The ESA 2018 Track B Experiment

An in-depth analysis of two parallel PCs reviewing the complete set of submissions independently

Hannah Bast

PC Chair of ESA'18 Track B @ Helsinki August 22 – 24

last updated 01.02.2019

UNI FREIBURG

ESA 2018 Track B Experiment

UNI FREIBURG

- Two independent PCs of 12 members each
  - Each PC reviewed the complete set of submissions
  - Each PC followed the same reviewing "algorithm"
  - The PCs had the same composition with respect to:

Gender	8 male, 4 female
Seniority	2 x PhD 1-5y ago, 4 x 6-10y ago, 6 x senior
Continent	8 x Europe, 4 x Americas, 0 x Asia (we tried)
Topic	1 x parallel, 2 x string, 2 x comp. geometry, 2 x operations research, 5 x algo in general

  - Goal of the experiment: **after** the PC work is done, investigate commonalities and differences of the result

PC Members

Paolo Ferragina	U Pisa	Martin Aumüller	ITU Copenhagen
Stefan Funke	U Stuttgart	Christina Büsing	RWTH Aachen
Michael Goodrich	UC Irvine	Pierluigi Crescenzi	U Firenze
Sungjin Im	UC Merced	Veronica Gil-Costa	UNSL San Luis
Silvio Lattanzi	Google Zürich	Inge Li Gørtz	DTU Lyngby
Tamara Mchedlidze	KIT Karlsruhe	Michael Kerber	TU Graz
Richard Peng	Georgia Tech	Jon Lee	U Michigan
Simon Puglisi	U Helsinki	Matthias Müller-H	U Halle
Melanie Schmidt	U Bonn	Petra Mutzel	TU Dortmund
Anita Schöbel	U Göttingen	Gonzalo Navarro	U Chile
Sebastian Stiller	TU Braunschweig	C. Schwiegelshohn	TU Dortmund
Carola Wenk	Tulane University	Darren Strash	Hamilton College

54 subreviews used

59 subreviews used

Submissions

UNI FREIBURG

- Overview
  - 51 valid submissions … 5 invalid format / withdrawn
  - 12-13 submissions per PC member
  - 313 reviews overall … 95 x 3 reviews, 7 x 4 reviews
  - Each PC accepted exactly 11 papers, both together 15 papers
  - Acceptance rate **21.6%** per PC and **29.4%** overall
  - Top countries wrt #submissions and #accepted:

Country	submitted	accepted	acc. rate
US of A	15.5	5.2	34%
Germany	12.6	5.1	40%
Austria	2.8	1.0	36%
France	2.7	0.7	25%

Scores

- Scores

+2	**accept**	good fit and no major weaknesses
+1	**weak accept**	significant weaknesses, but still acceptable
0	**borderline**	hovering between +1 and –1
–1	**weak reject**	significant weaknesses, lean to reject
–2	**reject**	bad fit or major weaknesses

- Additional semantics (last two because of the experiment)
  - Final score must not be 0
  - A submission needs at least one +2 to be accepted
  - Scores should be synced with reviews (during discussion)
  - Score set should reflect the final status of the discussion

Review Process

UNI FREIBURG

- Phase 1: Initial Reviews
  - PC members only see their own reviews in this phase
- Phase 2: Discussions (mostly) per paper
  - Read each other's reviews, discuss, keep scores and score sets in sync with reviews and discussion
  - Accept (in rounds): all scores +2 with conf ≥ 3 — **PC1 / PC2**: 5 / 4
  - Reject (in rounds): no +2 score until the end — 34 / 39
- Phase 3: Discussions between papers + Voting
  - Succinct summary for each remaining paper
  - Re-discuss + adapt scores — 0 / 2 more decisions
  - Vote on remaining papers — 12 / 6 votes

Results of the Experiment 1/6

- Questions
  - Overlap in accepted papers?
  - Percentage of clear accepts?
  - Percentage of clear rejects?
  - Papers where the two PCs strongly disagree?
  - Effectiveness of the discussion phase?
  - Effectiveness of the voting phase?
  - Most decisive aspects for reviewer decisions?
  - Reviewer agreement with respect to these aspects?
  - Consequences for the reviewing process?

Results of the Experiment 2/6

UNI FREIBURG

- Overlap in accepted papers
  - Percentages for 10 / **11** / 12 papers accepted per PC

After Phase 1:	50% / **55%** / 67%	Note: one paper out and another paper in can already change the percentage by 6-7%
After Phase 2:	70% / **73%** / 75%	
After Phase 3:	70% / **64%** / 58%	

  - Percentages for some probabilistic models (details for 11 papers acc.)

Fully random:	20% / **22%** / 24%	(100% x 0.22)
Random 20:	50% / **55%** / 60%	(40% x 0.55) (60% x 0.0)
Random graded:	**63%**	(10% x 0.8) (20% x 0.6) (20% x 0.1) (50% x 0.0)
Random graded:	**72%**	(18% x 0.8) (12% x 0.6) (4% x 0.1) (66% x 0.0)

Results of the Experiment 3/6

- Percentage of clear accepts
  - 10 papers with at least two +2 in one PC (with confid. ≥ 3)

+2 +2 +2	+2 +2 +2
+2 +2 +2	+2 +2 +2
+2 +2 +2	+2 +2 +1
+2 +2 +2	+2 +1 +1
+2 +2 +2	+1 +1 +1
+1 +2 +1	+2 +2 +2
+2 +1 +2	+2 +2 +2
+1 +1 +2	+2 +2 +2
−1 −1 −1	+2 +2 +2
+2 +1 +2	+2 +1 +2
+2 +1 +2	−1 −1 −1
+2 +2 +1	+1 +1 +1

out of 9 papers that were "clear accepts" in one PC

4 were **rejected** by the other PC

and only 2 were also "clear accepts" in the other PC

Results of the Experiment 4/6

- Confusion matrix between the two PCs
  - Let's categorize papers by their scores sets as follows:

+2	+2 +2 +2
+1	at least one +2
0	at least one +1, but no +2
–1	no +1 or +2, but no –2 either
–2	no +1 or +2, at least one –2

+2 or –2 only considered as such if reviewer confidence ≥ 3

  - Recall that:

All **+2** papers were accepted

we will refer to these as the "gray zone" papers

The **+1** papers were decided by more discussion, then voting

All **0** and **–1** and **–2** papers were rejected (after discussion)

Results of the Experiment 5/6

- Confusion matrix between the two PCs:

51	24	11	16
21	14	6	1
12	6	3	3
18	4	2	12

After Phase 1
(all reviews in, before discussion)

51	12	14	25
17	11	3	3
9	0	5	4
25	1	6	18

After Phase 2
(discussions per paper, scores updated)

51	11	15	25
17	10	4	3
9	0	5	4
25	1	6	18

After Phase 3
(discussion of papers in "gray zone", voting)

- Almost no strong disagreement:

	Ph3	Ph2	Ph1
– #Papers with **−2** in one PC and **>0** in the other:	**0**	**0**	**1**
– #Papers with **+2** in one PC and **<0** in the other:	**1**	**1**	**0**

Results of the Experiment 6/6

UNI FREIBURG

- Most decisive aspects for reviewer decisions (306 reviews)

W	quality of write-up	(202 reviews)	+ o – = 78 / 38 / **86**
R	quality of results	(263 reviews)	+ o – = **128** / 66 / 69
E	quality of evaluation	(181 reviews)	+ o – = 53 / 28 / **100**
T	technical depth	(63 reviews)	+ o – = 23 / 3 / **37**
C	correctness	(18 reviews)	+ o – = 6 / 0 / **12**

- Disagreement per paper (where ≥ 2 reviews mention the aspect)

W+ and W–	19 / 74 = 26%	if + o – were random: 44%
R+ and R–	21 / 94 = 22%	if + o – were random: 42%
E+ and E–	13 / 62 = 21%	if + o – were random: 42%
T+ and T–	1 / 16 = 6%	if + o – were random: 44%

Personal observations as PC Chair

UNI FREIBURG

- For ESA, the whole reviewing process is purely electronic (no physical meeting of the PC at any point)
- This works well for Phase 1 (it's always a hassle to get all reviews in time, but it can be done and usually works)
- For the various **discussions** and votes, this is a major problem:

  If a PC member does not reply (the usual case), it is impossible to know whether that is because they are sticking to their original review/score or because they forgot to answer

  For this PC, because of the experiment, I took extra-ordinary care to always get feedback from (almost) everybody

  So the agreement between the two PCs is probably a bit better than usual because of that

Conclusion from the Experiment 1/2

UNI FREIBURG

- Quick answers to the questions
  - Overlap in accepted papers? 50-75% not the best figure to remember
  - Percentage of clear accepts? Very few, if any
  - Percentage of clear rejects? About 40%
  - Papers where the two PCs strongly disagreed? 1 out of 51
  - Effectiveness of the per-paper discussions? Reasonable
  - Effectiveness of the "gray zone" discussions? Very little
  - Most decisive aspects for rejects? Write-up + Evaluation
  - Reviewer agreement with respect to these aspects? Moderate
  - Consequences for the reviewing process? See next but one slide

Conclusion from the Experiment 2/2

- Executive summary

  No clear distinction between "clear" and "possible" accepts, and the corresponding discussions are not very effective

  Note that the decision often **feels** just and fair to the PC, but that doesn't mean the decision is (much) better than random

  Almost **no** confusion of three score levels or more

  That is, of "strong accept" (+2) and "likely reject" (< 0)
  or of "possible accept" (> 0) and "strong reject" (−2)

  Moderate agreement concerning the individual aspects of a paper (quality of write-up, quality of results, quality of evaluation)

  Apparently good enough for a precision of two score levels

Consequences for the Reviewing Process

UNI FREIBURG

- Option A: Leave it as it is
  - It's certainly not bad and does an excellent job in identifying the "clear rejects" and giving the authors detailed feedback
  - Maybe drop or shorten the "gray zone" discussions
- Option B: Do not try to distinguish +1 and +2 papers
  - It looks like they are very hard to distinguish anyway
  - This would mean **doubling** the acceptance rate 25% → 50%
- Option C: Accept each paper with probability ~ score
  - Telling from this experiment and others of its kind, the process would be just as fair or even more fair (less biases)
  - The PC can focus on the effective part of the work

Happy discussions + thank you for your attention !