Recovering quality scores in noisy pairwise subjective experiments using Negative log-likelihood

Andréas Pastor, Lukáš Krasula, Xiaoqing Zhu, Zhi Li, Patrick Le Callet

RECOVERING QUALITY SCORES IN NOISY PAIRWISE SUBJECTIVE EXPERIMENTS USING NEGATIVE LOG-LIKELIHOOD

Andréas Pastor[⋆] *Lukáš Krasula*[†] *Xiaoqing Zhu*[†] *Zhi Li*[†] *Patrick Le Callet*[⋆]

[⋆] Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
[†]Netflix Inc., Los Gatos, CA, USA

ABSTRACT

To gather larger datasets to train data-angry deep learning quality assessment models, crowdsourcing has become essential to recruit participants. These participants are asked their opinion by directly rating stimuli, e.g., using single or double stimulus methodologies, or indirectly by ranking stimuli or comparing distances as in the Maximum Likelihood Difference Scaling method. In crowdsourcing, participants' behaviors and environmental distractions are not controlled. So, the researcher must pay attention to the answers' reliability. Cleaning methods exist for direct annotation subjective methodologies. However, solutions for indirect annotation methods are limited. In this work, we propose a method based on the negative log-likelihood to detect spammers among participants from their answers. To demonstrate its use, we applied it in a quadruplet preference-based scenario. The proposed method requires low computation and can be integrated into active-sampling strategies, where annotations available per comparison are small. We demonstrate that our method is robust to various spammer behaviors and accurate by removing only spammers. It helps reduce the gap between data collected in in-lab conditions (i.e., no spammer) and through crowdsourcing: our method reduces estimated uncertainties around data-points by 50%, and RMSE between estimations from an in-lab experiment and the same experiment performed in crowdsourcing by 1.8.

Index Terms— Spammer removal, crowdsourcing, pair comparison, active-sampling

1. INTRODUCTION

Relying on crowdsourcing can increase the reach of an experiment to a larger population, speed up the annotation process and gather opinions on stimuli at scale. To train supervised models and benchmark systems, these subjective data must be clean and accurate. Removing the noise introduced by spammer behavior becomes a priority.

ITU standards [1, 2, 3, 4] are defined to clean subjective data from direct rating subjective methodologies (e.g., Absolute Category Rating (ACR), Double Stimuli Impairment Scale (DSIS)). Other initiatives [5, 6, 7] exist and are improving on these standards to provide better interpretation and accuracy on spammer removal.

However, there is limited research on data cleaning for indirect annotation strategies such as ranking, n-Alternative Forced Choice (n-AFC), pairwise comparison (PC), or Difference Scaling [8]. *Active-sampling* solutions [9, 10, 11, 12, 13, 14] and more recently [15, 16, 17, 18, 19] proposed solutions to *efficiently* select the most informative comparisons to retrieve accurate estimations and minimize experimental effort. These algorithms should be able to recover estimates in most situations, including conditions with bad annotator behavior: the spammers.

In [13], the authors proposed a method to perform active-sampling and noise removal simultaneously, but the solver has a significant time complexity, see results of [15] for more details. With active-sampling in crowdsourcing, the number of participants can be huge. Moreover, the number of annotations for each possible pair can be low or even null (i.e., sparsity of PCMs). To perform the solving, retrieving stimuli estimates, and estimating reliability score for each crowdsourcer is impractical due to the large number of parameters to model the thousands of unique participants and the thousands of stimuli. In [20], a solution for spammer removal for PC experiments using Cohen's Kappa and Rogers-Tanimoto dissimilarity measures is proposed. However, this method requires a group of participants to complete the same set of annotations to obtain a minimum number of annotations per pair and a significant intersection between the pairs evaluated by two participants to provide dissimilarity scores with substantial reliability. This makes the technique impractical in active-sampling scenarios.

In this work, we are trying to answer the question: how can we compute a reliability score per participant in an active-sampling scenario where a participant may be the only one to annotate a set of stimuli?

2. NEGATIVE LOG-LIKELIHOOD FOR SPAMMER DETECTION

This section presents our method to detect spammers. For two measures X and Y (e.g., qualities in a pair, a perceived difference in a triplet or a quadruplet), any recovery models [21, 22, 23, 8, 24] transform the preferences from a group of participants to a probability P that the measure X is greater than Y.

For example in the solver for MLDS[8], the probability is computed as:

$$P\left(X > Y\right) = \phi\left(\frac{\delta(q)}{\sigma}\right), \quad (1)$$

with $\phi\left(x\right)$, the cumulative standard normal distribution CDF. X is the perceived difference L_{cd} between the two stimuli from the first pair of the quadruplet, and Y is the difference L_{ab} in the second pair: the quadruplet is measured as $\delta\left(q\right) = \delta\left(a, b, c, d\right)$.

$$\delta\left(a, b, c, d\right) = L_{cd} - L_{ab} = |\psi_d - \psi_c| - |\psi_b - \psi_a|, \quad (2)$$

Where the ψ_i are the parameters of the solver to estimate the values for the stimuli.

From this, we can derive the likelihood of a session: the likelihood of the set R of n preferences from a participant,

$$L\left(\psi, \sigma\right) = \prod_{k=1}^{n} \phi\left(\frac{\delta(q^k)}{\sigma}\right)^{1-R_k} \phi\left(1 - \frac{\delta(q^k)}{\sigma}\right)^{R_k}, \quad (3)$$

where R_k is his/her preference for the k-th quadruplet, value of R_k is 0 or 1.

By applying the negative logarithm, we obtain a score per session with a minimum of 0. The score increases significantly each time an unlikely preference is given, and otherwise with a small amount.

3. SUBJECTIVE EXPERIMENT DATA

This section describes the two datasets we used in this work. The method we propose in the previous section can be applied to any indirect annotation subjective methodology. We choose to demonstrate it in a quadruplet scenario.

The first dataset is from an "in-lab" experiment on the quality of video patches, and the second one, named "crowdsourced" is its reproduction on Prolific[1], a crowdsourcing platform. These two datasets are subsets of the dataset presented in [25] on 12 out of the 20 contents. The task of participants, see visual example in figure 1, was to provide a preference on a quadruplet (a,b,c,d): where do you perceive a greater difference between a first pair of video patches and a second pair of video patches? The differences are encoding distortions generated from libaom-AV1[2]. We applied MLDS methodology to estimate supra-threshold differences in these small videos, *tubes*. The tubes are of size 64×64 pixels, 400 ms in length, and are extracted from 1080p video sources (SRCs). A quadruplet contains four tubes, from the same reference tube for intra-content comparisons, see fig.1, or from two different reference tubes for inter-content comparisons.

The "in-lab" dataset has 5235 judgments from 173 annotations sessions. A session lasts 6 minutes on average, and a participant gives his preferences on a set of quadruplets. These sets are generated by the active-sampling algorithm proposed in [24] for inter-content difference scaling.

[1]Prolific website: https://www.prolific.co/

[2]AV1 encoder v3.1.2, from AOM Alliance Open Thishttps://aomedia.googlesource.com/aom/

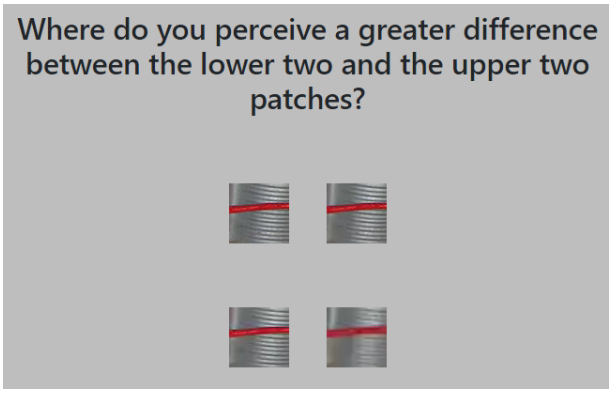

Fig. 1: Example of quadruplet presented to a participant to rate where he perceives a greater difference in the quadruplet.

There are 17 unique participants in this subjective study. They can be considered experts and non-spammers since they have been trained for the task and recruited in our laboratory with experience in quality assessment.

The "crowdsourced" dataset has 7650 judgments from 260 sessions performed by participants recruited on Prolific. Similarly, the participants were asked to give their preferences for quadruplets, following the same active-sampling procedure. These participants can be considered as naïve observers for quality assessment since the requirements to participate in the study are only on the spoken language and their approval rate. Although we have these requirements, we cannot ensure that all the participants aren't spammers.

In figure 2, we present the estimations obtained when we apply MLDS for inter-content scaling solver on the data. On the left, estimates were obtained from the judgments from experts from the "in-lab" dataset, and on the right, from the "crowdsourced" dataset. Each curve in the figures is a *perceptual curve* and models a perceived distance between a reference tube and the distorted versions of this reference tube. At distortion level 0 on the x-axis, the reference tube is encoded with quantization parameters (QP 0) of AV1. The other distortion levels correspond to different increasing QP values to compress the information while increasing the visibility of the distortions. An example of reading the figure will be: "the content modeled by the purple line is with most visible distortions when encoding with AV1".

In [27, 28], authors examine the evolution of subjective scores *discriminability* to show how well a subjective methodology can retrieve accurate estimates. A two-sample Wilcoxon test is performed on all the possible pairs of stimuli of the dataset. A p-value of 0.05 is used to compute the percentage of significantly different pairs. Here, we can use this discriminability ratio to indicate the quality of the collected data and how well the data points are separated.

In figure 2 (left), the Confidence Intervals (CIs) estimated by the solver are, on average, 0.05 across the 12 perceptual curves, and the discriminability ratio is 0.981. For the (right) one, the average CIs is 0.18, and the discriminability ratio is 0.970. With lower average CI values and a higher discrim-

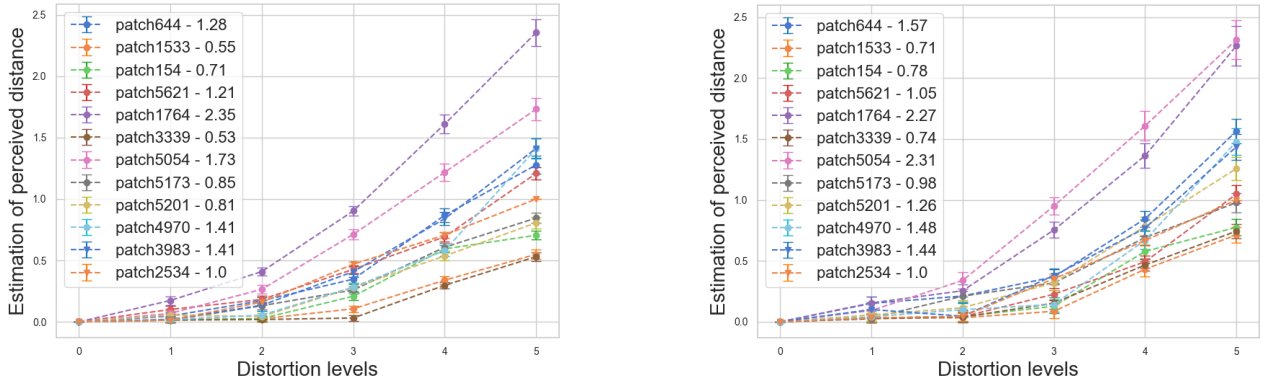

Fig. 2: MLE solving on the two datasets: (left) "in-lab" dataset and (right) "crowdsourced" dataset. Confidence Intervals (CIs) are obtained via bootstrapping [26] over 100 runs. The X-axis is the distortion levels applied on each reference tube, and the Y-axis is the estimated perceived distances by participants between the reference and the distorted tubes.

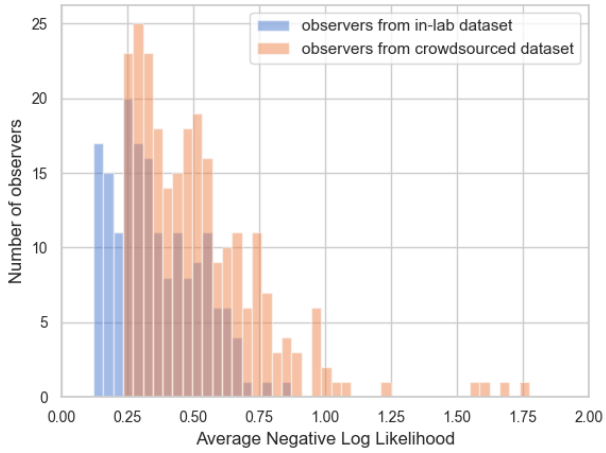

Fig. 3: Average negative log-likelihood on the sessions of the "in-lab" dataset in blue and the "crowdsourced" dataset in orange.

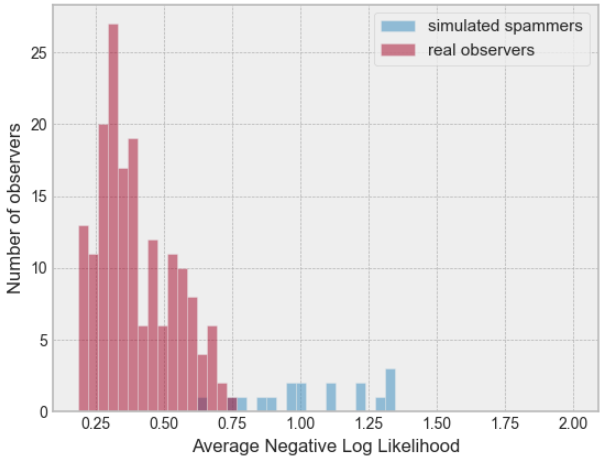

Fig. 4: Example of simulated spammers and how our method can correctly classify these sessions as spammers from their negative log-likelihood score: AUC = 0.997.

inability ratio for the "in-lab" dataset, we can conclude that the "in-lab" dataset produces more accurate estimates. This can be explained by the absence of spammers among the participants and their expertise in quality assessment.

In figure 3, the 260 sessions from the "crowdsourced" dataset are represented in orange as the average Negative Log-Likelihood (NLL) of the judgments of each session, using the ψ estimated from solving presented in figure 2 (right). In blue are the sessions from the "in-lab" dataset, with ψ shown in figure 2 (left). We can see that the NLL of the "crowdsourced" sessions is, on average larger compared to the session from the "in-lab" experiment. This is due to spammers (e.g., sessions with NLL scores above 1). The responses of these spammers increase the noise in data, affecting the estimation of probabilities by the solving algorithm and increasing the NLL of non-spammers sessions.

4. PERFORMANCE EVALUATIONS

In this section, we evaluate our proposed method for spammers detection and removal.

4.1. Simulated spammers detection

We simulate spammers inside the "in-lab" dataset to evaluate if the proposed method can detect spammers. With this simulation, we want to find a threshold in the NLL score that can consistently remove spammers.

There are different profiles of spammers: voters that answer at random, with patterns (e.g., ABAB, AAAA, BBBB), with inversion pattern where the votes are reversed compared to the expected task (e.g., participants that misunderstood the task), or profiles that are a mix of the previous ones.

Once we simulated these spammers, we added them to the "in-lab" dataset with the actual observers. We remind the reader that these natural observers are considered non-spammers from their expertise and the training they received.

We choose that these new sessions represent 10% of the dataset: 18 simulated sessions are added. We performed an MLDS recovery on this data ("in-lab" + spammers) and applied the method based on the negative log-likelihood described above to characterize each session. We used Receiver Operating Characteristic (ROC) and Area Under the Curve score (AUC) to analyze the method's performance and separate simulated spammers from natural observers.

An example is provided in figure 4, where in red is the distribution of sessions from the "in-lab" dataset and in blue the added sessions of simulated spammers. From ROC anal-

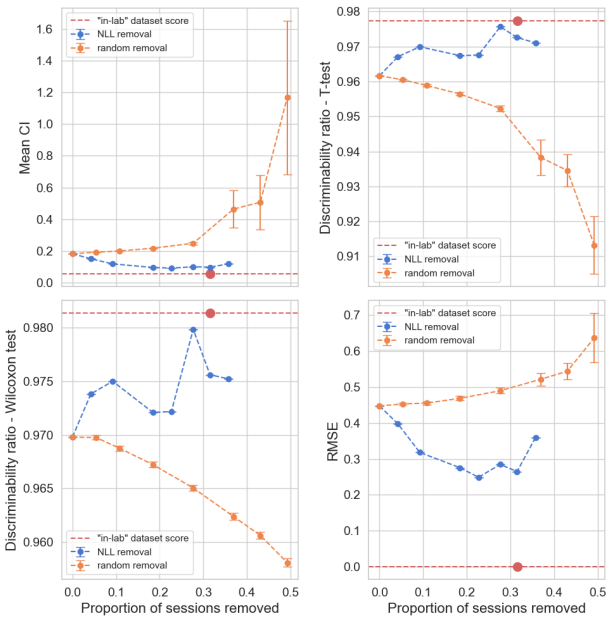

Fig. 5: Results obtained when removing from the "crowdsourced" dataset, sessions classified as spammers by our method in blue, or sessions selected randomly in orange. The orange curve is computed as the average on 1000 permutation runs. On X-axis is the proportion of sessions removed from the "crowdsourced" dataset. In red is the score obtained from the 260 sessions from the "in-lab" dataset, a 31.5% smaller dataset represented by a red dot.

ysis, the AUC score is 0.997, translating to a good detection of spammer sessions.

4.2. Real spammers dectection

The simulation presented in the previous section is reproduced 1000 times, each time generating new simulated spammers, to estimate a precise AUC score for our method and a set of thresholds $th_{X\%}$ corresponding to NLL values that can separate on average $X\%$ of simulated spammers from the real ones, with X varying from 10% to 100%. We applied these thresholds to the "crowdsourced" dataset and removed 11 to 96 sessions.

To compare the effect of removing the sessions that are flagged by our method and ensure that we are not eliminating non-spammer sessions, we design an experiment where we track the evolution of the size of the retrieved CI by MLE solver when we remove these spammers sessions versus when we remove the same number of sessions at random.

Improvement in average estimated CI size: in figure 5 (top left), we can see that when we remove sessions at random in the "crowdsourced" dataset (orange curve), the solving procedure estimate values with larger CI, which is expected since we are removing potentially valuable data.When we remove the session classified as spammers by our method (in blue), we can see that the mean CIs size is decreasing. In red is the mean CIs size obtained on the "in lab" dataset, 0.05, which is $3.6\times$ smaller than 0.18 for the "crowdsourced" dataset. After removing 48 spammers sessions (around 18% of the dataset), the gap between the "crowdsourced" dataset and the "in-lab" dataset decreased to 0.096: a reduction by 2.

Improvement in discriminability: in figure 5 (top right and bottom left), the discriminability ratio [27, 28] evolution is explored. In blue, removing spammers increases the ratio value and approaches the ratio value of the "in-lab" dataset in red. Proving that our method is improving, or maintaining at worse, the discriminability between stimuli of the dataset. In orange, discriminability decreases as we remove random sessions, which is expected.

Improvement in RMSE: in figure 5 (bottom right), the Root Mean Square Error (RMSE) evolution is explored. The RMSE is computed between the estimated values from the "in-lab" dataset and those from the truncated "crowdsourced" dataset. When removing random sessions, in orange, the RMSE and gap with the "in-lab" dataset increase. With our method, in blue, the RMSE reduces from 0.45 to 0.25: a 1.8 reduction.

One can see from all the indicators that there is a threshold after which removing more sessions decreases performance. Mean CI and RMSE start to increase again, and discriminability ratios start to decrease. This is because we start removing valuable sessions. In this particular experiment, removing more than 20% of the sessions hurts Mean CI and RMSE indicators performances.

5. RECOMMENDATIONS ON THE USAGE

To use the method, having a clean version of the dataset to find NLL thresholds is unnecessary. Our approach can be directly applied to an uncurated dataset. The simulation presented in the previous section can add simulated spammers in any dataset to derive thresholds. To validate these thresholds, the analysis shown in figure 5 is sufficient to tell if removing spammers' sessions is effective.

6. CONCLUSION

We presented a method based on negative logarithm likelihood to remove spammers from subjective experiments based on indirect annotation methodologies. To demonstrate the method's effectiveness, we first validated on a dataset without spammers, "in-lab", that our approach can detect and classify simulated spammers correctly. Secondly, we apply our method to a dataset collected in crowdsourcing with real spammers. We show how our method positively impacts the data quality obtained after removing the participants classified as spammers: reduction of CIs size by 2 and 1.8 on the RMSE between datasets while maintaining or slightly improving the discriminability between stimuli.

As a future work, this method could be integrated into any active-sampling procedure to rule out or not new incoming annotations of a participant before generating a new set of samples to be annotated. Thus, avoiding the noise of a spammer earlier in creating a dataset.

7. REFERENCES

[1] ITU Recommendation BT.500-14, "Methodologies for the Subjective Assessment of the Quality of Television Images," 2019.

[2] ITU-T Rec. P.910, "Subjective video quality assessment methods for multimedia applications," 2008.

[3] ITU-T Rec. P.913, "Methods for the subjective assessment of video quality, audio quality and audiovisual quality of internet video and distribution quality television in any environment," 2016.

[4] ITU-R Rec. BS.1534-3, "Method for the subjective assessment of intermediate quality levels of coding systems," 2015.

[5] Zhi Li, Christos G. Bampis, Lucjan Janowski, and Ioannis Katsavounidis, "A simple model for subject behavior in subjective experiments," *CoRR*, vol. abs/2004.02067, 2020.

[6] Zhi Li and Christos G Bampis, "Recover subjective quality scores from noisy measurements," in *2017 Data compression conference (DCC)*. IEEE, 2017, pp. 52–61.

[7] Jing Li, Suiyi Ling, Junle Wang, and Patrick Le Callet, "A probabilistic graphical model for analyzing the subjective visual quality assessment data from crowdsourcing," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 3339–3347.

[8] Kenneth Knoblauch, Laurence T Maloney, et al., "Mlds: Maximum likelihood difference scaling in r," *Journal of Statistical Software*, vol. 25, no. 2, pp. 1–26, 2008.

[9] Mark E Glickman and Shane T Jensen, "Adaptive paired comparison design," *Journal of statistical planning and inference*, vol. 127, no. 1-2, pp. 279–293, 2005.

[10] Thomas Pfeiffer, Xi Alice Gao, Yiling Chen, Andrew Mao, and David G Rand, "Adaptive polling for information aggregation," in *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[11] Jing Li, Marcus Barkowsky, and Patrick Le Callet, "Analysis and improvement of a paired comparison method in the application of 3dtv subjective experiment," in *2012 19th IEEE International Conference on Image Processing*. IEEE, 2012, pp. 629–632.

[12] Jing Li, Marcus Barkowsky, and Patrick Le Callet, "Boosting paired comparison methodology in measuring visual discomfort of 3dtv: performances of three different designs," in *Stereoscopic Displays and Applications XXIV*. International Society for Optics and Photonics, 2013, vol. 8648, p. 86481V.

[13] Xi Chen, Paul N Bennett, Kevyn Collins-Thompson, and Eric Horvitz, "Pairwise ranking aggregation in a crowdsourced setting," in *Proceedings of the sixth ACM international conference on Web search and data mining*, 2013, pp. 193–202.

[14] Peng Ye and David Doermann, "Active sampling for subjective image quality assessment," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 4249–4256.

[15] Jing Li, Rafal K. Mantiuk, Junle Wang, Suiyi Ling, and Patrick Le Callet, "Hybrid-mst: A hybrid active sampling strategy for pairwise preference aggregation," *CoRR*, vol. abs/1810.08851, 2018.

[16] Qianqian Xu, Jiechao Xiong, Xi Chen, Qingming Huang, and Yuan Yao, "Hodgerank with information maximization for crowdsourced pairwise ranking aggregation," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[17] Edwin Simpson and Iryna Gurevych, "Scalable bayesian preference learning for crowds," *Machine Learning*, pp. 1–30, 2020.

[18] Aliaksei Mikhailiuk, Clifford Wilmot, Maria Perez-Ortiz, Dingcheng Yue, and Rafał K Mantiuk, "Active sampling for pairwise comparisons via approximate message passing and information gain maximization," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 2559–2566.

[19] Hui Men, Hanhe Lin, Mohsen Jenadeleh, and Dietmar Saupe, "Subjective image quality assessment with boosted triplet comparisons," *IEEE Access*, vol. 9, pp. 138939–138975, 2021.

[20] Ali Ak, Mona Abid, Matthieu Perreira Da Silva, and Patrick Le Callet, "On spammer detection in crowdsourcing pairwise comparison tasks: Case study on two multimedia qoe assessment scenarios," in *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2021, pp. 1–6.

[21] Louis Leon Thurstone, "A law of comparative judgement," *Psychological Review*, vol. 34, pp. 278–286, 1927.

[22] Ralph A. Bradley and Milton E. Terry, "The rank analysis of incomplete block designs — I. The method of paired comparisons," *Biometrika*, vol. 39, pp. 324–345, 1952.

[23] Amos Tversky, "Elimination by aspects: A theory of choice.," *Psychological review*, vol. 79, no. 4, pp. 281, 1972.

[24] Andréas Pastor, Lukáš Krasula, Xiaoqing Zhu, Zhi Li, and Patrick Le Callet, "Improving maximum likelihood difference scaling method to measure inter content scale," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 2045–2049.

[25] Andréas Pastor, Lukáš Krasula, Xiaoqing Zhu, Zhi Li, and Patrick Le Callet, "On the accuracy of open video quality metrics for local decision in av1 video codec," in *2022 IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 4013–4017.

[26] Felix A Wichmann and N Jeremy Hill, "The psychometric function: Ii. bootstrap-based confidence intervals and sampling," *Perception & psychophysics*, vol. 63, no. 8, pp. 1314–1329, 2001.

[27] Randy F Fela, Andréas Pastor, Patrick Le Callet, Nick Zacharov, Toinon Vigier, and Soren Forchhammer, "Perceptual evaluation on audio-visual dataset of 360 content," in *2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. IEEE, 2022, pp. 1–6.

[28] Yana Nehmé, Jean-Philippe Farrugia, Florent Dupont, Patrick Le Callet, and Guillaume Lavoué, "Comparison of subjective methods for quality assessment of 3D graphics in virtual reality," *ACM Transactions on Applied Perception (TAP)*, vol. 18, no. 1, pp. 1–23, 2020.