POPL 2014 Program Chair's Report

Peter Sewell
University of Cambridge
Peter.Sewell@cl.cam.ac.uk

Abstract

This note describes the POPL 2014 paper selection process and its rationale.

1. Overview

We begin in this section with a summary of the main points of general interest. Later sections flesh out some of these with more details; these may be of use for future chairs and steering committee members.

Judgement and Process There has recently been much discussion in the community about the process that POPL should follow, with questions of single vs double blind reviewing, sources of external reviewers, supplementary material, and so on. But it is important to remember that fundamentally we are asking individuals (the chairs, PC, and external reviewers) to exercise their best judgement, to identify the papers that make the most substantial advances for the subject, and that this is necessarily subjective, relying on their expertise and insight. The role of all the process machinery, with all its scoring, anonymity rules, etc., is simply to help them do that as well as possible, and no amount of careful process design can remove the central need for good judgement. Accordingly, while most of this note is concerned with details of the process, all that should be seen as fine-tuning.

Topics and a topic-balanced PC POPL is a broad conference requiring a wide range of expertise from its PC, and we aimed to construct a PC that had expertise in each area in proportion to the number of expected submissions involving that area. We started with an analysis of the POPL 2013 submitted abstracts, constructing a list of named topics. It seems useful to clarify that the point of these is not to let an author describe their paper (which can lead to some over-general or over-specific topics), but rather to identify the appropriate set of reviewers. Hence, each topic should identify a community of potential reviewers with some particular useful body of expertise. Future chairs might want to re-use the same topic list, evolving it gradually over time, so that submission frequency from one year is useful for constructing the next year's PC. Then in PC selection the PC and General chairs manually annotated potential PC members with topics, and we aimed for a PC in which each topic was represented twice as often as the expected number of PC reviews required. This worked out well, at least in so far as most papers had a reasonable number of bids; there were only a couple of identifiable areas where there were several (2–4) submissions and a lack of expertise in the PC.

Double-blind reviewing (DBR) Following POPL 2012, and the survey by Hicks as PC chair that showed a clear preference in the community, we adopted double-blind submission, but in even more "light weight" form: to avoid inhibiting normal scientific discussion, authors were explicitly permitted to discuss their work on mailing lists, and, to let PC members find external reviewers and assess the context of a paper, they were permitted to de-blind papers when necessary. The point of DBR here is to help the PC and external reviewers avoid first-impression bias where possible; it is not to make it hard for them to discover authors if they try. Additionally, reviewers who happen to guess or know the authors of a paper should not be inhibited in reviewing it.

The argument for some form of lightweight DBR seems to have been resolved and perhaps need not be debated further, though in future I would suggest it be relaxed in one particular way: any PC member should be allowed to de-blind a paper in order to suggest an external expert reviewer. A click-box in the conference management system for such de-blinding would help.

That said, DBR does come with a significant downside: it makes it difficult to use the knowledge of the whole PC to identify suitable expert reviewers. It is hard to assess the cost of this relative to the benefits of DBR.

External reviewers After the choice of PC, the most important thing for the conference is to find good expert reviewers — people who will really understand the context and contribution of each paper. This used to be devolved to the PC as a whole, but that is at odds with the combination of the shift to DBR and the desire to have PC members form their own views about their assigned papers, not just hand off responsibility to subreviewers. An alternative is to use a preselected External Review Committee (ERC), of around 60 people for POPL 2013 and 2012, for all or most external reviews, asking each to do around 4 reviews. For POPL 2014 we instead took the view (following POPL 2011) that it would be better to seek external reviews from the community at large, to maximise the expertise available and give the best chance of finding externals who really understood each paper. To distribute the effort involved, one member of the PC was designated as 'guardian' for each paper and was responsible, with the PC chair, for selecting one or two externals for it. Guardians were expected to de-blind the papers they were responsible for early in the process. We also invited authors to nominate up to five candidate reviewers, making it clear that these should not be contacted by the authors and that they might or might be used; this proved useful. External reviewers were actively encouraged to engage in the electronic discussion of their papers. In all, 273 individuals contributed external reviews, so we did access much expertise that would not have been available from an ERC.

The downside of this process is the load on the PC chair, which was manageable but significant. The conference management system did not provide support to load-balance review requests across externals, but we did not want to ask any individual to do more than two (rarely three) reviews, while some were in demand for many papers (and so should be used where their expertise was most useful). That meant the PC chair was involved in most external review requests: approximately 438 requests to 350 distinct people, with around 128 declines (one paper had 7 declines). Better conference management support would let much of this be distributed across the PC guardians, each of which was responsible for only around 8 papers.

Supplementary material Making supplementary material (such as detailed definitions and proofs, proof scripts, or experimental data) available via URLs is problematic: there is a potential loss of confidentiality (or less informed reviews, if the reviewer avoids downloading material just because of that); sometimes there are errors in the URLs; and some authors game the system by completing or updating their supplementary material after the submission deadline, which is unfair to those that do not. We therefore required supplementary material to be uploaded *at submission time*, as a tarball or single pdf, which worked well (Eddie Kohler kindly added HotCRP support for this). Authors were also reminded to highlight their supplementary material. For material that is intrinsically non-uploadable (e.g. links to a running system on the web) URLs were permitted in the supplementary material. For simplicity, supplementary material was not required to be anonymised and was made available to reviewers only after they submitted their first-draft review. In future one should support both anonymised and non-anonymised supplementary material.

Bidding and assignment To clarify the bidding process we fixed on a particular semantics for bidding scores so that PC members could identify the papers that they would like to (or should) review. A first-draft assignment of papers to PC papers was made automatically (using the built-in HotCRP algorithm) and then manually tuned by the PC chair.

Page limits and deadlines The submission deadline and page limits were rigorously enforced, for fairness. Authors who had uploaded papers exceeding the page limit were warned on the day of the deadline, and a grace period of a few minutes was permitted to avoid any debate over clock skew. The abstract registration deadline was treated more liberally, as a number of authors had failed to read the deadline, despite it having been prominently stated, with world clock links, on the CFP. The deadlines were at close-of-business in the PC chair's timezone to ensure that support staff were available in case of problems with the submission server.

Scoring The usual ABCD/XYZ scoring was combined with scores for goals, execution, and presentation, calibrated with respect to typical accepted POPL papers; this seemed to be helpful in focussing reviews and discussion.

Author response Authors were asked to put the main points first in their response and encouraged to keep it short but no hard length limit was imposed (though a number of authors did misinterpret the soft limit as a hard one). Late-arriving reviews were manually sent to authors for quick responses.

PC meeting There was significant electronic discussion between author response and the physical PC meeting.

By the time of the PC meeting it becomes tempting to regard the scores that a paper has as an accurate and absolute measure of its quality (e.g. with remarks like "this is an AA paper"). But if that were true there would be no point in further discussion and one could just pick the accepted papers as the top of the score rank order: the whole point of the meeting is to discuss cases where the scores are not sufficient. Accordingly, we tried to focus in the meeting on reasons rather than scores. There also wasn't much discussion of authors as individuals.

The PC meeting was organised in two phases: a complete pass of 102 papers in a random order (but with related papers adjacent), on the first day and morning of the second, then a review of 17 papers that had been left on the accept/reject border and 7 more that PC members wanted to revisit in the early afternoon of the second. This worked well, but it would have been better still to more clearly identify and discuss the class of 'perfectly acceptable' papers: those for which there is a consensus that a reasonable PC could accept, even if there are some negative points. These seem to be those for which there is the most randomness in the decision making.

The PC and General chairs fixed an upper bound of 55 papers, both to make the conference schedule workable without requiring talks to be over-short and to ensure the acceptance rate did not go above 25%. In the event we reached a natural total of 51 accepts.

The accepted papers were rather far from a prefix of the pre-PC-meeting rank order (sorted by score counts per the HotCRP default): the highest rank of a rejected paper was 26 (with two As) and the lowest rank of an accepted paper was 95 (with three Bs).

PC submissions and chair conflicts PC submissions were permitted, otherwise there would be too large an impact on the students and colleagues of PC members, making some reluctant to serve. But PC members were not involved in any way in the discussion of PC papers, and indeed were not even informed of the outcomes for their papers until other authors were. Instead, the PC and General chairs together identified external reviewers (typically four) for each PC-member submission, moderated the electronic discussion among them, and came to a conclusion.

The *Principles of POPL* document states that *"SIGPLAN requires that PC papers be held to a higher standard than other papers. For POPL, the criterion for acceptability of a PC paper is clear accept."* It is debatable whether such a condition is appropriate for conferences where the PC are not involved in decisions of PC papers —it is arguably unfair— but in any case we were satisfied that it held.

We also had to deal with a number of papers for which the PC chair had a conflict. For these, the General Chair managed the selection of external reviewers, moderated the electronic discussion, and (except for those that were also PC submissions) appointed a deputy chair to handle that part of the PC meeting. This provided welcome clarity but did mean that the PC chair could not provide overall calibration for those papers.

Feedback to authors The main focus of the process was on making the best decisions, not on providing the best critical comment back to authors. Nonetheless, after investing (in many cases) considerable effort in electronic and physical discussion of a paper, it is wasteful to simply discard it, and authors of rejected papers typically prefer to know as much as possible of the reasons why. Accordingly, the PC meeting was assisted by two scribes (one for each day) who took notes about the discussion or reasons for decisions, and the guardian of each paper was asked after the meeting to add an author-visible comment summarising any of the discussion that would be useful. This could be improved still further by more explicitly recording the reasons for the final decision in each case. For many PC and reviewer comments from the electronic discussion there is no essential reason why they could not be made visible to the authors at author response time or after the process, if written with that in mind, and it would be worth encouraging the PC and reviewers to do so (and tag them as such in the system).

Semantic Mechanisation Survey We repeated Benjamin Pierce's survey (with minor changes) on the use of mechanised proof from POPL 2009, with some simple click boxes on the submission page. Around 10% of submissions were completely formalised, slightly more partially formalised, and the acceptance rates for these were in line with those for submissions as a whole. There does not seem to be a significant change in these proportions since 2009, though the absolute numbers of submissions are higher.

Author Survey We sent a survey to the authors (between author response and PC meeting), asking (a) their views on the double-blind process and the extent to which it had affected their behaviour, and (b) whether they thought the reviews would be helpful in improving presentation or technical content and future research,

and (c) whether they thought the reviewers understood their submission well enough to come to an informed judgement. The last is the most important: this is asking whether authors think that we (as the POPL organisation, program committee, and external review community) are doing a satisfactory job.

The authors are clearly in favour of a DBR process: 67% *yes* or *strong yes* vs 15% *no* or *strong no* (and 18% *don't care*). That said, many authors will not be in a position to assess the impact of a DBR process on finding expert reviewers who really understand each submission.

The detailed policy on what is permitted is clearly not getting across to all authors: a significant fraction (25% of survey respondants) refrained from putting the paper on a web page because of DBR, despite this being explicitly permitted in the 2014 DBR FAQ.

Most (86%) respondants thought the reviews would be helpful in improving presentation, and a smaller but reasonable fraction (62%) thought that the reviewers' remarks would be *"helpful in improving the technical content of their paper or for their future research"*.

Finally, for the main question, of whether they thought the reviewers would collectively understood their submission (taking the author response into account) well enough to come to an informed judgement, 60% replied *yes* or *strong yes*, with 29% *"not sure"*, and 11% *no* or *strong no*. Without comparable data from other conference instances, it is hard to know whether one should consider this good or bad. The fact that 40% were not confident that the reviewers will understand their work is less than ideal, at least.

As for the 11% that answered *no* or *strong no* (referring to 22 papers), a priori these might represent either cases where we have made a serious error in the review process or dissatisfied authors of submissions that in fact were not up to the required standard. Looking at the reviews and discussion of these papers, there are some clear cases of the latter, a few cases where an extra review came in late, and several cases where there was extensive discussion (and hence where different reviewers or PC members might have come to a different conclusion).

Acknowledgements It was instructive and a pleasure to serve as POPL 2014 PC chair. I would like to thank the General Chair, Suresh Jagannathan, for his unstinting support, the POPL SC for their advice (especially Mike Hicks and Mooly Sagiv), the program committee and external reviewers for their sterling work, Eddie Kohler for providing and supporting the HotCRP conference management system, Piete Brooks for managing our HotCRP installation, and Gabriel Kerneis and Dominic Mulligan for their assistance at the PC meeting. And, of course, all the authors of the submitted papers.

2. Topics

Conference management systems typically give authors a list of click-box named topics for paper registration. These might be useful for:

1. bidding
2. paper assignment
3. program scheduling
4. subject-balancing the next PC

For POPL 2014 topics were not really useful for bidding or assignment: the paper titles and abstracts provide much better information for PC members and the PC chair. They were useful in scheduling the program, putting related papers in the same session and avoiding clashes. But the main use was in building a subject-balanced PC. POPL is a broad conference, and we wanted to construct a PC that had expertise in each area at least in proportion to the number of expected submissions involving that area.

It seems useful to clarify that the point of named topics is not to let an author describe their paper (and especially not to let them completely describe their paper), but rather to identify the appropriate set of reviewers. Hence, each topic should identify a community of potential reviewers with some particular useful body of expertise, so that you can say, for some particular paper, "an appropriate reviewer for this paper should know something about X". Some topics that have been used are too generic for this, e.g. *"operational semantics"* or *"language design"*, and some are too specific, with only a few reviewers and papers matching them. To produce a good set of topics and to get data on how many submissions there might be for each, we went through all the submitted abstracts from POPL 2013 (which just had free-form topics) and manually abstracted them. This was time-consuming and surely imperfect, but instructive; combining it with discussions with a few experts in areas we had less expertise in gave a set of topics to use for POPL 2014. The topics are shown in Fig. 1 together with:

- the PC-chair annotated counts from 2013 and the author-annotated counts from POPL 2014 (in some cases significantly different);
- the number of 2014 accepted papers (and acceptance rate) for papers tagged with each topic;
- the number of PC members who could cover each topic, both as initially identified by the PC and General chair during PC selection and also as self-identified by the PC members at the start of bidding (a big difference here shows the chairs' misjudgements; PC members generally self-identified as being able to cover more topics than we guessed); and
- the number of submissions per PC member who had self-identified as able to cover that topic.

Note that this is all treating each topic in isolation. Topics are obviously correlated and it might be worth taking that into account in the quantitative PC balancing.

Future PC chairs might want to re-use the same set of topics to give data that is comparable from year to year, evolving it gradually as the subject changes.

Topic balancing appeared to worked out well: most papers had a reasonable number of bids and in constructing the paper assignment it was rather often the case that there were three available high bids. There were only a couple of identifiable areas where there were several (2–4) submissions and a lack of expertise in the PC.

3. PC selection

Recent PC sizes from 2009–13 have been 25,20,26,27,27. For 2014 we aimed for 28 people (not including the chair), which for the 237 submissions of POPL 2013 and 3 PC reviews/paper would be 25 reviews per PC member, and a guardian load of 8.5 papers per PC member. In the event one PC member had to withdraw at a late stage, leaving us with a PC of 27, and there were 220 submissions, giving a reviewing load of 21–25 reviews per PC member.

We considered that POPL PC members should normally have previously published in the conference, and should not have served on the PC too recently. To build an initial pool of potential PC members, we used code from Mike Hicks to pull POPL authorship data from DBLP for 2000-2012, and collected PC chairs, PC membership and ERC membership from 2007–2013. Combining these and hand-normalising variants of names in the data gave a file with names associated to sets of tags (`POPL:nn`, `POPL20nn-PC-CHAIR`, `POPL20nn-PC`, and `POPL20nn-ERC`).

POPL 2014 topic	POPL 2013 submissions	POPL 2014 submissions	POPL 2014 accepted	POPL 2014 acceptance ratio	POPL 2014 PC members (pre-identified by chairs)	POPL 2014 PC members (self-identified)	POPL 2014 papers / PC member
(all program analysis)	63				10		
Static Program Analysis		72	17	.24		10	7
Abstract Interpretation		30	8	.27		6	5
Dynamic Program Analysis		15	4	.27		6	3
Model Checking		17	5	.29		5	3
Decision Procedures (including SAT and SMT)		22	8	.36		3	7
Shared Memory Concurrency	47	26	5	.19	10	6	4
Message Passing Concurrency	13	18	3	.17	2	6	3
Type Systems (including Inference, Type Theory)	44	65	17	.26	13	14	4
Types and Effects		29	5	.17		14	2
Program Logics	25	44	10	.25	5	16	3
Proof Assistants	13	18	4	.22	7	9	2
Semantic Models (Logical Relations, Categories, Domains, etc.)	23	56	16	.29	6	7	8
Functional Languages	30	70	21	.30	7	9	8
Object Oriented Languages	19	20	2	.10	4	5	4
Dynamic Languages	9	12	4	.33	1	3	4
Security	17	22	5	.23	4	7	3
Complexity	15	10	1	.10	1	1	10
Compiler Optimisation and Design	13	21	5	.24	3	3	7
Synthesis	15	17	3	.18		7	3
Verified Compilation	5	18	5	.28	2	10	2
(none)		4					

Figure 1. Topic Counts

Ideally most PC members would combine a broad view of the subject together with depth in several specific areas, together with the essential but impossible-to-quantify good judgement. Then there are things one can quantify. A good PC should be balanced in several ways:

- by topics, in proportion to the expected submissions;
- by gender, as far as possible;
- by institution, without too many from the same institution;
- by country, with the major POPL local communities represented; and
- by seniority, with a good mix of junior and senior members.

We regarded judgement and topic-balancing as the most important of these. Discussion between the PC chair and general chair gave us around 45 initial candidates (others were added in later discussion with the SC). We annotated those with additional tags, for topic expertise, gender, institution, country, and seniority, and wrote a simple script to analyse proposed subsets of these people by those five criteria, listing those with each tag and comparing with the numbers we were aiming for; that was invaluable as we fine-tuned the PC, producing a report for each change under consideration. We also considered whether any one would be superseded by another, whether there were too many close colleagues of the chair, and how we thought they would interact electronically and in the PC meeting.

We spent most effort on balancing by topics. For a topic with N submissions in POPL 2013, we aimed for a minimum of $N * 3/28$ PC members with expertise in that topic, and in most cases twice that (equivalently, to compare with the last column of Fig. 1: for each topic at most 9 papers for each relevant PC member and ideally no more than 5).

For balancing by institution, there was discussion with the SC and with the SIGPLAN chair and co-chair about what policy is appropriate for large multi-site institutions such as MSR, IBM Research, and INRIA, which is a sensitive question (POPL 2013 ended up with 6 from MSR, which some argued was excessive, though pro rata per acceptances or per submissions would have been 5.8 and 4 respectively). For POPL 2014 we had 3 MSR members after PC selection, but another PC member added an MSR affiliation during the process.

Of the first round of PC invitations, 23/28 accepted (note that this took up to a month). We proposed a second round to the SC after most of the responses were in, of which 5/6 accepted.

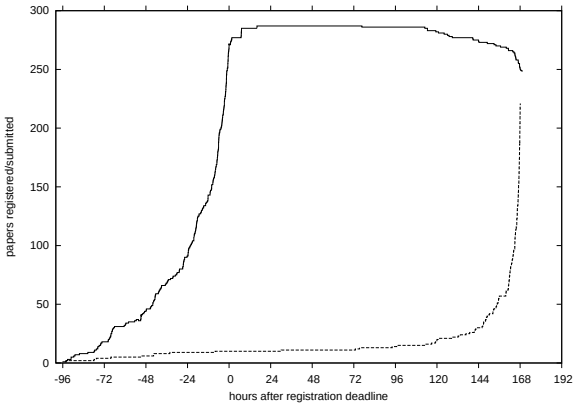

Figure 2. Numbers of registered (solid) and submitted (dashed) papers

4. Submission numbers

There were approximately 288 papers registered, which turned into 220 actual submissions (76% of those registered). 4 papers were withdrawn between submission and the start of the author response period (two by the PC chair, for duplicate submissions, and two by the authors). That gave 216 papers in play. 7 papers were withdrawn by authors after the start of the author response period (explaining the confusing "209 submitted" that HotCRP now reports). 51 papers were accepted, which gives an acceptance rate of 51/220 or 23%. Figure 2 shows the registration and submission numbers over time.

5. Single vs Double-blind Submission

We adopted a lightweight double-blind process. The choice between single- and double-blind has been extensively debated in the community in the last few years, and the surveys by Mike Hicks for POPL 2012 show a clear preference in the community for this, which we found persuasive and which was confirmed in the POPL 2014 author survey. In more detail:

Paper submissions were required to have the author names unlisted, references to previous work in the third person, and so on. We made clear in the CFP and a DBR FAQ that authors were not required to "hide" their submissions – they could put them on their web pages, give talks about them, etc., as usual. In a difference from POPL 2012, posting or discussing papers on mailing lists was explicitly permitted. The point of DBR here is to help PC members review papers with minimal bias, not to make it hard for them to discover authorship if they try. There seems to be some confusion on this point in the community, with several people asking whether the fact that they could identify the authors of a paper meant that they should not review it. There is also some confusion about what authors are allowed to do in a DBR regime, as we see in the author survey.

Authorship was revealed to the reviewing PC member after they submitted a review (which they could subsequently update), and PC members could, if they think it necessary, submit placeholder reviews, e.g. to view non-anonymous supplementary material.

Guardians necessarily had to de-blind their papers early to identify suitable external reviewers. This was also done by submitting placeholder reviews (including the "PLACEHOLDER" keyword to make searching possible). This did make monitoring the review process more awkward, and explicit conference-management-system support for optional de-blinding would be helpful.

Three submissions were not anonymised, apparently by oversight; they were reviewed as normal.

6. An ERC or external experts?

We chose not to have a preselected external review committee (ERC), but instead to proactively seek out expert reviewers early and throughout the review period and to ask them directly, and also to involve them in individual paper discussions. Our rationale for this was that restricting external reviews to members of a preselected ERC limits possibilities for finding the right expert reviewers, and in the PC meeting ERC reviews are effectively similar to external reviews in any case. The experience in POPL 2011 was that people were largely rather responsive to direct requests for expert reviews, which suggested that this would be feasible. The other main reason put forward for an ERC, of reviewing PC submissions, was handled by the chairs seeking external reviewers directly.

To help identify externals, we asked the PC during bidding (and later) to suggest candidates, for any papers where they know one or more people who "should" review the paper. We expected each PC member to suggest one or two candidates for each paper that they bid for, for example. This is a modest extra load on the PC members early in the process, but it seems to be reasonable.

We also invited the authors to nominate, at submission time, a list of up to five candidate reviewers that they think would be experts. This is a departure from normal practice for conference reviewing, though common elsewhere (e.g. for grant applications). We made it clear that these suggestions might or might not be used, at the discretion of the PC and PC chair. We also made clear that the authors should not contact their suggested reviewers directly:

> *"Please list the names and emails of up to 5 potential reviewers that you believe have expertise in the area of this paper. Do not include any PC members or people who would be conflicted, and please do not discuss this with your suggested reviewers (in particular, do not ask them whether they would be prepared to review the paper). The PC may or may not call for reviews from any of those suggested."*

During paper assignment, the PC chair identified a "guardian" among the PC for each paper, tasking them with selecting external reviewers (aided by the above data, and by the PC chair if necessary). HotCRP now has some support for this ("paper managers"), added after POPL 2014 was underway.

The PC were asked to write their own reviews rather than farm them out to subreviewers, but to suggest additional external reviewers (to the guardian and PC chair) during the process, if it became clear that an additional opinion would be helpful. In the event most PC members did this. A few did not, asking subreviewers directly and combining them with their own reviews. In hindsight, that should be permitted — the important point is that PC members should form their own opinion, and sometimes that might best be done with such a combination. Though it does need a better mechanism for load-balancing across externals: sometimes the effort of a particular external would be better spent on a different paper.

External reviewers were allowed and encouraged to participate in the electronic discussion of the papers they have reviewed (but they were not allowed to see all the rest of the HotCRP data, e.g. reviewer assignment for other papers).

We aimed for three PC reviews and at least one external review for each paper. In cases where the PC lacked expertise we sought additional external experts, and in a few cases we deemed three to be sufficient. Out of 220 submisions, the majority of papers (146, 66%) received four reviews; 53 (24%) received five, 15 (7%)

received three, and 2 (1%) received six (the remaining 4 were withdrawn before the author response period). A small number of additional reviews were solicited close to or after the rebuttal period; in those cases the PC chair forwarded the reviews directly to the authors for a quick response.

For PC submissions the PC and General chairs sought to identify four external reviews for each paper.

The conference management system did not provide support to load-balance review requests across externals, but we did not want to ask any individual to do more than two (rarely three) reviews, while some were in demand for many papers (and so should be used where their expertise was most useful). That meant the PC chair was involved in most external review requests: approximately 438 requests to 350 distinct people, with around 128 declines (one paper had 7 declines). Better conference management support would let much of this be distributed across the PC guardians, each of which was responsible for only around 8 papers.

In all, 273 individuals contributed external reviews, so we did access much expertise that would not have been available from an ERC.

Looking at the replies to review requests, it is clear that the aggregated load of reviewing across multiple conferences is a problem and many candidate reviewers are overloaded. Of the declined requests (simplifying somewhat), 68 were overcommitted, 18 considered themselves conflicted, 16 considered themselves insufficiently expert, 9 did not comment, and 6 were on vacation. All the individuals contacted did eventually respond (either positively or negatively) except for around 10.

7. Review and Reviewer Analysis

Of the 216 papers in play at the start of the author response period, there were 50 for which the authors did not suggest reviewers. There were 704 suggestions in total, covering around 400 distinct people. Some were very popular (one reviewer was suggested 12 times) while 282 were suggested once.

The author-suggested reviewers often had a large overlap with those thought of by the PC or chairs, but by no means always. Sometimes they added names we would not have thought of, while sometimes they seemed to be unreasonably close to the authors. We ended up with at least one review from among those suggested for approximately half of the papers for which the authors did make a suggestion; 90 of 299 non-PC reviews were by reviewers in the author-suggested lists.

One might imagine that the author-suggested reviewers would be uniformly positive, but that was not at all the case. Figure 3 compares the reviewer expertise and overall merit scores between the different kinds of reviewers: PC members, externals who were suggested by the authors, and other externals. The suggested reviewers were somewhat more likely to count themselves as expert and did give a somewhat higher percentage of As, however — emphasising the need to treat review scores with caution and look closely at the review texts. One can also see that external reviews (of both kinds) more often had X and A scores than PC member reviews.

In total there were 907 reviews, divided among 431 (48%) X reviews, 344 (38%) Y reviews, and just 132 (15%) Z reviews. That seems reasonable for a conference as broad as POPL, though obviously one would like as high a proportion of X reviews as possible.

65% of the 217 submissions that were reviewed received at least two X reviews:

5X	4	2%
4X	17	8%
3X	48	22%
2X	71	33%
1X	57	26%
0X	20	9%

(note also that a number of the 0X and 1X papers actually had high-confidence reviews).

8. Conflicts of interest

We relied on authors to identify their conflicts of interest, with tick-boxes on the submission page for PC members and a free text field for others. We also asked PC members to list their conflicts, as a backup, but that information was not used in the end (which was confusing on a couple of occasions).

Several authors did not identify all their conflicts (a few did not list any), so we did discover some conflicts late, after a review was written. Sometimes we re-assigned papers among the PC because of this, which was awkward but manageable.

In future, one might explicitly require conflicts to be entered at paper registration time and do some sanity checking of those before the submission deadline, allowing the PC chair to prompt authors. But a better solution is really needed, and ideally it would make use of the existing co-authorship information available from databases such as DBLP or the ACM DL so that authors and PC members do not need to duplicate that.

9. Bidding

HotCRP just provides an arbitrary numerical scale for bidding, but to give a clear semantics for bidding scores, to help the PC chair manually tune the paper assignment, we asked the PC to use these values for bidding:

- 20 : "I really want to (or really should) review this paper"
- 11 : "I have expertise in the area and would be prepared to review this paper"
- 10 : "I'd quite like to review this paper"
- 0 : "I could review this paper"
- -20 : "I really don't want to review this paper"
- -100 : "I have a conflict with this paper"

10. Scoring

Here we combined the usual ABCD/XYZ scoring as below with additional questions:

- *Goals: Are the authors trying to do something worth doing (in the POPL context)?*
- *Execution: Have the authors done what they attempted well? (Is the work mathematically rigorous and elegant, experimentally solid, and so on, as appropriate to the topic.)*
- *Presentation: Is the work presented well?*

each on a scale of:

1. *The [X] is lacking.*
2. *The [X] is fine, to the standard of a serious conference, but not competitive here.*
3. *The [X] is good. To the standard of a perfectly acceptable POPL paper.*
4. *The [X] is great! To the standard of the best third of previous POPL accepted papers.*

PC	608	A: 72 (12%)	B:196 (32%)	C:249 (41%)	D: 91 (15%)	X:211 (35%)	Y:274 (45%)	Z:123 (20%)
Ext:suggested	90	A: 29 (32%)	B: 31 (34%)	C: 23 (26%)	D: 7 (8%)	X: 79 (88%)	Y: 10 (11%)	Z: 1 (1%)
Ext:unsuggested	209	A: 36 (17%)	B: 62 (30%)	C: 82 (39%)	D: 29 (14%)	X:141 (68%)	Y: 60 (29%)	Z: 8 (4%)

Figure 3. All review scores, by kind of reviewer

The idea here was that the lexicographic order of those should focus discussion, so, for example, if something failed on Goals, discussion could stop at that point. They seemed to be useful in prompting reviewers to explain their judgements more clearly. Note that the text provides a calibration that reviewers familiar with POPL can share, rather than attempting some absolute but vague measure.

For the usual ABCD/XYZ scoring, we adapted the text of the former slightly, to speak of "having expertise" rather than "being an expert", as below. Experience shows that reviewers often understate their expertise. It might be preferable to have a more explicit "confidence" score instead of expertise.

Reviewer expertise

X. I have expertise in the subject area of this paper.

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.

Please summarise your judgement about whether the paper should be accepted for this POPL.

A. Good paper. I will champion it at the PC meeting.

B. OK paper, but I will not champion it.

C. Weak paper, though I will not fight strongly against it.

D. I will argue to reject this paper.

11. Reading the papers

As PC chair one has a choice of how much of the submissions and reviews to read, and whether to review any papers personally. For 2014, that was:

1. reading all the paper abstracts for paper assignment (looking at the actual papers where necessary, as the abstracts unfortunately often do not give a good sense of what is in the paper);
2. looking in more detail at papers where necessary to identify good candidate external reviewers;
3. reading all the paper introductions, and sometimes the next section or more, before the PC meeting;
4. properly reading very few papers (less than ten);
5. writing no paper reviews; and
6. reading all the reviews and comments as they arrived and sometimes prompting further discussion.

Points 3 and 6 were time-consuming but worthwhile. Though having done that, but still not properly having read the papers, it was necessary to be cautious not to rely on one's first impressions, or over-weight them with respect to the judgements of the PC members and externals.

12. PC meeting

There was a physical PC meeting, as is usual for POPL, with electronic discussion beforehand. The meeting was in Cambridge UK, though we also considered an East-coast USA location to minimise total travel. We rescheduled the PC meeting slightly to avoid ICFP (whose dates were announced between when we created our schedule and when PC members were invited). One PC member was unable to attend in the end, so we tried to advance the electronic discussion for their papers and designated alternate guardians to lead the physical discussion.

The PC meeting discussed 102 papers: all of those except PC submissions that had at least two Bs for which a clear consensus to reject had not been reached in the pre-meeting electronic discussion. That seems to be a fairly robust boundary.

By the time of a PC meeting it becomes tempting to regard the scores that a paper has as an accurate and absolute measure of its quality (e.g. with remarks like "this is an AA paper"). But if that were true there would be no point in further discussion and one could just pick the accepted papers as the top of the score rank order: the whole point of the meeting is to discuss cases where the scores are not sufficient. Accordingly, we tried to focus in the meeting on reasons rather than scores. We also were also quite prepared to accept papers that at the start of the meeting did not have a champion, to avoid leaving good papers that happened only to have received B scores on the floor.

The discussion was in two passes. The first pass classified papers into *accept*, *'acceptish'*, *'rejectish'*, and *reject*, exploiting the existence of the second pass to cut short discussion by making a tentative decision for the middle two. The second revisited those two categories along with papers that any member of the PC wanted to revisit (allowing some time for them to consider, but perhaps not quite enough).

To reduce expectation bias, the first-pass discussion order was random (though fixed before the meeting) except that: (a) we started with two probable-accept papers, to set a good tone; (b) groups of related papers were discussed together (17 such groups, of 2–7 papers each, had been identified during the review process, of which 9 groups survived to the PC meeting); and (c) the PC-chair-conflict papers were discussed together and chaired by a member of the PC selected by the General chair. The second pass went through the 'acceptish', 'rejectish' and revisit papers, changing the decision for a few of each. In hindsight, we should also have kept more careful track of all the papers for which there was a consensus that they were in principle 'acceptable' and ensured that they were all reconsidered; that would have added just a few to the set.

We maintained the normal conflict-of-interest protocol, with conflicted individuals (including the PC chair) leaving the room as necessary, throughout the meeting. It was thought that this might be too awkward for the second pass, but in fact it was fine and much preferable to the converse.

The overall schedule turned out as below (with a few short breaks of 10–20 minutes not noted). Changing the rate dynamically is difficult without damaging the discussion, but it was important to keep focussed on the accept/reject decisions and to make transitions quickly when decisions became clear.

pass		time	papers	minutes/paper
1	Sunday	9.05 – 12.50	34	6.6
		1.35 – 16.00	26	5.6
		4.30 – 18.30	17	7.1
	Monday	9.10 – 12.45	25	8.6
2		13.30 – 15.30	24	5.0

All non-conflicting scores and review data was made visible to the PC at the same time as reviews are sent to the authors, and they were encouraged to contribute to the discussion of any papers where they were particularly expert.

In the end the accepted papers were rather far from a prefix of the pre-PC-meeting rank order, as one can see in Fig. 4: the highest rank of a rejected paper was 26 (with two As) and the lowest rank of an accepted paper was 95 (with three Bs). (These ranks are with respect to the default HotCRP sort order, by counts of high scores.)

It may also be interesting to look at acceptance rates for papers with different numbers of A reviews (these counts are from the post-PC-meeting scores, but few were changed).

	submitted	accepted
$\geq$ 3 A	13	13 (100%)
$\geq$ 2 A	34	29 (85%)
$\geq$ 1 A	85	48 (56%)
0 A	131	3 (2%)

13. PC workshop

We took advantage of the presence of the PC to host a one-day workshop after the meeting, which proved popular (around 10 talks by PC members and good attendance from nearby researchers).

14. HotCRP

The HotCRP conference management system was invaluable. We switched from an install from a release version to one from the github repository, to take advantage of fixes to bugs identified during the process. We tested the configuration with a small dry-run, taking a few dummy papers through the process with a couple of colleagues. There were two failures of the system, one in the night before submission (when it was unattended) and one of a few minutes on the day of submission. The latter was associated with a rise in memory usage for no apparent reason. Fine-tuning of the paper assignment and management of external reviewers (keeping track of review requests and declines for each paper and external reviewer) were done off-line with some simple text files and scripts, as HotCRP did not seem to be sufficiently flexible. The install ran as a part of a small virtual server. It used around 1GB of disc for mysql and was configured with 1–2GB memory.

15. Pearls?

ICFP has traditionally had a category of "Functional Pearl" submissions: "elegant, instructive, and fun essays on functional programming". For POPL, the CFPs of recent conferences have also had a Pearl category, but it is not clear that it works well. POPL is broader, so the idea of a POPL pearl seems to be a paper that "explains an old idea in a new way". But POPL papers have many different kinds of contribution, and for none of the others do we have special categories; we believe that it suffices for authors who wish to argue that they have made a substantial contribution with a new presentation of an old idea to do so in the normal way, in their abstract and text. POPL 2013 had just one submission mentioning "pearl" in its title, which was rejected, and no others mentioning "pearl" in their abstracts. POPL 2014 had three submissions mentioning "pearl" in the title and a couple of others that were not self-identified as Pearls but where the concept came up during PC discussion; these are very small numbers compared with the total number of submissions, and none of them were accepted. On the whole the concept confused the discussion of those papers rather than clarifying it.

16. Other remarks

One should be clear on the web page and text CFP what is mandatory at paper registration time, and ideally ensure that the conference management system enforces it. This should include title, abstract, authors, topics, and conflicts. One should also be clear whether the conference management system permits submission updates. The text and web page CFPs need to be carefully checked against each other. One might consider requiring authors to include the paper number prominently in the frontmatter of each submitted paper. Several authors forgot to press the "submit" button on an apparently complete response; the PC chair did that on their behalf.

17. Timetable

The detailed timetable for POPL 2014 as it happened, from the PC chair's point of view, is in Fig. 5. A few things would have been better done earlier: the first round of PC invitations (as one or two candidates had already accepted too many other commitments), the program schedule and session chairs, and the invited speakers (we discussed at the PC meeting briefly then by email over the next few weeks).

It would be desirable to have more time for external reviewers, especially given that the review period overlaps many people's holiday period. One might consider extending the whole process by a week, and/or asking the PC to identify externals earlier.

18. Semantic Mechanisation Survey

We repeated Benjamin Pierce's survey on the use of mechanised proof from POPL 2009, with some simple click boxes on the submission page:

We would like to discover how many POPL submissions were developed using a proof assistant or other mechanised semantics tool, to express their definitions and/or to mechanically check the proofs.

- *No response. Check this box if you prefer not to answer this question.*
- *No. Check this box if you have not used a proof assistant or related semantics tool in this paper.*
- *Partly. Check this box if you have used a proof assistant or related semantics tool in some way in developing the results in your paper - e.g., for formalising and sanity-checking definitions.*
- *Completely. Check this box if the proofs of your main results have been fully mechanically checked.*

The answers were not visible to PC members or reviewers.

2014	submitted		accepted		acceptance rate
no response	73	33%	15	29%	21 %
no	97	44%	26	51%	27 %
partly	28	13%	5	10%	18%
completely	21	10%	5	10%	24%
total	219		51		23 %

Summarising, around 10% of submissions were completely formalised, slightly more partially formalised, and the acceptance rates for these were in line with those for submissions as a whole.

For comparison, here are the questions and responses from 2009:

- *Check this box if you have used a proof assistant in some way in developing the results in your paper – e.g., for formalizing and sanity-checking definitions. (Your responses to these questions will be used only for informational purposes; they will not affect your chances of acceptance.)*

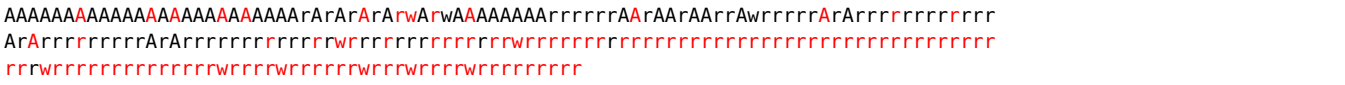

Figure 4. Outcomes in rank order (Accept, reject, withdrawn; those in black were discussed in the physical PC meeting)

- *Check this box if the proofs of your main results have been fully mechanically checked.*
- *Check this box if you may or may not have used a proof assistant in some way but prefer not to say which.*

There was also another option to explicitly decline to specify.

2009	submitted		accepted		acceptance rate
No response	112	70%	27	75%	24%
Declined to specify	5	3%	2	6%	40%
Machine assisted	30	19%	5	14%	17%
Fully verified	12	8%	2	6%	17%
Total	159		36		23%

There seems to be no major difference between the 2014 and 2009 proportions (the absolute numbers are slightly higher in 2014), though anecdotally formalisation is more common than it was.

19. Author Survey

We sent a survey to the authors, asking (a) their views on the double-blind process and the extent to which it had affected their behaviour, and (b) whether they thought the reviews would be helpful in improving presentation or technical content and future research, and (c) whether they thought the reviewers understood their submission well enough to come to an informed judgement. The last is the most important: this is asking whether authors think that we (as the POPL organisation, program committee, and external review community) are doing a satisfactory job.

The survey (a link to a Google Docs survey) was sent between the author response period and the PC meeting, so they had seen the reviews (and written any response) but did not yet know the decisions, with the idea that this would give the most accurate assessment.

The questions were discussed with the SC in advance, who suggested additional questions and wording changes. The whole was kept short (just 8 substantive questions) to keep the response rate high.

The survey was sent to all corresponding authors. Practice varied: some papers had just one corresponding author whereas others had all their authors listed. Some papers are therefore multiply represented in the results, but normalising with respect to this does not affect the conclusions.

The survey was not anonymous: the first three questions asked for the paper number, paper title, and responder name, so that we could identify multiple responses for a paper and correlate the answers against paper outcomes. However, authors were told that

> *The responses won't be shown to anyone except the Program and General chairs except in anonymised summary form; in particular they won't be shown to the POPL 2014 PC before the PC meeting and won't affect the decisions for POPL 2014.*

It is possible that one would get more or more detailed or accurate answers for an anonymous survey, but the high response rate suggests otherwise. Two responses were anonymous despite the instructions.

In total there were 236 responses, covering 159–161 (74%) out of the 216 submissions in play at the start of the author response period.

19.1 Survey Questions and Results

This subsection gives the survey questions and result data. Our conclusions from that are in the next subsection.

Rubric: *POPL 2014 Author Survey. This is a short survey for authors of POPL 2014 submissions about the POPL process, to help assess and improve it for future years. The responses won't be shown to anyone except the Program and General chairs except in anonymised summary form; in particular they won't be shown to the POPL 2014 PC before the PC meeting and won't affect the decisions for POPL 2014. Please complete it by 26 September.*

1 *Paper number*

2 *Paper title*

3 *Your name (please ensure this exactly matches what is in the HotCRP system)*

4 *The lightweight double-blind process has advantages (helping reviewers avoid initial bias) and costs (in writing the paper for authors and in finding nonconflicted reviewers for the PC). Do you think future POPLs should use it?*

strong yes	53	23%	67%
yes	106	45%	
don't care	42	18%	18%
no	25	11%	15%
strong no	9	4%	

5 *Have you done any of these? (select any/all that apply)*

- *talking informally about your work outside your immediate group.* 183 (78% of responses)
- *giving talks about your work.* 92 (39%)
- *putting the paper on a web page.* 66 (28%)

6 *Did the double-blind process affect your actions by causing you to not do any of these? (select any/all that apply)*

- *talking informally about your work outside your immediate group.* 12 (5%)
- *giving talks about your work.* 21 (9%)
- *putting the paper on a web page.* 60 (25%)

7 *Seeking more reviews per paper is informative but is a significant load on the community. For this paper, do you think we had:*

- *unnecessarily many reviews.* 7 (3%)
- *a reasonable number of reviews.* 209 (89%)
- *too few reviews.* 20 (8%)

8 *Will the reviewers' remarks be helpful in improving presentation?*

strong yes	51	22%	86%
yes	152	64%	
no	20	8%	14%
strong no	13	6%	

9 *Will the reviewers' remarks be helpful in improving the technical content of your paper or for your future research?*

2012	
late May	informal discussion of PC chair nomination
05 June	formal invitation to serve as PC chair
...	discuss process and PC with GC
3 October	send draft process proposal to SC
28 October	mail draft PC proposal to SC
1 November	agreed PC proposal with SC
2 November	send first round of PC invite requests
15 November	send second round of PC invite requests
3 December	PC complete
11 December	send draft CFP to PC
2013	
16 January	POPL 2014 web page online; CFP advertised
30 June (approx)	submission site online
Friday 5 July	paper registration deadline (16:00 UTC)
Friday 12 July	paper submission deadline (16:00 UTC)
12–17 July	chairs consider externals for PC papers
Wednesday 17 July	first batch of external requests for PC papers
Thursday 18 July	bidding deadline (8:00 UTC)
Saturday 20 July	paper review and guardian assignment done
Thursday 1 August	external suggestions due from guardians
1–4 August	PC chair goes through all suggestions, load-balancing and picking more
Sunday 4 August	main external requests for normal papers
5–7 August	PC chair deals with many declines
Tuesday 6 August	main external requests for PC-chair papers
12 August	reminder to ∼70 externals to accept or decline
...	more requests and reminders until PC meeting
Monday 2 September	external Reviews due (08:00 UTC)
	sent 119 review reminders for non-PC papers
	sent 34 review reminders for PC papers
Monday 9 September	PC Reviews due (08:00 UTC)
Tue. 10 – Fri. 13 Sept.	author response period
10 September	open response period; mail authors and PC
10 September	change HotCRP config to let PC see all reviews and discussions (except conflicts)
12 September	PC chair finished reading paper intros
17 September	mail PC and 249 externals to look at reponses
...	two weeks electronic discussion time
...	chairs prompt discussion on some papers
...	chairs ask for a few more externals
19 September	sent out author survey
22 September	change HotCRP config to de-blind all
22–25 September	PC chair making early reject decisions and prompting discussion (esp. with externals)
Sun. 29/Mon. 30 Sept.	PC meeting
Tuesday 1 October	author notifications sent
Tuesday 1 October	PC workshop
...	PC update reviews and comments
Wednesday 2 October	advertised author notification deadline
Saturday 5 October	final reviews sent to authors
...	fix invited speakers
9 November	Sheridan camera-ready deadline
10 November	draft program to PC and ask about attendance
...	fix session chairs
19 November	program on www and to Sheridan
21 November	Sheridan deadline for front matter
2014	
Wed. 22–Fri. 24 Jan.	POPL 2014

Figure 5. Timetable

strong yes	27	11%	62%
yes	120	51%	
no	73	31%	38%
strong no	16	7%	

10 *Do you think the reviewers (collectively, and taking the author response into account) will understand your submission well enough to come to an informed judgement?*

strong yes	27	11%	60%
yes	115	49%	
not sure	69	29%	29%
no	17	7%	11%
strong no	8	3%	

11 *Comments* [free text box]. 59 (25%) comments

19.2 Survey Conclusions

Double Blind The authors are clearly in favour of a DBR process: 67% *yes* or *strong yes* vs 15% *no* or *strong no* (and 18% *don't care*). That said, many authors will not be in a position to assess the impact of a DBR process on finding expert reviewers who really understand each submission.

One concern with a DBR process is that it may impede normal scientific communication. For POPL 2012, Mike Hicks formulated a *lightweight double-blind* policy, producing a DBR FAQ that made clear that this was undesirable, but asked:

> *that you not attempt to deliberately subvert the double-blind reviewing process by announcing the names of the authors of your paper to the potential reviewers of your paper. It is difficult to define exactly what counts as "subversion" here, but some blatant examples include: sending individual e-mail to members of the PC or ERC about your work (unless they are conflicted out anyway), or posting mail to a major mailing list (e.g. TYPES) announcing your paper.*

going on to say

> *On the other hand, it is perfectly fine, for example, to visit other institutions and give talks about your work, to present your submitted work during job interviews, to present your work at professional meetings (e.g. Dagstuhl), or to post your work on your web page.*

POPL 2014 adopted a still lighter-weight policy, with an FAQ based on that for 2012 but making explicit that discussion of a submission on a mailing list was legitimate.

Questions 5 and 6 show that these messages are not getting across to all authors: a significant fraction (25% of survey respondants) refrained from putting the paper on a web page because of DBR, and smaller numbers (9% and 5%) refrained from giving talks or talking informally about their work. This is unfortunate and worth addressing by future chairs. It is clear that stating a policy in such a FAQ does not necessarily get it across to all authors.

Number of Reviews The *number* of reviews to seek per paper is one of the variables that a PC chair can easily control. There is a balance between seeking more reviews and imposing an unjustifiable review load on the community — recall that each extra review per paper amounts to 200+ more reviews. For POPL 2014, out of 220 submissions, the majority of papers (146, 66%) received four reviews; 53 (24%) received five, 15 (7%) received three, and 2 (1%) received six (the remaining 4 were withdrawn before the author response period). Question 7 shows that the survey responders overwhelmingly consider that their paper had a *reasonable number of reviews* (209, 89%) with just a few saying *unnecessarily many* (7, 3%) or *too few* (20, 8%).

Review Quality Questions 8, 9, and 10 ask whether authors thought the reviews would be helpful in improving presentation (8) or technical content and future research (9), and whether they thought the reviewers would collectively understood their submission (taking the author response into account) well enough to come to an informed judgement (10).

A large majority (86%) of respondants answered *yes* or *strong yes* to the first. This is good to see but not surprising: POPL reviews very often contain helpful suggestions for presentation improvement.

A smaller but reasonable fraction (62%) thought that the reviewers' remarks would be *"helpful in improving the technical content of their paper or for their future research"*.

Finally, 60% replied *yes* or *strong yes* to Question 10, with 29% *"not sure"*, and 11% *no* or *strong no*. Without comparable data from other conference instances, it is hard to know whether one should consider this good or bad. The fact that 40% were not confident that the reviewers will understand their work is less than ideal, at least.

The 11% that answered *no* or *strong no* (referring to 22 papers) might, a priori, represent either cases where we have made a serious error in the review process or dissatisfied authors of submissions that in fact were not up to the required standard. Looking at the reviews and discussion of these papers, there are some clear cases of the latter, a few cases where an extra review came in late (and was forwarded to authors for quick comments), and several cases where there was extensive discussion (and hence where different reviewers or PC members might have come to a different conclusion).

One might expect a strong correlation between whether a paper is ultimately accepted and whether the authors consider that the reviewers to have understood it well. To examine this we consider, for each possible outcome of Q. 10, the set of distinct papers for which such a response was given, and calculate the acceptance rate for those. For comparison, recall the overall acceptance rate was 23%.

Q10 response	distinct papers	accepted	acceptance rate
strong yes	21	11	52%
yes	90	33	37%
not sure	57	16	28%
no	14	1	7%
strong no	8	1	13%

The textual comments are included in an appendix, but there do not seem to be any clear conclusions to be drawn from them.

A. Survey comments

This appendix records the textual comments from the author survey. This should be read in conjunction with the numerical scores: only 59 out of 236 responses included a textual comment, and those who gave low scores for the review quality question are, unsurprisingly, disproportionately represented among those 59.

For anonymity, parts of comments that would have identified the authors have been redacted, indicated *[. . .]*. Comments have been subdivided into those on the double-blind process (positive, negative, and neutral), reviewing (positive, negative, and neutral), the number of reviews, the author response, and others.

Double Blind (positive)

- *I am favor of continuing the lightweight double-blind process coupled with the guardian process. I believe the cost of the lightweight double-blind process is minimal, especially with a guardian who can solicit expert reviews early.*

Double Blind (negative)

- *Disadvantages of blind process: pointers to research report and web page (experiments, tool) should be anonymized as wel, not convinced of its effectiveness anyway.*
- *The lightweight double-blind process did not seem useful; as an external reviewer, I saw one PC member submitted a "dummy" review. By doing that, he/she can know the identity of authors before coming to a strong opinion about the paper. Generally, in my opinion, a double blind process (whether heavy weight or lightweight) often bring more harms than benefits. It does not bring much benefit because reviewers can often guess the identity of authors from their past research record.*
- *absolutely against dobule-blind process as the people that may be more at risk of giving biased reviews are the ones most likely to be able to de-anonymise the submission anyway. double blinding gives them an additional advantage in terms of information asymmetry.*
- *Our paper was about a software artifact (a program and proofs regarding the program). Not having the program available to the reviewers for the initial review seemed to be problematic (a reviewer complained that we referred too often to the appendix, i.e., the program, which is what the paper is about). Double blind refereeing, even in this light form, discourages the publication of such results.*
- *I found the anonymity constraint not to be helpful. Without it I would have added pointers to extra material that might have helped the reviewers and possibly covered some of their remarks.*
- *The double-blind process discouraged me from advertising the artifact that came out of this work.*
- *Double blind does not make much sense since people can easily read out who wrote the paper because some authors do have their style of writing the papers;*
- *Double-blind review is hard to enforce; it is typically the case that papers/ideas are presented to the community before submissions. Also, it is usually not hard to guess who the authors are from the content of their papers. More than a double-blind process, it would be more useful to have thorough reviews. With the notifications, it would be helpful to receive a brief account that justify the decision of the PC (and how the response of authors have been considered).*

Double Blind (neutral)

- *The lightweight double-blind process was not a hindrance.*
- *I'm willing to go along with DBR, but I don't think it adds anything - as author or as reviewer. But perhaps that means that I have precisely the unconscious biases it is designed to fix...*
- *The public talk I gave about the work was prior to the submission deadline. Since submitting, I essentially only talked about the work with people who're conflicted with the paper, graduate students of such people, and representatives of funding agencies.*
- *I doubt that the lightweight double blind version of the peer review process had an impact positively or negatively on the result. The bottom line is that there is no perfect peer review process, but I applaud the attempt to make the process as fair and as effective as possible.*
- *I was left with some confusion about what did or did not constitute "lobbying", but feel like it wasn't too hard to behave ethically while still discussing the technical work.*
- *Talks about this work have been previous to POPL submission*

- *Double blind review only purpose is to kill bad papers by famous authors. If the reviewer is really an expert of the area s/he will know your work, or at least your group, no matter if you put the work on the web or not. However, people slightly outside the area might not know that you are one of the good and the great, and thus reject your bad paper.*

Reviewing (positive)

- *I greatly enjoyed reading the scores with the new rating scheme. I believe it is sensible to convey what the PC chair expects a score to represent rather than let the reviewer use some internal scheme.*
- *My first time submitting to POPL. I found the process simple and smooth. The reviews of my work were very detailed and thorough and much appreciated. Thank you.*
- *The reviewer comments were very useful, although the first reviewer seem to have missed the point of the paper. May be perhaps we could've explained the purpose of the research more clearly. Other than that, we're really happy about the reviews. We didn't try to defend the paper because we had to rush the paper for this deadline, and we discovered a few errors on the submission later on (some were not caught by the reviewers). POPL is awesome!*
- *Exceptionally well-informed, fair and balanced reviews. It is not always that good.*
- *I believe that even negative reviews are given in good faith - we all try to be professional. They are quite helpful, pointing to some issues that I have not considered or simply identifying things that I need to express and communicate better, be it for a particular community or in general.*
- *I was very impressed by the attention to detail of four out of our five reviews.*
- *The reviewing process seems good overall this year, though the proof is in the pudding :-)*
- *The reviews – both positive and negative – were extremely thorough, thoughtful, and constructive. This tradition of super-high reviewing standards is one of POPL's greatest strengths!*
- *This is my first submission to POPL and, so far, a very nice experience. I received 5 reviews [. . .], which is a LOT! Clearly the two experts gave the most useful reviews/comments. However, the others were also very useful to understand the point of view of not-experts. POPL: very good!*
- *The reviews are critical for the authors to (1) understand how they have to present their work in a clear way and (2) to learn in which ways their work can (or has to be) improved. I feel the reviews for our paper accomplished these tasks.*

Reviewing (negative)

- *Most comments were of the generic variety .. what is the usefulness, more examples please, too technical etc. While of course authors need to address these issues, sometimes reviewers hide behind these comments and their genericity meant I didnt find them particularly useful.*
- *It seemed like some reviewers didn't know what grade corresponded to what merit. It would be great to have some introductory message to all reviewers to make sure all of them have same consesus on merit/grade.*
- *I wish there would be greater education for reviewers about natural biases and failings in reviewing different kinds of work. Especially in conceptual work that is different from existing trends, I find stronger negative reactions and unjustified skepticism. There is by now documented evidence for various psychological biases people have when making judgements. It would be good to keep this in mind.*
- *I have answered "not sure" to the last question because from the reviews that I received it seems like two reviewers did not read certain parts of the paper, or were biased by their own works. Yet I don't know if changing the review process could help in avoiding situations like this.*
- *Many of the reviewers were not expert, and an expert one completely missed a main point of the paper*
- *We had only 3 reviews. Since the standard is 4 reviews (3 from the PC and 1 external), and reviewers had all different opinions, I wonder what the additional review will/would be. I would have prefered to get the review later (in the response to reviews window) than not at all. Since at the PC meeting the initial scores may count for at least as much as the PC discussions, one missing review when reviews scores are spread can have a real impact on the outcome (one way or the other:-).*
- *Several reviews requested extra information that was already in the article, including topics that were discussed for entire paragraphs or sections. This leads us to wonder whether the reviewers really read the article in detail, or just skimmed through it.*
- *This was a highly technical paper, and I'm not convinced that any of the reviews "got" it. I do have some sympathies. I wonder whether I should accept that some kinds of work, although entirely on topic, may be better suited to journal than conference exposition? Incidentally, it's not helpful that HotCRP has hidden the reviews, just when this survey comes along to ask about them!*
- *There was serious misunderstanding of a key element of the paper by one reviewer that hopefully the rebuttal has fixed, we'll see. In general, reviewers seemed to like the paper but were not sure how to place its contribution, so their understanding of the paper might be good, but not of the context.*
- *Our paper suffers from lacking a true expert reviewer at this time. The one who claims he is an expert is not, really, and seems not gloss over key technical contributions of the paper.*
- *I thought the reviews of the paper were remarkably low quality for POPL, except for the last one. In the past I've had great experiences, where even papers that are clearly not getting in received substantial and useful feedback, which is one of the reasons I think POPL is a great conference and community. This year, and on this paper in particular, I thought that three or the four reviews were cursory, dismissive, not thoughtful, and non-constructive. Let me emphasize that this is about the review text, and not the scores assigned to the paper—I'm quite used to having papers receive low scores, so that doesn't bother me :). Overall, I feel rather disappointed by the process.*
- *I submitted this paper at POPL2013 (last year). I managed to get [. . .] and I know it was on the fence at the PC. I improved it quite a lot for this year. I got [. . .] (Unfortunately the reviews are not available anymore, so I have to remember them). But I did not know what to do with those reviews. There was no real complaint on the content or the presentation, only a few remarks that could have been clarified easily since the answers were already in the paper. There was no advice or direction that would help me improve it. This may be due to two aspects: - either the paper is not interesting at all even if well formed, with no errors etc. In this case how come that I got [. . .] last year? - either those reviewers did not understand the* nature *of the contribution. This case is happening in other conferences: when*

confronted to alien contributions, reviewers tend to give low ratings. But I do not think that a rebuttal would have changed their mind, hence [. . .]. Still, it is too bad that: - I feel like I have to explain not only the content of the contribution and its relevance, but also the nature of the contribution: it seems to me that it's basic epistemology. - reviewers tend to completely forget the positive aspects of the paper. Again, there were no complaints in the reviews. In my opinion, they should have stated the positive aspects (understandable?, well written?, well argued/supported?) , and rate it accordingly i.e. with a more balanced view.

- *A single, biased review in the initial process made it virtually impossible for my paper to receive the attention it deserved. I accept this as a consequence of "the process", but [...]*
- *The reviewers claimed expertise in the area but it is clear that the majority of them were either ignorant of the field, or amateurs, or both. POPL has a standard to maintain, and the PC should be chosen to abide to the highest professional standard. The nature of the reviews indicate that the reviewers might not either have time, or were not mature enough to be part of POPL PC.*
- *The reviewers seem to think that [. . .]. How they came to that conclusion, given our citations of previous work and the technical content of the paper, is beyond me. Had my name been visible, the reviewer would likely have doubted his/her erroneous conclusion. The quality of our "expert" reviews was abysmal to the point of incompetence.*
- *One reviewer stated that [. . .]. I believe the reviewer was simply biased against the work, and found a reason to downgrade our score.*
- *While the number of reviews were good, the quality of most (but not all) of the reviews our paper received were quite poor and will not be terribly useful for improving either the work or the presentation.*
- *It seems that none of the reviewers had the time to read through our formalism. The general sentiment expressed by all four reviewers is that they could not follow the theory, but [. . .]. These and similar observations lead us to believe that the reviewers were rushed in reading our paper and explain our feedback above.*
- *I am not sure what the point is in providing 4 reviews when 3 of them are extremely superficial. From the questions I doubt 3 of the referees actually read the paper at all, never mind carefully. (However, the fourth referee was quite thorough and made a couple of cogent comments.) The reviews seem to be no longer accessible on HOTCRP (why?) but if I remember correctly I didn't have one single 'expert' review. The peer review system for very broad yet very selective conferences such as POPL has become a ridiculous lottery. Sometimes I drew the winning referee (i.e. one that has some interest in my line of work) and sometimes, such as this time around, I didn't. The reports are mostly statements of (dis)interest in technical details rather than technical comments. I think small improvements such as double-blind reviews are a step in the right direction, but the system is so flawed that nothing short of a radical rethink will fix the problems. I hope someone will have the courage to do it because, tragically, for some people, especially young ones, a lot is riding on the outcome of this lottery.*
- *Unfortunately, the reviewers appear not to have understood the contributions. I realise that (a) it is difficult, in technical and diverse conferences like POPL, to find appropriate reviewers; (b) it is sometimes the authors' fault if reviewers don't understand. However, in this case I believe the reviewers did a bad job.*
- *I'm not so sure how informed the judgement can with two reviewers giving low marks accompanied by very flimsy review reports.*
- *Of four reviews of our submitted paper, three were short and seemed to indicate that only little time was spent on reviewing the paper. Consequently, the reviewer comments usually amounted to "Looks OK, but doesn't interest me" and were thus not helpful in improving the paper. This may indicate that using more external reviewers (and thus spreading the reviewing load over more people) could improve this situation.*
- *I thought the reviewers on the paper did a reasonable job—they all clearly invested time into reading and understanding the paper, and trying to give us constructive feedback. The main issue was lack of expertise. Three of four reviewers were definitely not experts, and the most expert reviewer actually came from a rather different segment of the community and seemed unfamiliar with work closest to our paper. So, while on the one hand we should definitely change our paper to address that reviewer's community better (and hence we are grateful for that review), we would really have liked to get more expert feedback.*

Reviewing (neutral)

- *The fact that the reviewers might not understand our submission well enough might be due to our presentation of the paper: we propose a new framework for a [. . .] problem, but we have not yet found the "good" way to present it. We believe that the reports of the reviewers will help us for improving our presentation in the future.*
- *We hope less confident/informed reviewers discuss with other reviewers and adjust their reviews sincerely.*
- *Two of the four reviewers clearly had difficulty assessing the paper; the other two were experts in the area. In any case, our paper is not in a large research, so I think the double blind approach was a bit 'transparent' in our case.*
- *We were disappointed by the reviewers responses to the paper. True, the paper is very technical; but POPL is a technical conference. We were accused of being "niche"; but we feel that POPL is our niche - we can't think of a better place to submit this work. Having said that, one reviewer and one subreviewer clearly understood the content, and made many helpful small suggestions to improve the presentation - those reviews were very useful.*

Reviewing (number of reviews)

- *Number of reviewers:3 is OK if there is really a discussion in case of mismatch (something I do not know). But as soon as one review is superficial it might be annoying.*
- *More expert reviews are always appreciated to improve the quality of POPL reviews.*
- *So 4 reviews is a right number: you get experts (that will know you) and a bit borderlines (who will not know you). Either way there will always be somebody who will not read carefully your paper, no matter if blind or not, but this is part of life.*

Author response

- *I have doubts about author responses being cost effective. They delay a definitive decision for 2 weeks and seldom result in any decision change. It would help if the process was interactive, if during these 2 weeks authors could see and react to reviewers' comments, after sending their response.*

- *Some reviews did not clearly identify the 'questions to authors' part. Other reviews included parts of the review in the questions part. This was hard to understand which questions are really important for the reviewers. Notice that we couldn't answer to all questions (from review + questions parts) or give short and, sometimes, incomplete answers due to the limit of 500 words (avg. 100 words/review). We believe that this can affect the understanding of the paper. Worse, we also believe that the lack of possibility to answer to a question, judged less important for us but maybe considered as very important for the reviewers, is a good reason to reject the paper. We would have appreciated to have the order of importance of the questions already established by the reviewers. We think that if there is no limit to the number of questions for the reviewers, there should be no limit for the number of words in the 'answers' part.*
- *I think it is appropriate that the main author response should be word-limited. However, I would favour (not just for POPL) a system where the authors submit a 500-ish word response that will definitely be read by the PC, and optionally a longer response that may optionally be looked at. I think the second response would be a nice way to allow reviewers to meticulously respond to small, but nevertheless important (if not make-or-break) comments, (a) to show that they have taken them in and care, and (b) because in many cases the PC member or reviewer who made the comment will likely be genuinely interested in a response, even if the point in question is rather specific. I saw in the response instructions that there was the option of submitting a longer list of responses at the end of the rebuttal, which is basically what I'm suggesting, but then the resonse system said "try to keep to within 500 words"; I think this could have been clearer.*

Other

- *Seeing the remark "Never submit passwords through Google Forms." below, I am reminded that the author response notification from HotCRP contained the passwords in clear text! Clearly, this should not be the case.*
- *We forgot to anonymize our submission. A reminder on the submisssion page (or automactic detection extracting the authors name from the PDF) could have avoided this.*
- *If the reviews happen to be awful in the small subset of the community, I think it makes sense to not ask authors to write response at all and the best would be allowing them to resubmit elsewhere rather than waiting for the rejection email and having wasted some other paper submission deadlines?*
- *I would like to see consistency in the reviewing process from year to year. Programs should try to stick with the model that has been used in previous years. In general, I don't think that relying on authors to suggest reviewers is a good idea.*
- *In some of the reviews it was difficult to understand the comment "OK but I will not champion it" with respect to the positive comments. I guess there is a limited number of papers that each PC member is able to champion. With respect to helpful comments, I think it is always better n+1 than n readings. There is no unique way to read a paper so feedback from different members of the PC could be really interesting in some cases.*
- *It is essential that authors be permitted to submit supplementary material along with their papers that will be viewable by the PC *before* entering reviews. In one of our reviews, a reviewer explicitly noted that they had revised their assessment of the difficulty of what we had done based on looking at our technical appendix only after they submitted their initial review. Why keep the reviewer in the dark in the first place? The issue is even more important for authors who submit mechanized proofs along with their paper as a way of avoiding the need to spell out boring technical details in the paper itself. The PC should have the mechanized proofs on hand *when reviewing*, in order to properly assess the work.*
- *In general, I think POPL is becoming a very insular community. They are not very welcoming of ideas that are at the intersection of PL and some other area. This has to change. Otherwise the community will shrink.*
- *The submission system hotcrp itself worked great. Especially when submitting revisions up to the last minute. I would love to have another round of reviews or rebuttal response, as the rebuttal was challenging to write, as I'm scared that even a small syntax misunderstanding could wreck our chances!*
- *The results of the survey could be biased or the number of responses suppressed because of having to submit the answers before the final decision on the paper and having the responses sent to the PC chair (even given the note at the top of the survey). If you had a 3rd party collect the answers, you might get better responses (perhaps someone on the SIGPLAN EC unconnected to the POPL review process).*
- *I worry a bit about the handling of PC papers. Without the PC or an ERC being involved, I wonder if the external reviewers will be as well calibrated, or willing to give out top scores – e.g., a reviewer with a stack of 25 papers will likely to be inclined to give out a few As; a reviewer with just one, will be much less likely to do so.*