Comparison of crowdsourcing and laboratory settings for subjective assessment of video quality and acceptability & annoyance

Ali Ak, Abhishek Gera, Denise Noyes, Hassene Tmar, Ioannis Katsavounidis, Patrick Le Callet

COMPARISON OF CROWDSOURCING AND LABORATORY SETTINGS FOR SUBJECTIVE ASSESSMENT OF VIDEO QUALITY AND ACCEPTABILITY & ANNOYANCE

Ali Ak[*], *Abhishek Gera*[‡], *Denise Noyes*[‡], *Hassene Tmar* [‡], *Ioannis Katsavounidis* [‡], *Patrick Le Callet*[*§]

[*]Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
[†] Meta Platforms, Inc., Menlo Park, USA
[§]Institut Universitaire de France (IUF), France

ABSTRACT

User satisfaction is significantly influenced by their expectations of video quality. Even when users are presented with identical video stimuli, the Quality of Experience (QoE) can vary based on the context. The acceptability and annoyance paradigm serves as a tool to understand this relationship by measuring QoE as a function of user expectations and video quality. Traditionally, subjective experiments assessing QoE have been conducted in controlled laboratory settings. While the extension of traditional video quality experiments to crowdsourcing settings is well-explored, the impact of crowdsourcing on QoE studies has not been thoroughly examined. This study explore the potential use of crowdsourcing platforms for acceptability & annoyance experiments. To this end, video quality and acceptability & annoyance experiments were conducted in both laboratory and crowdsourcing settings. The findings reveal a more linear relationship between video quality and QoE in crowdsourcing settings. Subjects in crowdsourcing settings tend to have higher expectations of video quality, resulting in a slight increase in acceptability & annoyance thresholds compared to laboratory experiments. Analyses suggest that extending acceptability & annoyance experiments to crowdsourcing is not as straightforward as extending traditional video quality experiments. In crowdsourcing settings, priming subject expectations with instructions is not as effective as it is in laboratory conditions.

Index Terms— acceptability and annoyance, quality of experience, video quality, crowdsourcing, user generated content

1. INTRODUCTION

In the past decade, An impressive number of subjective video quality datasets has been developed and published. However, for a video streaming service provider, such as those operating on online social media platforms or video streaming platforms, perceived video quality alone is not the optimal metric. What holds greater importance for service providers is understanding whether the video quality meets user expectations (avoiding annoyance), or at least surpasses acceptable levels. In response to this need, the acceptability & annoyance paradigm has gained popularity in recent years [1, 2, 3]. The acceptability & annoyance paradigm is particularly relevant in the context of online social media platforms, where an immense volume of videos is streamed every second.

Acceptability and annoyance thresholds define the video quality levels at which a video stops being acceptable or starts being annoying, respectively. It's well-established that video quality alone doesn't directly express the acceptability & annoyance of video content, requiring consideration of user expectations and the consumption context [2]. Previous efforts have been made to establish mappings between video quality scores and acceptability & annoyance scale for estimating these thresholds [4, 5, 2]. However, this process typically involves two distinct experiments and needs to be repeated for each context. As a consequence, the required number of participants significantly increases, making the collection of acceptability & annoyance labels for large-scale datasets practically challenging in laboratory settings.

To address the limitations of laboratory experiments, numerous efforts have been made to extend perceptual quality experiments to crowdsourcing [6, 7, 8]. Following the recommendations outlined in these studies, several datasets [9, 10] have been successfully collected on crowdsourcing platforms like Prolific [11] and Amazon Mechanical Turk [12]. However, the extension of Quality of Experience (QoE) studies, particularly within the acceptability & annoyance paradigm, to crowdsourcing remains under-explored.

In acceptability & annoyance experiments, user expectations are typically primed with carefully designed instructions [2, 3]. For instance, in [2], authors employed two user profiles—Basic users costing €6 per month and Premium users costing €12 per month for the video streaming service. Subjects were primed with written instructions before the experiment. Another study [3] primed subjects based on their remaining quota in their mobile data plan, demonstrating the effectiveness of this approach in stimulating different contexts. However, we hypothesize that priming subjects similarly in a crowdsourcing setting may not be as effective due to the potentially lower attention span of the subjects.

In this study, we aimed to assess the impact of crowdsourcing on acceptability & annoyance labels by conducting video quality and acceptability & annoyance experiments in a crowdsourcing setting. We then compared the results with experiments conducted in controlled laboratory conditions. Our findings indicate that, in both settings, participants exhibit comparable discriminatory power in video quality experiments. However, for acceptability & annoyance experiments, we observed a slightly lower discriminative power of subjects in the crowdsourcing setting.

Furthermore, our analysis revealed that the relationship between video quality and acceptability & annoyance is more linear in the crowdsourcing setting, suggesting a potential lack of understanding of the task by participants. Consequently, we observed a slight increase in acceptability & annoyance thresholds obtained in crowdsourcing settings. These results shed light on the nuances of conducting acceptability & annoyance experiments in crowdsourcing setting.

Fig. 1. Example of selected SRCs from the IPI-VUGC Dataset.

2. SUBJECTIVE EXPERIMENTS

In this study, we build upon the publicly available IPI-VUGC[1] dataset [13]. The IPI-VUGC dataset provides subjective opinion scores for both video quality and acceptability & annoyance. We choose 18 source content (SRC), each accompanied by 6 processed video sequences (PVS), from the IPI-VUGC dataset. Subsequently, we replicate the same experiment on Prolific [11] crowdsourcing platform, ensuring minimal differences between the two settings.

In summary, the study consists four distinct experiments: "InLab ACR", "InLab AccAnn", "Crowdsourcing ACR", and "Crowdsourcing AccAnn". The content used in each experiment is identical, and there is no deviation in the experiment design between the InLab and Crowdsourcing studies. This consistent approach enables a robust comparison across different experimental settings.

2.1. Content

As mentioned earlier, we employed 18 SRCs from the IPI-VUGC dataset for our crowdsourcing experiments. These SRCs have a duration of 5 seconds and a resolution of 1080p. The dataset also includes 6 PVS for each SRC, utilizing the h264 [14] coding algorithm at varying spatial resolutions and constant rate factors (CRFs). Notably, the videos are vertically oriented and they don't contain an audio channel. They are predominantly user-generated and recorded by mobile phones with a few exceptions of drone footage. Some SRCs in the dataset were edited with text and sticker overlays—a characteristic feature of User-Generated Content (UGC) commonly found on online social media platforms. Figure 1 presents 4 example of the selected SRCs from the IPI-VUGC Dataset.

2.2. Experiment Methodologies

Video Quality Experiment relies on Absolute Category Rating with Hidden Reference (ACR-HR) methodology with the classical scale ["Bad", "Poor", "Fair", "Good", "Excellent"]. This scale is numerically represented in the range [1, 5], where 1 corresponds to "Bad" and 5 corresponds to "Excellent." Participants are simply instructed to provide a rating for the video quality within this numerical range, without any additional instruction. Similar to the laboratory experiment from IPI-VUGC Dataset, four videos were selected for non-explicit training at the beginning of each session. These videos were intentionally selected to cover the video quality range found in the entire dataset to help aligning the expectations of the users. To prevent repetition with the experiment stimuli, the training videos were excluded from the dataset.

[1]IPI-VUGC Dataset: https://zenodo.org/doi/10.5281/zenodo.10475209

Acceptability and Annoyance (AccAnn) Experiment adopts the single-step acceptability & annoyance procedure proposed in [2]. This method simplifies the classical multi-step approach [15, 1] by combining the acceptability & annoyance questions into a single question. The scale is then presented to subjects as "Not Annoying", "Annoying but Acceptable", and "Not Acceptable", color-coded for clarity.

Given the importance of instructions in acceptability & annoyance experiments on setting user expectations, the crowdsourcing experiment employed the following instructions, following the instructions proposed in the IPI-VUGC dataset:

"You are going to participate in an experiment determining the acceptability and annoyance of videos. You will need to imagine yourself scrolling through your preferred social media platform (e.g., Facebook, Instagram, TikTok, etc.) and encountering these videos. Based on your expectations of the video quality in these encounters, you will need to rate the quality of the video in terms of Acceptability and Annoyance.

- *The video is not annoying when its quality satisfies or exceeds your expectations.*
- *The video is annoying but acceptable when its quality is acceptable but not completely satisfies your expectations.*
- *The video is not acceptable when its quality does not meet your expectations. Such video quality makes you think about skipping to the next video."*

2.3. InLab Experiments

As stated earlier, we use the subjective video quality scores and acceptability & annoyance labels provided in IPI-VUGC dataset [13] and refer these as the InLab experiment data.

Video quality scores in the IPI-VUGC dataset were obtained through an ACR-HR experiment conducted in controlled laboratory conditions. The AccAnn labels were gathered using a single-step design within the context of an online social media platform. The subjective experiments occurred in controlled laboratory conditions, utilizing an iPhone 14 Pro device with the native AV Player[2]. Subjects held the device freely with an armrest during the experiment, and screen brightness was maintained at a fixed level for consistency. The remaining experiment details adhered to the recommendations outlined by the ITU [16].

2.4. Crowdsourcing Experiments

Crowdsourcing experiments were conducted on Prolific [11] utilizing the participant pool available on the platform. Following the recommendations in previous studies, the dataset was divided into three sessions to accommodate the potentially shorter attention span of participants in crowdsourcing settings [6, 7]. Both video quality and AccAnn experiments were conducted across three sessions. Participants were required to complete 1000 tasks, ensuring an approval rate exceeding 99.5% on Prolific. On average, each stimulus was rated by 35 unique participants.

In contrast to InLab experiments, crowdsourcing experiments were carried out on computer screens, with display specifications limited to 1080p resolution to regulate the presentation of video stimuli. To maintain control over the video stimulus presentation, the videos were downscaled to 1080 pixels in height before the experiment, thereby avoiding reliance on the native sampling algorithms of participants' devices.

[2]https://developer.apple.com/documentation/avfoundation/avplayer/

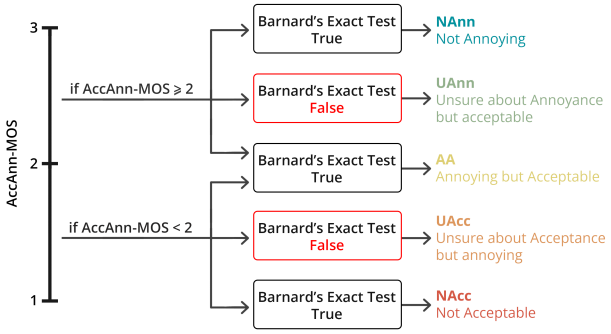

Fig. 2. Overview of the algorithm that determines the AccAnn category of a stimulus based on its AccAnn-MOS and distribution of individual AccAnn labels.

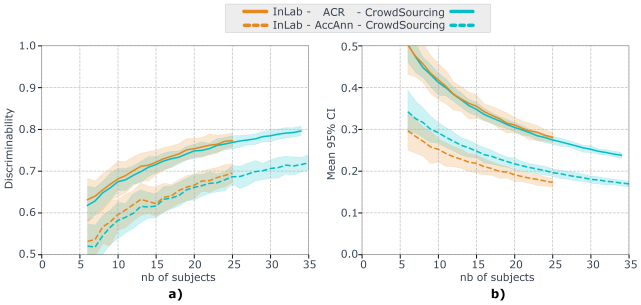

Fig. 3. The comparison between the four experiments is illustrated in terms of a) discriminability and b) mean 95% confidence intervals, with solid and dashed lines representing the ACR and AccAnn experiments, respectively. The color fills around each line indicate the 95 percentile range of the bootstrap iterations. InLab experiments are depicted in orange, while crowdsourcing experiments are represented in teal.

2.5. Representation of Subjective Opinions

The subjective opinions gathered in the experiments are represented in three distinct forms. Traditionally, video quality scores are expressed as Mean Opinion Scores (MOS). In this study, MOS is denoted as ACR-MOS to prevent the confusion with MOS of acceptability & annoyance. ACR-MOS represents the video quality on a continuous scale within the range of [1, 5], where higher values indicate superior video quality. To mitigate bias and inconsistencies linked to raw subjective opinions, the subjective scores were processed using ZREC [17], a MOS recovery algorithm.

Likewise, acceptability & annoyance scores are represented on a continuous scale and referred as AccAnn-MOS. The AccAnn-MOS scale spans the range of [1, 3], where higher values correspond to a superior QoE (*i.e.,* "Not Annoying").

A categorical representation is also employed for acceptability and annoyance labels. Similar to previous studies, each stimulus is categorized into one of five groups based on the distribution of individual AccAnn labels [2, 3]. Figure 2 provides a summary of the algorithm. Each stimulus is assigned to one of the three primary categories (NAnn, AA, and NAcc) or one of the two threshold categories (UAnn and UAcc). If the distribution of individual opinion scores for a given video exhibit statistically significant agreement, it is assigned to one of the main categories based on the majority opinion. Otherwise, the stimulus is assigned to one of the threshold categories. UAnn is the threshold category for stimuli where the video quality starts to be annoying, while UAcc is the threshold category for stimuli that start to become unacceptable. The color codes used for the categories in Figure 2 are consistent throughout the paper.

3. DISCRIMINABILITY AND MEAN-CI ANALYSIS

In this section, we make a comparison of the discriminatory power between crowdsourcing and laboratory settings, utilizing two metrics: discriminability and mean 95% confidence interval (CI) [6, 18, 19]. Higher discriminability is desired to design cost-efficient and reliable experiments. Moreover, lower mean 95% CI is indicative of higher clarity in the collected data [17].

To measure the discriminatory power with varying number of observers, we randomly select n subjects from the total pool of participants (24 for InLab and 34 for crowdsourcing experiments) through a bootstrap procedure with 100 iterations. At each iteration, we conduct two-sample Wilcoxon test on ACR-MOS and AccAnn-MOS values of all possible pairs of stimuli in the dataset. Pairs with a p-value of 0.05 is considered to be statistically significant. Additionally, we compute the mean 95% CI for all stimuli at each iteration.

In Figure 3-a, the discriminability of each experiment is plotted as a function of the number of subjects.Both ACR and AccAnn experiments in both settings exhibit similar discriminability. Furthermore, the discriminability is higher in ACR experiments compared to AccAnn experiments. This observation can be attributed to the finer granularity of the ACR scale (5 levels in ACR scale vs 3 levels AccAnn) and the inherent ambiguity associated with the AccAnn task.

Similar conclusions can be drawn based on the mean 95% CI values, as depicted in Figure 3-b. Notably, there is a distinction between the mean 95% CI values of InLab-AccAnn and crowdsourcing-AccAnn experiments. Since the same distinctions is not observed between the mean 95% CI values InLab-ACR and crowdsourcing-ACR experiments, this cannot be explained just by the varying viewing conditions in the crowdsourcing setting. In fact, it implies that the AccAnn experiment was better understood by the subjects in InLab experiment since with same discriminability, it results in lower mean 95% CI compared to Crowdsourcing-AccAnn.

4. COMPARISON OF SUBJECTIVE OPINIONS

In this section, we conduct a comparison of subjective ratings to determine whether similar conclusions can be drawn in crowdsourcing and laboratory settings.

4.1. ACR-MOS and AccAnn-MOS

In Figure 4, the correlation of ACR-MOS and AccAnn-MOS between InLab and crowdsourcing experiments is presented. Notably, compared to AccAnn-MOS values, ACR-MOS values exhibit a slightly stronger correlation between the laboratory and crowdsourcing settings. In the crowdsourcing experiment, acceptability & annoyance of videos in the mid to high-quality range are slightly lower. The category "Annoying but Acceptable" (where AccAnn-MOS is 2) is also less frequently used in crowdsourcing settings. This difference in the utilization of the AccAnn scale might have a

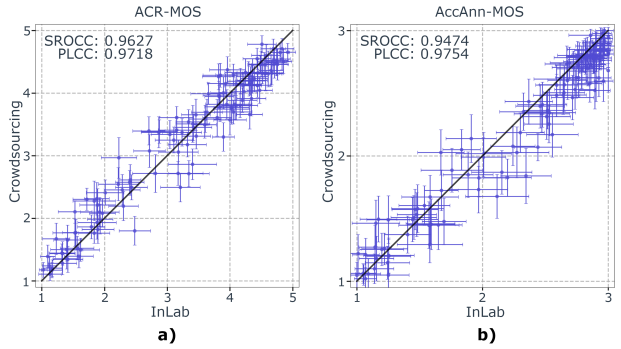

Fig. 4. Correlation of a) ACR-MOS and b) AccAnn-MOS between the InLab and crowdsourcing settings. Each point represent a video in the dataset, where the horizontal and vertical axes represents the scores acquired in InLab and crowdsourcing experiments, respectively.

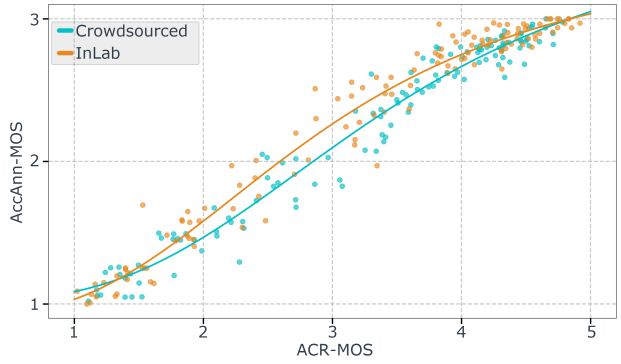

Fig. 5. Relation between ACR-MOS and AccAnn-MOS values in laboratory and crowdsourcing settings. A 4-parameter logistic function is fitted for each setting. Horizontal and vertical axes represents the ACR-MOS and AccAnn-MOS values, respectively.

minor impact on the mapping between video quality and acceptability and annoyance.

The mapping between ACR-MOS and AccAnn-MOS is visualized in Figure 5. Relationship between ACR-MOS and AccAnn-MOS is more linear in the crowdsourcing experiment. In addition, in the mid and high-quality range, subjects in the crowdsourcing experiment appear to have slightly higher expectations of video quality compared to subjects in the InLab experiment, consistent with observations in Figure 4.

4.2. Acceptability & Annoyance Categories

In addition to continuous scale comparisons (ACR-MOS and AccAnn-MOS), we can compare the acceptability & annoyance categories between the laboratory and crowdsourcing settings. The algorithm illustrated in Figure 2 is employed to determine the acceptability & annoyance categories of stimuli.

Figure 6 presents categorical comparisons between InLab-AccAnn and crowdsourcing-AccAnn experiments. Stimuli are ordered from left to right based on the AccAnn-MOS values in the InLab experiment. The same color coding is used for acceptability & annoyance categories in both settings, with circles representing the laboratory setting and squares representing the crowdsourcing experiment. For each stimulus, the distance between the circle and rectangle represents the difference in terms of AccAnn-MOS, as shown on the vertical axis. Lines between InLab (circles) and crowdsourcing (squares) samples are black if the two settings are in agreement and red if there is a disagreement.

Table 1. Acceptability and annoyance thresholds in terms of ACR-MOS and UVQ. ACR-MOS and UVQ values has the theoretical range of [1, 5]

	Acceptability		Annoyance	
	InLab	crowdsourcing	InLab	crowdsourcing
ACR-MOS	1.9804	2.1157	3.4001	3.6543
UVQ	3.1891	3.1616	3.6508	3.6898

The InLab and crowdsourcing experiments appear to be in agreement for most stimuli. Among the total 126 stimuli in the dataset, only 23 stimuli are categorized differently. Importantly, none of the mismatches places the content more than one category away. In other words, mismatches in categories occur only within neighboring categories. Meaning that the two settings are only in disagreement of the statistical significance of the stimuli, rather than its main category. Moreover, the majority of stimuli are categorized at a lower Quality of Experience (closer to "Not Acceptable") in the crowdsourcing experiment, aligning with previous observations in continuous scale comparisons.

5. ACCEPTABILITY AND ANNOYANCE THRESHOLDS

In this section, we compare the metric thresholds for acceptability & annoyance in laboratory and crowdsourcing settings. The acceptability threshold is defined as the value (within the theoretical range of the metric) below which the video quality is deemed unacceptable. The annoyance threshold is defined as the value below which the video quality starts to become annoying.

UGC videos exhibit considerable variation in quality without access to a pristine reference, presenting unique challenges for quality prediction. Given that the dataset comprises UGC content with non-pristine references (see second and third images in Figure 1), full-reference video quality metrics (e.g., VMAF [20]), designed to predict the difference mean opinion score, are not suitable for this task. Therefore, we rely on UVQ [21], a no-reference video quality metric designed to predict the perceptual quality of UGC videos. It is important to note that while the experiments can be extended with other suitable no-reference metrics, such extensions are beyond the scope of the current study.

Similar to [2, 3], acceptability & annoyance thresholds are defined as the mean score of all stimuli in “UAcc” and “UAnn” categories, respectively. The categories “UAcc” and “UAnn” are determined by Barnard’s test based on subjective annotations. From a statistical perspective, “UAcc” category represents the condition where half of the subjects perceive the video quality as “Not Acceptable” while the other half thinks it is “Annoying but Acceptable”. Similarly, “UAnn” is placed between the “Annoying but Acceptable” and “Not Annoying” categories.

Table 1 presents the acceptability & annoyance thresholds in terms of ACR-MOS and UVQ values for both InLab and crowdsourcing experiments. Consistent with earlier observations, acceptability & annoyance thresholds in terms of ACR-MOS are slightly higher in the crowdsourcing setting. For instance, an ACR-MOS

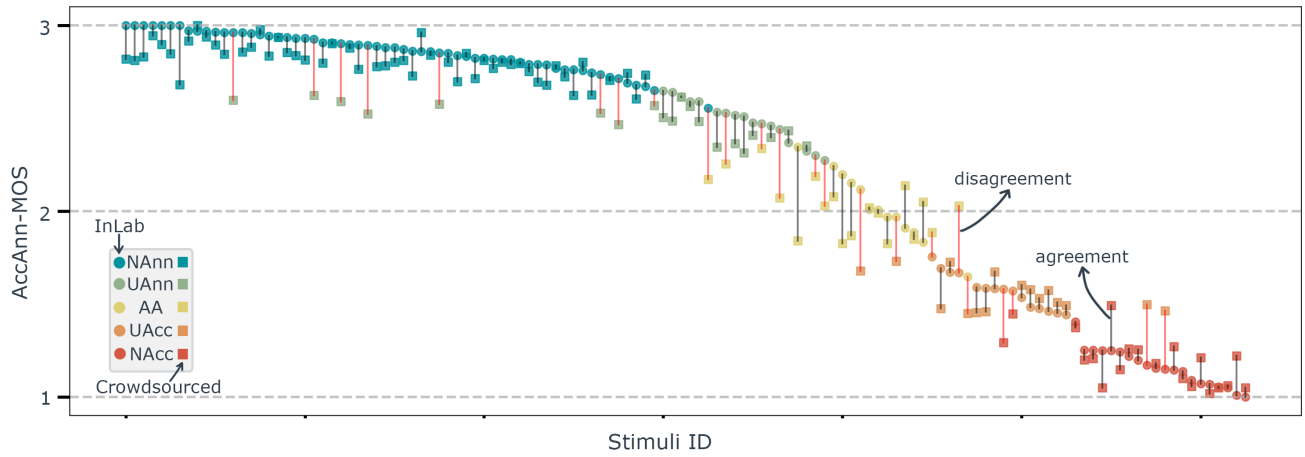

Fig. 6. Comparison of acceptability & annoyance categories between the inlab and crowdsourced experiments. Stimuli are ordered based on the AccAnn-MOS values (represented in the vertical axis) acquired from the InLab experiment. Acceptability & annoyance categories are color-coded. For each stimuli, the disagreement and agreement between experiments is marked with red and black lines, respectively.

value of 1.9804 is required for an acceptable video quality in the laboratory setting, whereas in the crowdsourcing study, this threshold is 2.1157. Videos start to be annoying at an ACR-MOS value of 3.4001 in laboratory settings, while 3.6543 ACR-MOS is required for a satisfactory experience in the crowdsourcing experiment. This suggests that subjects were more tolerant to decrease in video quality in the laboratory experiment. UVQ thresholds indicate a similar trend in annoyance thresholds, while an inverse effect is observed on the acceptability threshold. Although not within the scope of this study, this might be explained by the slightly lower accuracy of UVQ in the low-quality range.

6. CONCLUSION

Accurate assessment of acceptability & annoyance in video quality holds immense significance for service providers, especially on online social media platforms. It quantifies the variations that can be introduced into video encoding pipelines in different contexts, all while maintaining a consistently accepted Quality of Experience (QoE) for observers. The ultimate goal is to develop a context-neutral, objective QoE model that can be dynamically adapted based on specific contexts, including the streaming platform, display device, remaining battery, signal strength, and more. However, a notable challenge in achieving this objective is the increased need for participants in acceptability and annoyance experiments. To address this need, this paper focuses on exploring the feasibility of extending AccAnn experiments to crowdsourcing platforms.

The reliability of crowdsourcing in this context was examined from various aspects. Findings indicate a more linear relationship between video quality and QoE in the crowdsourcing setting. Additionally, subjects in crowdsourcing experiment displayed slightly higher expectations of video quality at the acceptability and annoyance thresholds. Comparison of acceptability & annoyance categories revealed that 23 out of 126 stimuli categorized differently between the two settings.

Our results imply that, in crowdsourcing settings, subjects exhibit a limited ability to adjust their expectations in response to instructions, unlike in controlled laboratory experiments. Despite the reduced ability of subjects in crowdsourcing setting, the impact remains minimal and predictable on acceptability & annoyance thresholds. These insights can contribute valuable information for the design of future acceptability & annoyance tests in crowdsourcing settings. As part of our future work, we aim to explore alternative methods for priming user expectations in acceptability & annoyance experiments.

7. REFERENCES

[1] Satu Jumisko-Pyykkö and Timo Utriainen, "A hybrid method for quality evaluation in the context of use for mobile (3d) television," *Multimedia Tools Appl.*, vol. 55, no. 2, pp. 185–225, nov 2011.

[2] Jing Li, Lukáš Krasula, Yoann Baveye, Zhi Li, and Patrick Le Callet, "Accann: A new subjective assessment methodology for measuring acceptability and annoyance of quality of experience," *IEEE Transactions on Multimedia*, vol. 21, no. 10, pp. 2589–2602, 2019.

[3] Ali Ak, Anne Flore Perrin, Denise Noyes, Ioannis Katsavounidis, and Patrick Le Callet, "Video consumption in context: Influence of data plan consumption on qoe," in *Proceedings of the 2023 ACM International Conference on Interactive Media Experiences*, New York, NY, USA, 2023, IMX '23, p. 320–324, Association for Computing Machinery.

[4] Anne Oeldorf-Hirsch, Jonathan Donner, and Ed Cutrell, "How bad is good enough? exploring mobile video quality trade-offs for bandwidth-constrained consumers," 10 2012.

[5] Toon De Pessemier, Katrien De Moor, Wout Joseph, Lieven De Marez, and Luc Martens, "Quantifying subjective quality evaluations for mobile video watching in a semi-living lab context," *IEEE Transactions on Broadcasting*, vol. 58, no. 4, pp. 580–589, 2012.

[6] Abhishek Goswami, Ali Ak, Wolf Hauser, Patrick Le Callet, and Frederic Dufaux, "Reliability of crowdsourcing for subjective quality evaluation of tone mapping operators," in *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*, 2021, pp. 1–6.

[7] Tobias Hoßfeld, Matthias Hirth, Judith Redi, Filippo Mazza, Pavel Korshunov, Babak Naderi, Michael Seufert, Bruno Gardlo, Sebastian Egger, and Christian Keimel, "Best Practices and Recommendations for Crowdsourced QoE—Lessons learned from the Qualinet Task Force "Crowdsourcing"," .

[8] B. Gardlo, S. Egger, M. Seufert, and R. Schatz, "Crowdsourcing 2.0: Enhancing execution speed and reliability of web-based QoE testing," in *2014 IEEE International Conference on Communications (ICC)*, 2014, pp. 1070–1075.

[9] Yilin Wang, Sasi Inguva, and Balu Adsumilli, "Youtube ugc dataset for video compression research," in *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*. Sept. 2019, IEEE.

[10] Ali Ak, Emin Zerman, Maurice Quach, Aladine Chetouani, Aljosa Smolic, Giuseppe Valenzise, and Patrick Le Callet, "Basics: Broad quality assessment of static point clouds in a compression scenario," *IEEE Transactions on Multimedia*, pp. 1–13, 2024.

[11] "Prolific," https://www.prolific.co/, Accessed: Feb 2024. [Online].

[12] "Amazon Mechanical Turk," https://www.mturk.com, Accessed: Oct 2020. [Online].

[13] Ali Ak and Patrick Le Callet, "IPI-VUGC: Acceptance/Annoyance and Video Quality of Vertically Oriented User Generated Videos," Jan. 2024.

[14] ITU-T, "Reference software for itu-t h.264 advanced video coding," ITU-T H.264-2, 2016.

[15] Satu Jumisko-Pyykkö and Miska M. Hannuksela, "Does context matter in quality evaluation of mobile television?," in *Proceedings of the 10th International Conference on Human Computer Interaction with Mobile Devices and Services*, New York, NY, USA, 2008, MobileHCI '08, p. 63–72, Association for Computing Machinery.

[16] ITU-R, "Methods for the subjective assessment of video quality, audio quality and audiovisual quality of internet video and distribution quality television in any environment," ITU-R Recommendation Recommendation P.913, 2021.

[17] Jingwen Zhu, Ali Ak, Patrick Le Callet, Sriram Sethuraman, and Kumar Rahul, "Zrec: Robust recovery of mean and percentile opinion scores," in *2023 IEEE International Conference on Image Processing (ICIP)*, 2023, pp. 2630–2634.

[18] Andréas Pastor, Pierre David, Ioannis Katsavounidis, Lukas Krasula, Hassene Tmar, and Patrick Le Callet, ""Discriminability-Experimental Cost" tradeoff in subjective video quality assessment of codec: DCR with EVP rating scale versus ACR-HR," working paper or preprint, Dec. 2023.

[19] Andréas Pastor and Patrick Le Callet, "Towards guidelines for subjective haptic quality assessment: A case study on quality assessment of compressed haptic signals," pp. 1667–1672, 2023.

[20] Zhi Li, Anne Aaron, Ioannis Katsavounidis, Anush Moorthy, and Megha Manohara, "Toward a practical perceptual video quality metric," *The Netflix Tech Blog*, 2016.

[21] Yilin Wang, Feng Yang, Balu Adsumilli, Neil Birkbeck, Joong Gon Yim, Junjie Ke, Hossein Talebi, Peyman Milanfar, Ross Wolf, Jayaprasanna Jayaraman, Carena Church, and Jessie Lin, "Uvq: Measuring youtube's perceptual video quality," *The Google Research Blog*, 2022.