



FORUM ACUSTICUM EURONOISE 2025

OPTIMAL PAIRWISE COMPARISON PROCEDURES FOR SUBJECTIVE EVALUATION

Jack Webb^{1*} Lorenzo Picinali¹

¹ Dyson School of Design Engineering, Imperial College London, United Kingdom

ABSTRACT

Audio signal processing algorithms are frequently assessed through subjective listening tests in which participants directly score degraded signals on a unidimensional numerical scale. However, this approach is susceptible to inconsistencies in scale calibration between assessors. Pairwise comparisons between degraded signals offer a more intuitive alternative, eliciting the relative scores of candidate signals with lower measurement error and reduced participant fatigue. Yet, due to the quadratic growth of the number of necessary comparisons, a complete set of pairwise comparisons becomes unfeasible for large datasets. This paper compares pairwise comparison procedures to identify the most efficient methods for approximating true quality scores with minimal comparisons. A novel sampling procedure is proposed and benchmarked against state-of-the-art methods on simulated datasets. Bayesian sampling produces the most robust score estimates among previously established methods, while the proposed procedure consistently converges fastest on the underlying ranking with comparable score accuracy.

Keywords: pairwise comparisons, subjective evaluation, active sampling, Bayesian inference, tournaments

1. INTRODUCTION

Subjective assessments of processed audio signals can be obtained through direct numerical ratings, or implicit ratings derived from pairwise comparisons. In the latter

case, participants compare two audio signals and select the one they prefer or perceive as higher on a specific quality dimension, such as brightness. While direct ratings provide an immediate, fine-graded score, they also suffer from noisiness due to the difficulty of ensuring that rating scales are understood uniformly between participants [1]. Pairwise comparisons generally offer higher sensitivity and reduced noise [2], as well as more efficient testing procedures due to the low complexity of the comparison task [3].

Several psychoacoustic studies have deployed pairwise comparisons [4–7], often via round-robin tournaments where each stimulus is compared against all others. However, a round-robin tournament between n stimuli requires $n(n - 1)/2$ comparisons, which quickly becomes impractical for large n . Furthermore, pairwise comparisons are still subject to noise from differences in observer expertise and preferences, random errors and experimental biases.

Multiple rounds of comparisons can mitigate these issues, but efficient sampling is essential to reduce the number of necessary comparisons while maintaining accuracy. This paper compares several canonical and state-of-the-art sampling methods, introducing a novel sorting approach, Sort-MST. Preliminary results indicate that Sort-MST yields the strongest ranking performance with score accuracies comparable to optimal Bayesian sampling approaches.

Recent studies have focused on ranking the top- κ stimuli from a larger set [8]. In audio quality evaluation, it is often preferable to both rank all stimuli, and obtain accurate quality scores for each stimulus.

Extracting scores from pairwise comparison data requires a model to relate the observer responses to an arbitrary quality scale of interest. Two of the most common

*Corresponding author: j.webb24@imperial.ac.uk.

Copyright: ©2025 Webb & Picinali. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.





FORUM ACUSTICUM EURONOISE 2025

models are the Bradley-Terry (BT) model [9], and Thurstone's model [10]. Both models parameterise the n stimuli as having underlying true scores $s = \{s_1, \dots, s_n\}$, $s \in \mathbb{R}$ on a given subjective scale such as quality, distortion or loudness. The BT model defines the probability of stimulus i winning over stimulus j , P_{ij} , as:

$$P_{ij} = \frac{e^{s_i}}{e^{s_i} + e^{s_j}} \quad (1)$$

Thurstone's model assumes each score s_i is normally distributed with mean μ_i and standard deviation σ_i , giving P_{ij} :

$$P_{ij} = \Phi \left(\frac{\mu_i - \mu_j}{\sqrt{\sigma_i^2 + \sigma_j^2}} \right) \quad (2)$$

where Φ is the cumulative standard normal distribution function. Both models can be considered examples of generalised linear models, where the BT model is the logistic case, and Thurstone's model is the probit case. Estimated quality scores \hat{s}_i can be calculated from comparison data using maximum likelihood estimation (MLE). This work focuses on the BT model, examining the robustness of different sampling algorithms against random judgment errors and natural variances in observer expertise that occur in a real pairwise comparison context.

2. METHODS

Six algorithms are implemented in MATLAB and investigated through computational simulation.

2.1 Standard procedures

Standard pairwise comparison procedures follow fixed rules for selecting comparisons. Here, we include:

1. *Random sampling*. This acts as a baseline for minimum performance.
2. *Single-elimination knockout tournament*. A standard knockout (KO) tournament is implemented. Tournament trees are randomly initialised and run until the tournament is complete. New tournaments are initiated until the maximum number of comparisons is reached.

2.2 Active sampling

Active sampling involves adaptively selecting comparisons based on previous comparison results. We examine sorting and information gain approaches.

2.2.1 Sorting methods

Following an initial set of comparisons, scores can be calculated by fitting BT or Thurstone models to the results. Sorting methods assume that pairs with the smallest difference in scores are the most informative to compare. Three algorithms are tested:

3. *Swiss tournament*. A Swiss-style tournament pairs stimuli with similar scores in each round without repeating pairings. This design has been shown to be the most efficient among standard tournament formats [11]. Our implementation uses opponent match win percentage (OMW) as a tiebreaker. To ensure robust rankings, tournaments run for $\lfloor \log_2(n) \rfloor + 2$ rounds before restarting.
4. *Tree selection*. The binary tree selection procedure detailed in [8] is implemented as a modified knockout tournament. Though tree selection is primarily intended for top- κ ranking, it is extended here to sort all stimuli.
5. *Sort-MST*. We propose a novel sorting method that leverages Elo scores, commonly used in chess, to pair stimuli with similar scores. Pairs are ranked by smallest difference in Elo scores, and the inverse of this ranking is used as edge weights to construct a minimum spanning tree. This returns a group of pairs in each round, ensuring a balanced design that avoids oversampling specific pairs.

2.2.2 Information gain methods

Most optimal active sampling algorithms rely on information gain maximisation. The posterior distribution of quality scores is calculated at each stage, and the next pair is selected to maximise the information gained by observing the result of that comparison. Often, the information gain is calculated using the Kullback-Leibler (KL) divergence between the current distribution and the new distribution given the outcome of a potential future comparison, under a Bayesian statistical model.

6. *Hybrid-MST*. Hybrid-MST and related approaches [12, 13] are fully Bayesian active sampling procedures that have recently demonstrated optimal performance. We implement Hybrid-MST to indicate the performance of state-of-the-art methods.

2.3 Procedure

Monte Carlo simulations are performed on simulated datasets of 8, 16, and 32 stimuli. Each stimulus i is assigned a ground truth score, s_i , randomly sampled from





FORUM ACUSTICUM EURONOISE 2025

a uniform distribution $U(0, 5)$. Standard deviations in each score, σ_i , are sampled from three distinct ranges: $U(0, 0.4)$, $U(0, 0.7)$, and $U(0, 1.0)$. Smaller σ values emulate clear distinctions between stimuli, while larger values simulate greater noise and observer uncertainty.

In a comparison between stimuli i and j , observed scores r_i and r_j are drawn from normal distributions: $r_i \sim \mathcal{N}(s_i, \sigma_i)$, $r_j \sim \mathcal{N}(s_j, \sigma_j)$. The match is won by the stimulus with the higher sampled score, unless a random judgment error flips the result. Judgment errors occur with probabilities $\epsilon \in \{0.1, 0.2, 0.3, 0.4\}$ to evaluate the robustness of each method to random noise. Each algorithm is tested over a range of 0 to 15 standard trials, where a standard trial consists of $n(n-1)/2$ comparisons. This process is repeated 100 times for each procedure.

2.4 Evaluation

We assess the algorithms by fitting BT scores every $n(n-1)/8$ comparisons. Ranking quality is assessed against the ground truth using Spearman's rank-order correlation coefficient (ROCC), while score accuracy is evaluated with Pearson's correlation coefficient (PCC) and root mean square error (RMSE). Since BT scores are scale-invariant, the estimated scores are aligned with the ground truth scores using sigmoid regression before RMSE is calculated, ensuring fair assessments.

3. RESULTS

While the biggest factors in performance are the number of stimuli (n) and the noisiness of the comparisons (σ, ϵ), there are clear differences between the various methods. Bayesian methods, such as Hybrid-MST, are optimised to achieve the lowest RMSE, and consequently demonstrate high correlations with ground truth scores across different n and varying σ and ϵ . However, the proposed method, Sort-MST, obtains comparable PCC to Hybrid-MST in a range of conditions. With $n \geq 16$, the Swiss system also achieves strong correlations (Fig. 1).

Sort-MST also consistently achieves the highest ranking accuracy across different conditions (Fig. 2). Utilising a minimum spanning tree prevents imbalanced designs, meaning the algorithm can obtain accurate rankings without excessive distortions in PCC. By contrast, repeated KO tournaments yield outcomes equal to or worse than random sampling. Tree selection is generally an improvement on random sampling, but rarely performs optimally.

In conditions where uncertainty in relative stimulus strengths is high, sorting algorithms can outperform

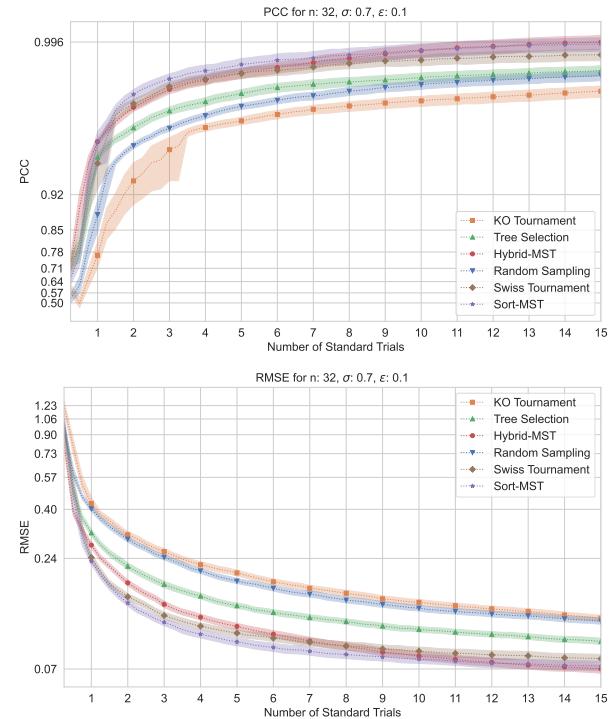


Figure 1. Mean PCC for and RMSE $n = 32$. All graphs show 95% confidence intervals bootstrapped over results from 100 repeats. Unless stated otherwise, all graphs show $\sigma \sim U(0, 0.7)$, $\epsilon = 0.1$.

Bayesian methods (Fig. 3). Increased noisiness in pairwise comparison outcomes may lead to misleading information gain predictions. In this case, methods that do not explicitly model score uncertainties — such as the Swiss system or Sort-MST — may perform significantly better.

4. CONCLUSION

Various pairwise comparison procedures are assessed using Monte Carlo simulations across a range of conditions. A novel sorting algorithm, Sort-MST, is shown to achieve the strongest rankings, with comparable score accuracy to Hybrid-MST, a Bayesian sampling approach that tends to yield the most robust scores. When uncertainty in the relative strengths of the stimuli is high, sorting methods can even outperform Bayesian sampling, with decreased computational expense. Future listening tests will validate these methods in real-world scenarios and investigate how scores can be aggregated across different observers.





FORUM ACUSTICUM EURONOISE 2025

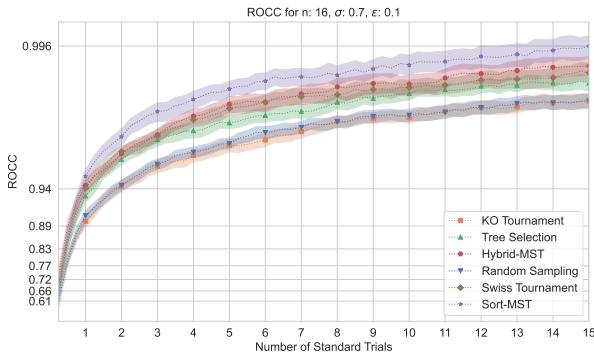


Figure 2. Mean ROCC for $n = 16$.

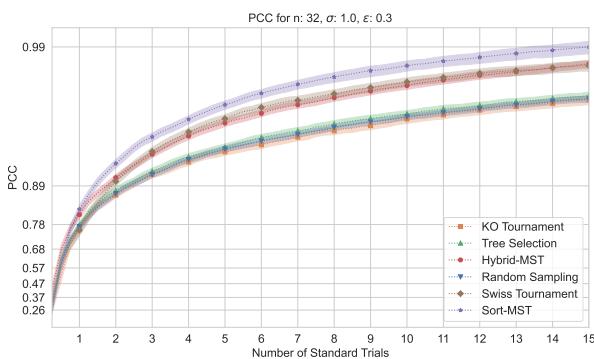


Figure 3. Mean PCC for $n = 32$, $\sigma \sim U(0, 1.0)$, $\epsilon = 0.3$

5. ACKNOWLEDGEMENTS

This work was supported by an EPSRC CASE conversion PhD scholarship, co-funded by Sonova.

6. REFERENCES

[1] J. Li, J. Wang, M. Barkowsky, and P. Le Callet, “Exploring the effects of subjective methodology on assessing visual discomfort in immersive multimedia,” *Electronic Imaging*, vol. 30, pp. 1–6, 2018.

[2] N. B. Shah, S. Balakrishnan, J. Bradley, A. Parekh, K. Ramch, and M. J. Wainwright, “Estimation from pairwise comparisons: sharp minimax bounds with topology dependence,” *Journal of Machine Learning Research*, vol. 17, no. 58, pp. 1–47, 2016.

[3] N. Stewart, G. D. Brown, and N. Chater, “Absolute identification by relative judgment,” *Psychological Review*, vol. 112, no. 4, pp. 881–911, 2005.

[4] E. Burke, S. Uppenkamp, and C. Koch, “A psychoacoustical study to investigate the perceived unpleasantness of infrasound combined with audio-frequency sound,” *Acta Acustica*, vol. 4, no. 5, pp. 1–12, 2020.

[5] M. Van Eeckhoutte, P. Folkeard, D. Glista, and S. Scollie, “Speech recognition, loudness, and preference with extended bandwidth hearing aids for adult hearing aid users,” *International Journal of Audiology*, vol. 59, no. 10, pp. 780–791, 2020.

[6] G. Volandri, F. Di Puccio, P. Forte, and L. Mattei, “Psychoacoustic analysis of power windows sounds: correlation between subjective and objective evaluations,” *Applied Acoustics*, vol. 134, pp. 160–170, 2018.

[7] B. C. Moore, C. Füllgrabe, and M. A. Stone, “Determination of preferred parameters for multichannel compression using individually fitted simulated hearing aids and paired comparisons,” *Ear and Hearing*, vol. 32, no. 5, pp. 556–568, 2011.

[8] N. Meyer-Kahlen and P. Hyvärinen, “Psychoacoustic ranking and selection using modified knockout tournaments,” *The Journal of the Acoustical Society of America*, vol. 156, pp. 879–890, Aug. 2024.

[9] R. A. Bradley and M. E. Terry, “Rank analysis of incomplete block designs: I. the method of paired comparisons,” *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.

[10] L. L. Thurstone, “A law of comparative judgment,” *Psychological Review*, vol. 34, no. 4, pp. 273–286, 1927.

[11] B. R. Sziklai, P. Biró, and L. Csató, “The efficacy of tournament designs,” *Computers & Operations Research*, vol. 144, 2022.

[12] J. Li, R. Mantiuk, J. Wang, S. Ling, and P. Le Callet, “Hybrid-MST: A hybrid active sampling strategy for pairwise preference aggregation,” in *Neural Information Processing Systems (NIPS 2018)*, pp. 1–11, 2018.

[13] A. Mikhailiuk, C. Wilmot, M. Perez-Ortiz, D. Yue, and R. K. Mantiuk, “Active sampling for pairwise comparisons via approximate message passing and information gain maximization,” in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 2559–2566, IEEE, 2021.

