accepted, the feedback was valuable in helping us improve the research and the presentation - just like peer review should be :-)

Most reviewers made no comments on the meat of the paper and instead focused on how the work related to prior literature (especially which literature) and the implications for future work. So, it's not enough to do good work itself to get a paper into ICSE. You have to situate it in the reviewer's conceptions of the past, and their hopes and dreams for the future, neither of which you could know when you submit the paper.

My paper was 'early' rejected. I don't mind the rejection in general be it early or according to the regular schedule.

I have concerns regarding the whole concept of early rejection. First of all, I think it is subject to bias (see conflict of interest). I explain: If a PC member submits a paper to ICSE, it is at his or her advantage to early reject as many papers as possible. It is not clear from the ICSE website whether or not the same people performed the two round of reviews. If so, this is a problem. How can we be assured that the early rejection does not favor PC members who are also authors? The early rejection comments are very shorts (I base this on the comments I receive usually from ICSE - normal review) compared to normal reviews. They appear to be performed in a rush (I guess this has to do with the number of papers received).Other issues: A large percentage of the PC is from the U.S. Why is that? Some PC members have served on the PC for multiple years (sometimes in a row). Why is that? Please feel free to share these comments with the steering committee. My aim is to improve the conference and not to simply criticize its review process. There is a growing impression that ICSE is biased. I don't share this impression. But it is there, so it has to be addressed. Thanks,

Reviews were weird. For example, the first review said that the tool was very hip, really cool and nifty design and that we are better off going for a startup - and that it is not suitable for ICSE. The only other complaint from that reviewer was that minority/diversity were not considered - this is something we agree with, but cant really be a ground for rejection. The other 2 reviewers largely commented about the number of "test subjects" that the users evaluated in the study. Given that each study already required more than an hour long, having more test subjects in the study would be infeasible.

Two of the reviewers provided legitimate and hepful reasons for rejecting the paper, so I have no problem with the result. On the other hand, one reviewer listed as one of his two major reasons for rejection that we had not cited our previous tech report on the same topic, with essentially the same content. To my understanding of the ACM republication policy there is no need to cite tech reports, and tech reports do

not count as prior publications. The reviewer should have been instructed to remove this comment from his review.

It should be better that if the reviewers can give their point of view about the problem.

Strongly Disagree: Reviews reflected sufficient level of knowledge on the part of the reviewers This is based on substantial evidence across two of the paper's three reviews. One reviewer made numerous claims regarding the application of machine learning in software engineering research in repudiation of our approach, yet each and every one of the claims was undeniably false. The fact that his/her claims are false is clearly and comprehensively supported by the literature. The disconcerting part of the review is not the simple missteps of the reviewer (demonstrated by his/her claims), it's that the reviewer appeared to pass himself/herself as an expert in the area. This would have been tempered had there been an opportunity to provide a rigorous and professional rebuttal. In fact, we compiled a rebuttal in the course of analyzing our three ICSE reviews and considering improvements to the paper. In another review, it was evident the reviewer was not proficient in the area by the nature of his/her questions as well as the lack of constructive and valuable feedback. While our goal is to strive for a clear and impactful paper with compelling results, we also believe it is paramount for reviewers to have demonstrated some level of proficiency in the relevant area(s) to serve as reviewers. Notably, a third review was largely constructive and provided good feedback.

One review was somewhat biased from the reviewer's perspective, in favor of a tool they created or helped create. With only two reviews, a balanced reviewing process is not guaranteed anymore. We think going to a review process with three reviews per paper would ensure fairness in such situations.

Reviewers failed to understand the problem correctly. I am dissapointed that ICSE reviewers lack in general knowledge about distributed systems. I thought that software engineering, middleware, and distributed programming abstractions were competence of this conference. But the reviewers did not understand basic concepts like Actors, single-threaded models or Chord overlays. Only one reviewer could provide some related work that was in fact not a competitor or even replacement to our proposal. They just provided subjective opinions without references or concrete critiques.

It is my first submission and I will not try again in the future. I am really dissaponted by this conference and I am publishing in top ones.

I think it was a difficult paper for reviewing, and I think the reviewers did a good job with it.

The reviews showed a very good understanding of the paper and provided constructive and helpful feedback for further improvements. The reason for rejection was reasonable.

Reviews were not constructive or useful at all. In fact, none of the review helped to find any improvement point in the paper. Large parts of the paper, according to the received reviews, were not understood and it deeply influenced the evaluation of the paper. Overall, the whole experience has been really disappointing!

Thank you.

This is the first time we've submitted to ICSE – even though our paper got rejected in the second round, I was happy with the reviews and it will help us to improve the paper further.

Overall assessment is significantly different from the individual assessment of the reviewers.

The general quality was mixed.

The Technical Research Paper submission: "Synthesizing Code from Free-Form Queries"

The Tool Demonstration paper: "Interactive Synthesis using Free-Form Queries"

Many of the problems addressed by the reviewers could be handled before sending a camera-ready review. So, the paper might be accepted. But I understand that this is ICSE and agree that improvements can be made.

The reviews help us with inputs to improve our work. We plan to apply them in our next submission.

It seems that big names are easier to go through the review process. Some papers with big names are just so-so. Why not conduct double-blinded review process?

Reviewer 1 was fine. Reviewer 2 was fine in principle, but had not taken enough time to understand the structure of the writeup. Reviewer 3 was totally the wrong person (strongly anti-empirical attitude for an empirical work).

I strongly disagree with the notion of the reviewers on what is a 'novelty' or what is a significant contribution. Our paper is focused on confirming findings on code smells in further different contexts, thus should not be discarded as a non significant contribution. This is quite symptomatic of the current mindset of Software Engineering Research, where replication studies are not valued or are deemed not interesting. Moreover, I find little or none suggestions about potential improvements on the paper, which is quite disappointing given the high expectations when submitting to ICSE.

We are satisfied with one review. It was constructive and reflected sufficient level of knowledge. However, the other review did not make any of the criteria above. The reviewer didn't seem to carefully read nor understand the contents of the paper—he or she criticized about something that we didn't even mention.

The reviewers recognized many strengths of the paper and suggested some constructive improvements. However, the reasons why the paper was rejected were weak and not fully motivated, especially considering the kind of paper submitted and the available pages for the submitted paper.

Best reviews I've had from a conference this year in terms of thoroughness and constructiveness. This is not to say most positive; in fact, the reviewers brought several important criticisms of the work, which we appreciated, as well as a couple of misunderstandings; however, this is o.k., I'd rather have the feedback!

I believe the reviewers were more accustomed to analysis of traditional languages (e.g. Java). So they had missed most of the JavaScript-specific features and challenges that distinguished our work. The second reviewer was more thorough and had tried to provide useful feedback. The first reviewer, however, had missed the main contributions of our work. Moreover, some of the weaknesses that the first reviewer had mentioned about the paper, were addresses in the paper, or were just not valid. These misunderstandings happened even in the simple statistical arguments. Here are a couple of examples: Reviewer 2: "The paper mentions that the differences between participants in the experiment and control groups for T3 are not significant. However, from Figure 4, for T3-Exp, the accuracy is 100% with little variation, while for T3-Ctrl, the accuracy is 0% with little variation. Why would the difference be insignificant ? Please clarify this in the paper." In the paper, we report the means of both groups, as well as the results of the statistical test. The results don't show a "statistically significant difference" (p-value ¿ 0.05), and that is what we describe and discuss. The reviewer mistakes the difference between means with statistical difference of two distributions. Reviewer 2: "In the last paragraph of Section IV.C, two tests, i.e., t-test and Mann-Whitney U tests, are performed on the same dataset. The outcome of these two tests are different (one significant, and another insignificant). Why not perform Shapiro-Wilk test to see whether the data is normal and then only run one of these tests? Please provide more information in the paper." In the paper, we clearly mention that we test the data for normality. Based on the results, we choose two different tests for "two different datasets" (accuracy and duration data). The results are reported separately in two different sections for these two different dependent variables. Overall, I believe that this year's 2-phase review process was not as fair as previous years. If one reviewer had a negative feedback for any reason, it would cause a rejection. And there was no rebuttal phase to argue some issues that could easily be fixed by referring the reviewer

to the correct place on the paper or more elaboration.

One of these reviews is two sentences. Another rejects the paper because the problem it addresses is too difficult. None of the reviews comment on research methodology, logic, quality of literature review or reasonableness of conclusions. Get your act together and stop enlisting amateurs and as*** to do reviews.

The reviewers are simply commenting on the text, rather than agreeing or disagreeing with the central point and offering specific suggestions. Their disagreements are mainly tangential to the central point of the paper. Their comments reveal that they are engineers at heart, with limited understanding of science or its methods.

When one reviewer rejects a paper because "this is all wrong" and another rejects it because "everyone already knows this is true", maybe that paper is worth a second look and those reviews shouldn't be taken seriously.

I submitted four papers to various ICSE tracks this year. The reviews from the Technical Papers track were thorough, fair and constructive. The reviews from NIER and SEIS were unmitigated sh**.

Although our paper was not accepted, the reviews were quite thorough and shall be useful for advancing our work. Thanks for the valuable feedback.

On three reviewers only one provided valuable feedback to improve the paper.

I have no doubt about the reviewers' expertise - the reviews were clearly written to show it. Still this expertise was not in the specific domain where the paper's contribution is. They thus made their evaluation based on their a-priori expertise and expectations, not relevant to the submission and its competing related work. Also some reviewers clearly tried to promote citations to their own previous work, not justified by relevance to paper.

Based on the reviews, we had to conclude that there ICSE is not interested in looking beyond solving immediate problems of industrial software engineering.

It's hard to think of ICSE as a premier conference when there is no room for foundational research. We believe that will gain in stature with some exploratory papers in it, those that have not been evaluated as rigorously as papers on some of the more customary topics.

I think the reviewers failed to get the main point of my paper but their decision to reject was still justiied based on the lack of empiical evidence that the method proposed really works. I would recommend ICSE to use double-blind reviewing, like recently adopted by ECOOP and PLDI. Quote from review: "The problem the paper tackles is not exciting." I guess we should all write papers like Hollywood writes films? The work is actually quite exciting to me, my authors, and our industry partners. Maybe this reviewers is just poorly informed about the area. But I don't think this has a place in a review (given the other two reviewers felt i was quite interesting). These types of reviewers are what gives reviewing a poor name. In the long run, the reviews did not tell me what was scientifically wrong with the paper. Just that it wasn't exciting enough.

These were, by far, the best batch of reviews I have ever received in my four years as a graduate student. Uniformly, they were extremely clear and detailed. Criticisms were constructive in nature, and each was backed with an objective line of reasoning rather than a strong opinion. Our camera-ready version is a much better paper than what we submitted, largely because of the helpfulness of these reviews.

Since individual reviewers have little incentive to stand up for such papers, the program chairs should. May be there should be a column somewhere on the review form asking "is this paper off the beaten track", and if so, the program chair should try to use judgment whether that paper enriches the conference despite

its other shortcomings.