Standard Operating Procedure (SOP)

Procedure for Selecting 45 Extended Abstracts + 15 Waiting List
(Weighted Percentile-Based Method)

YA AESOP 2026 PhD Conference - LOC

Version 2.0 — December 1, 2025

Contents

1 Overview and Roles

This SOP defines a rigorous, bias-resistant selection workflow for 147 extended abstracts using six PhD reviewers and three professors as senior validators.
The updated method uses:

1. calibrated double-blind peer review,
2. **weighted percentile transformation** to correct reviewer harshness/leniency,
3. domain-aware banding,
4. proportional domain quotas,
5. a reproducible weighted lottery for borderline cases.

Domain	Submissions	Share of total (%)
Qualitative	74	50%
Quantitative	50	34%
Theory	23	16%
Total	**147**	**100%**

Table 1: Distribution of extended abstract submissions by domain.

Roles

— **PhD Reviewers (LOC members):** conduct two blind reviews per abstract.
— **Professors:** adjudicate disagreements; apply limited wildcards.
— **Communications Lead (Kolouchova):** manages RSVP processes.
— **Record Keeper (Hadravova):** maintains audit trail and method documentation.

2 Evaluation Rubric (5-point scale)

Each abstract is scored 1–5 on the following four criteria:

1. **Fit to Track Theme** (weight 0.8)
2. **Originality / Contribution** (weight 1.0)
3. **Research Design / Argument Quality** (weight 1.0)
4. **Clarity of Writing** (weight 0.7)

Weights are applied after percentile conversion (Section 6).

3 Domain-Aware Assignment

Each abstract receives:

— one *in-domain* review, and
— one *cross-domain* or second in-domain review depending on workload distribution.

Assignments for domain–cross reviews use a fixed random seed to maintain reproducibility.

Domain	Domain–domain	Domain–cross	Total	Share domain–domain
Qualitative	55	19	74	74%
Quantitative	37	13	50	74%
Theory	23	0	23	100%

Table 2: Review assignment structure by domain.

4 Data Structure (Scoring Sheet)

Each review contains:

— `abstract_uid`, `domain`, `reviewer_id`
— four criterion scores (1–5)
— derived weighted percentile score (Section 6)

5 Calibration and Conflicts

— All abstracts anonymised.
— Reviewers declare conflicts of interest.
— A calibration round of 10 shared abstracts aligns scoring behaviour.

6 Per-Reviewer Percentile Transformation

Let reviewer r assign raw criterion scores $s_{i,c,r} \in \{1, \ldots, 5\}$.
For each reviewer and each criterion, compute a within-reviewer percentile:

$$p_{i,c,r} = \text{PercentileRank}_r(s_{i,c,r}), \quad p_{i,c,r} \in [0, 1].$$

Apply criterion weights:

$$\text{WPS}_{i,r} = 0.8\, p_{i,\text{fit},r} + 1.0\, p_{i,\text{orig},r} + 1.0\, p_{i,\text{design},r} + 0.7\, p_{i,\text{clarity},r}.$$

This corrects for harshness/leniency and yields a scale-free, comparable score across reviewers.

7 Aggregation to Abstract Level

With two reviews per abstract, the abstract's final score is:

$$\text{score}_a = \frac{\text{WPS}_1 + \text{WPS}_2}{2}.$$

Define disagreement as:

$$|\Delta|_a = |\text{WPS}_1 - \text{WPS}_2|.$$

Abstracts with:

$$|\Delta|_a > 1.50$$

are flagged for adjudication.
With three reviews (after adjudication):

$$\text{score}_a^{(3,\text{trim})} = \text{mean}(WPS_{(1)}, WPS_{(2)}).$$

8 Domain-Specific Banding

Rank abstracts within each domain by their final percentile score.
Define domain bands:

— **Top band (T):** domain percentile P65–P100
— **Middle band (M):** P30–P65
— **Bottom band (B):** P0–P30

Banding reduces sensitivity to minor numerical differences near selection thresholds.

9 Proportional Domain Quotas for 45 Selections

Compute quotas using the largest remainder method:

$$q_t^* = 45 \times \left(\frac{n_t}{147}\right).$$

Assign:

1. $\lfloor q_t^* \rfloor$ guaranteed seats per domain,
2. remaining seats allocated by descending fractional remainder.

10 Selecting the 45 Abstracts

1. **Automatic selection:** Fill each domain's quota starting from the Top band.
2. **Cut pool:** If quotas remain unfilled or ties occur, create a domain-specific cut pool including:
   — lower Top-band abstracts,
   — upper Middle-band abstracts.
3. **Weighted lottery:** For borderline items, apply a reproducible weighted lottery with weights:
$$w_a = \max\left(0.01,\ \text{score}_a - \min(\text{score}_{\text{pool}}) + 0.05\right).$$

11 Waiting List (15 Abstracts)

After selecting 45 abstracts:

— Compute domain-proportional quotas for 15 waiting-list slots,
— Apply the same Top/Middle band logic as above,
— Use lottery only when ties or clusters occur.

The waiting list is ranked and used sequentially during the RSVP stage.

12 RSVP Process

Selected authors (45) receive notification with a seven-day RSVP deadline. Non-respondents or decliners are replaced by waiting-list candidates in domain-proportional order.

13 Documentation and Equity Checks

— **Public method note:** percentile method, weights, quotas, lottery seed.
— **Internal audit pack:** anonymised raw scores, percentiles, final scores, bands, quota calculations, adjudication log.

Appendix A — 5-Point Scale Guide by Rubric Criterion

Scores apply per criterion. Examples are phrased to work across qualitative, quantitative, and theoretical submissions.

A. Fit to Track Theme

Score	Meaning	When to assign (examples)
1	Poor	Off-topic or question/argument unclear; no evident link to the track focus.
2	Weak	Loose or generic link; aims are vague or drift from the track's stated scope.
3	Adequate	Clearly relevant; research question/argument is identifiable but not sharply framed.
4	Strong	Tight alignment; well-specified question/argument directly addresses the track brief.
5	Outstanding	Central to the track's agenda; clearly advances the conversation the track was designed for.

B. Originality / Contribution

Score	Meaning	When to assign (examples)
1	Poor	No novelty; purely descriptive or well-known result re-stated without insight.
2	Weak	Minor twist on a familiar idea; marginal extension without clear added value.
3	Adequate	Modest originality (new case, small methodological tweak, or incremental theoretical point).
4	Strong	Clear contribution: non-trivial dataset/method, novel angle, or useful theoretical synthesis.
5	Outstanding	Distinctive, potentially field-shaping idea/dataset/model/ theory with clear implications.

C. Research Design / Argument Quality

Score	Meaning	When to assign (examples)
1	Poor	Method/argument missing or untenable (e.g., claims without support; logical gaps).
2	Weak	Partially appropriate but with major flaws (sampling threats; validity concerns; thin theory).
3	Adequate	Generally appropriate; limitations noted but core logic/method holds.
4	Strong	Well-matched design or compelling argument; awareness of assumptions/limits.
5	Outstanding	Rigorous and transparent design/logic; exemplary credibility (reflexivity/validity/consistency).

D. Clarity of Writing

Score	Meaning	When to assign (examples)
1	Poor	Hard to follow; disorganised; grammar issues obscure meaning.
2	Weak	Readable in parts but uneven; key terms or structure unclear; missing signposting.
3	Adequate	Understandable overall; some roughness or missing connectors but argument can be followed.
4	Strong	Clear, coherent, well-structured 1000–1500 words; appropriate terminology and flow.
5	Outstanding	Concise, elegant, and engaging; excellent structure and wording; zero ambiguity.