Subjective Test Environments: A Multifaceted Examination of Their Impact on Test Results

Jingwen Zhu, Ali Ak, Charles Dormeval, Patrick Le Callet, Rahul Kumar, Sriram Sethuraman

Subjective Test Environments: A Multifaceted Examination of Their Impact on Test Results

Jingwen Zhu*
Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004
Nantes, France
jingwen.zhu@etu.univ-nantes.fr

Ali AK*
Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004
Nantes, France
ali.ak@univ-nantes.fr

Charles Dormeval
Nantes Université, CAPACITÉS SAS
Nantes, France
charles.dormeval@univ-nantes.fr

Patrick Le Callet
Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004
Nantes, France
patrick.lecallet@univ-nantes.fr

Kumar Rahul
Amazon Prime Video
Seattle, USA
krhmz@amazon.com

Sriram Sethuraman
Amazon Prime Video
Bangalore, India
sssethur@amazon.com

ABSTRACT

Quality of Experience (QoE) in video streaming scenarios is significantly affected by the viewing environment and display device. Understanding and measuring the impact of these settings on QoE can help develop viewing environment-aware metrics and improve the efficiency of video streaming services. In this ongoing work, we conducted a subjective study in both laboratory and home settings using the same content and design to measure QoE in Degradation Category Rating (DCR). We first analyzed subject inconsistency and confidence intervals of the Mean Opinion Scores (MOS) between the two settings. We then used statistical models such as ANOVA and t-test to analyze the differences in subjective tests on video quality between the two viewing environments. Additionally, we employed the Eliminated-By-Aspects (EBA) model to quantify the influence of different settings on the measured QoE. We conclude with several research questions that could be further explored to better understand the impact of the viewing environment on QoE.

CCS CONCEPTS

• **Applied computing**; • **Human-centered computing**;

KEYWORDS

Subjective test, HD, Video Quality Assessment

ACM Reference Format:
Jingwen Zhu, Ali AK, Charles Dormeval, Patrick Le Callet, Kumar Rahul, and Sriram Sethuraman. 2023. Subjective Test Environments: A Multifaceted Examination of Their Impact on Test Results. In *ACM International Conference on Interactive Media Experiences (IMX '23), June 12–15, 2023, Nantes, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3573381.3596470

*Both authors contributed equally to this research.

IMX '23, June 12–15, 2023, Nantes, France

ACM ISBN 979-8-4007-0028-6/23/06.
https://doi.org/10.1145/3573381.3596470

1 INTRODUCTION

Previous studies have identified several factors that can significantly impact Quality of Experience (QoE) for multimedia content. These factors include, but are not limited to, video quality, device, observer's emotion *etc.*

The physical experiment environment is a major factor that influences QoE, and standardization efforts have been made to propose methodologies and recommendations for experiment conditions in subjective QoE studies. For instance, ITU has developed standards such as BT.500 [3], BT.910 [5], and BT.913 [4]. In a previous study by Jumisko-Pyykö *et al.* [8], it was found that the acceptability of content varied significantly when measured in laboratory environment compared to real-life scenarios. The results indicated that subjects were more critical during laboratory experiments when evaluating the acceptability of mobile videos.

In another study, Li *et al.* explored the impact of the influence of devices on QoE (Acceptability/Annoyance) for video streaming [9]. The results of their analysis showed that the devices in comparison (Tablet vs TV) has a significant impact on the measured QoE in terms of acceptability and annoyance. The authors relied on observer uncertainty and EBA to analyze and quantify the influence.

In this ongoing work, we explore the impact of experiment environments (laboratory vs home settings) on the collected MOS for high resolution videos. We conduct the same subjective experiment on quality evaluation of compressed videos in these two environments, namely "InLab" and "AtHome". We analyze the results in terms of the correlation between the collected MOS as well as the comparison of confidence intervals. Furthermore, we conduct an ANOVA test to determine the statistical impact of the experiment environments. Finally, we conducted an advanced analysis, namely Eliminated-by-Aspects (EBA) to measure the influence as a function of MOS.

2 SUBJECTIVE EXPERIMENTS

Two subjective experiments were conducted for this study with identical contents and experiment design. The only difference between the two experiments was the experiment setting where one of the experiments was conducted in a controlled laboratory environment and the other was in home environment of each participants.

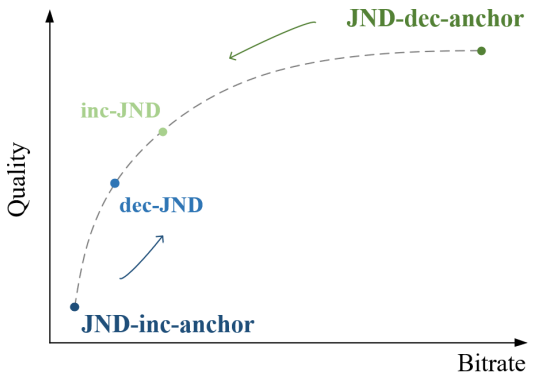


Figure 1: 2-direction JND search

2.1 Content

Content selection plays a crucial role in enhancing the efficiency of subjective tests by enabling the selection of representative contents [11]. Spatial information (SI), Temporal information (TI)[2] and Ambiguity[10] are computed on 229 HD (1080p) videos contents provided by Amazon Prime Video. K-means clustering was applied to the extracted features to group the contents based on their similarities. From these content clusters, 10 HD contents C_i were selected, with each content having a duration of 10 seconds.

Each selected contents are compressed by High Efficiency Video Coding (HEVC) with 3 different encoding resolutions (1080p, 720p and 540p). For each encoding resolutions, 13 different level of distortions are used. There are in total 39 (3×13) encoding recipes $R_{i,j}$ for each content C_i. However, it is time and money consuming to conduct subjective test on all the generated PVS (Processed Video Sequence). To select significant PVS for the subsequent subjective test, we conducted a 2-direction JND search by experts(golden-eyes) as described in Fig.1, following the framework proposed by Zhu et al. [17]. For each resolution, we selected the PVS with the highest quality as the anchor (referred to as JND-dec-anchor) and searched through the remaining PVS to identify the point at which the observer just begins to perceive a difference in quality between the anchor and the PVS (referred to as inc-JND points). Likewise, we selected the PVS with the lowest quality as the JND-inc-anchor to identify the JND points where the observer first perceives a quality difference. To determine these JND points, we used a binary search algorithm [13]. There are a total of 12 stimulus for each content, consisting of one source (SRC) and 11 PVS.

2.2 Experiment Design

We conducted the subjective experiment using the degradation category rating (DCR) test methodology both at home and in the lab. The DCR test involved presenting the test sequences in pairs, where the first stimulus in each pair was always the source reference, and the second stimulus was the processed video stimuli (PVS) of the same content. Five-level scale for rating was used as recommended in ITU-T P.910 [2]. The stimuli were presented to the subjects in a random order, with a constraint that the same content was not presented successively. The entire test took approximately 45 minutes to complete.

2.3 Experiment Settings

2.3.1 *InLab Experiment.*

- Display: a 4K calibrated UHD Grundig Finearts 55 FLX 9492 SL with a 55-inch screen size.
- Viewing distance: 3H for HD (1080p) video, with H the height of the screened video, as recommended in ITU-R BT.1769 [6]
- Ambient light: the illuminance level of the subjective environment was set as recommended by ITU-R BT.2013-1 [7]
- Subjects: A total of 24 participants, who were non-experts in subjective experiments, image processing, or related fields, took part in the study. All participants had either normal or corrected-to-normal visual acuity, which was ensured prior to the experiment using a Monoyer chart. Ishihara color plates were used to test color vision, and all viewers passed the pre-experiment vision check.

2.3.2 *AtHome Experiment.*

- Displays: 10 display of 55-inch are used at 10 different home environment:
  - SONY KD-55XH8094 x4
  - SAMSUNG QE55Q74TATXXC x2
  - SAMSUNG UE55TU8075U x2
  - LG nanocell 55NAN091 x2
- Viewing distance: instructions are given to participants, in which they are asked to watch the screen 2 meters away.
- Ambient light: the instructions suggest that participants close the curtains/blinds and turn off any lights that directly face the screens. Additionally, participants should turn on some lights in the room to create a dimly lit atmosphere while avoiding complete darkness.
- Subjects: 10 screens are set in 10 different home, each home have 2 participant, who have passed the pre-experiment vision check to make sure that they have normal or corrected-to normal visual acuity.

PVS with various encoding resolutions where up-scaled by the video player to match the resolution of the source content (i.e., viewing resolution).

3 ANALYSIS

In this section, we present the preliminary results of our analysis of measuring the impact of experiment settings on QoE.

3.1 MOS and CI of the MOS Comparison

In this analysis, we analyze the correlation between the MOS collected from the two experiments and compare their Confidence Intervals (CI). We expect a smaller CI for the InLab experiment. We use 3 different MOS Recover (observer screening) methodologies to analyze the correlation between the MOS collected in InLab and AtHome experiments. Summary of each method can be found below:

BT500: ITU-R BT.500 Recommendation [3] defines the simple and commonly used observer screening procedure. Subjects are

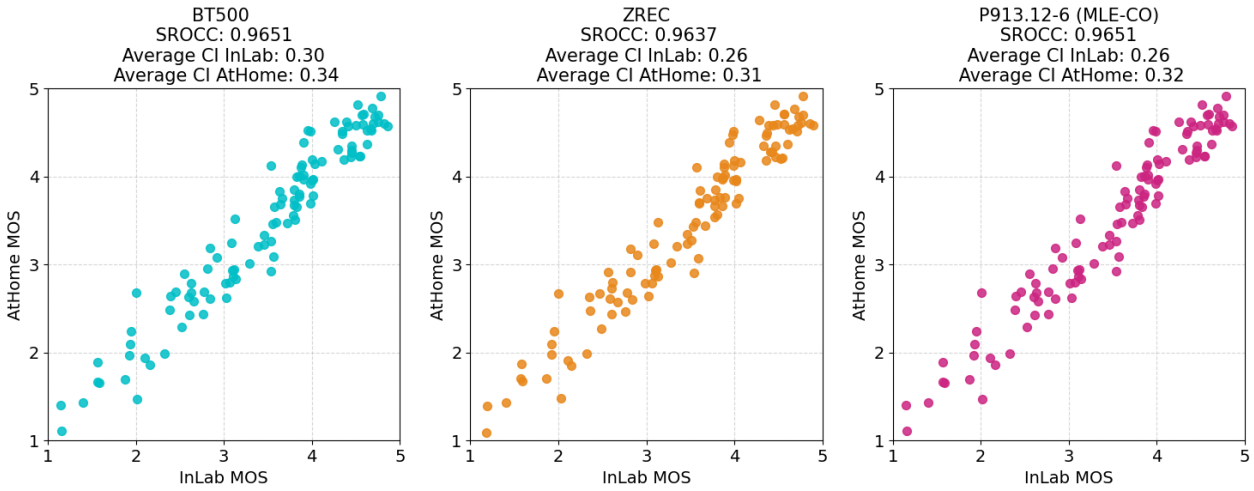


Figure 2: Scatter plots between InLab MOS and AtHome MOS with 3 different MOS recovery (observer screening) methods. Below the title of each plot, Spearman's Rank Order Correlation Coefficients (SROCC) and average CI 95% for InLab and AtHome experiments are given.

rejected based on the number of opinion scores outside of the predefined amount of standard deviation range of the population. If a subject found to be an outlier, all of his/her opinions are removed from the dataset. MOS is calculated as the mean of remaining subjects.

ZREC: is proposed in [16] by Zhu et al. and relies on estimating subject bias and inconsistency to recover the MOS. It doesn't require any solver and the evaluations show a smaller CI over the tested dataset compared to alternative methods.

P913-12.6 (MLE-CO): is introduced in ITU-R P.913 Recommendation clause 12.6 [4] and defines a procedure where MOS is recovered by bias removal and subject inconsistency based weighted average. The procedure defines the individual opinion scores of a subject as the combination of subject bias, inconsistency and the true quality of the stimuli and jointly solves these three parameters.

With each method describe above, MOS from the InLab and AtHome experiments and their CI (95%) are calculated. Figure 2 presents the results as a scatter plot between the InLab and AtHome MOS for each method. SROCC values indicate that the MOS acquired from experiments are highly correlated with all MOS recovery methods. On another front, we observe slightly lower CI for the InLab MOS compared to AtHome MOS. Considering the uncontrolled experiment environments in AtHome experiment, the results are not surprising. With more sophisticated MOS recovery methods, we can acquire lower CIs for both experiments however the slightly higher CI for AtHome experiment remains true.

3.2 ANOVA

We conducted an Analysis of Variance (ANOVA)[1] and a Student's t-test[15] on the 110 original ratings collected from both the InLab and AtHome experiments. The results are presented in Table 1. The p-value obtained from both tests was 0.3880, which indicates that the InLab and AtHome test environments did not have a significant impact on the subjective experiment results. It is worth noting that 24 participants took part in the InLab experiment. However, to ensure the validity of the results, we performed ANOVA and t-tests on the data obtained from 20 observers among them, after excluding the 4 participants who had the highest bias and inconsistency values calculated from the ZREC.

ANOVA		Student t-test	
F-statistic	0.7455	T-statistic	0.8634
p-value	0.3880	p-value	0.3880

Table 1: Results of ANOVA and Student t-test on InLab and AtHome experiment

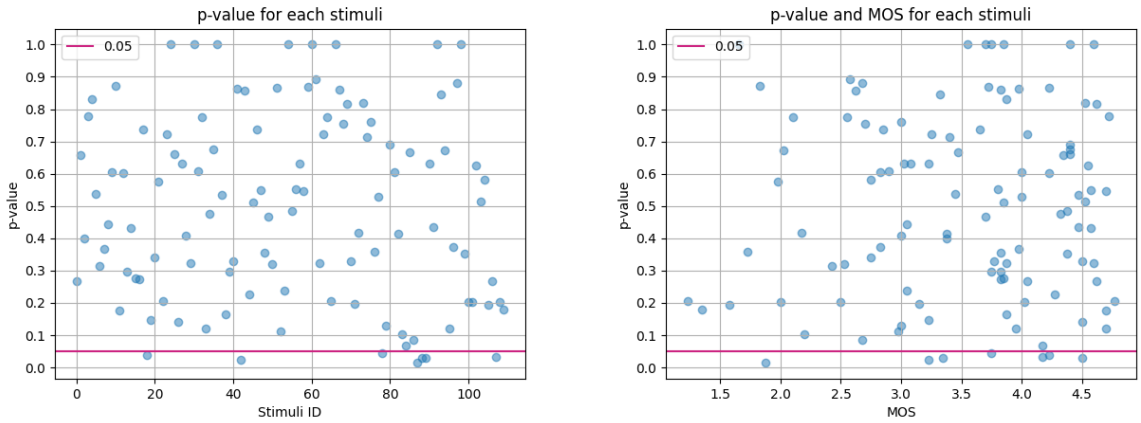
We also conducted ANOVA on a per-stimulus basis, and the results are presented in Fig. 3. As shown in Fig. 3 (a), out of the 110 stimuli, only 7 exhibited a p-value smaller than 0.05. This indicates that for the majority of stimuli, the InLab and AtHome environments did not have a significant impact on the subjective experiment results in terms of quality. In Fig. 3 (b), we plotted the MOS scores and p-values obtained by ANOVA for each stimulus to examine the MOS distributions of the stimuli that exhibited significant differences.

3.3 Eliminated-by-Aspects

In this part, we rely on EBA analysis the impact of experiment settings on the QoE and quantify it for various quality ranges.

EBA model[12] is used to analyze the subgroups containing same stimuli. It assumes that, in a subjective experiment, a subject chose a stimulus better than another stimulus in comparison due to set of attributes being present in the higher quality stimulus.

In QoE experiments, each video sequence i has a quality attribute defined by $u(q_i)$. If no other influence is considered, the



(a) (b)

Figure 3: Illustration of the p-value of ANOVA per stimuli (a) and the relationship between p-value and MOS for each stimuli (b)

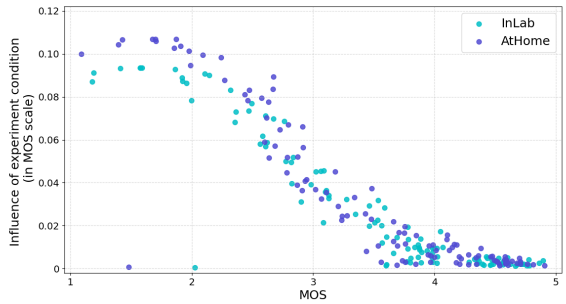


Figure 4: The influence of experiment environments $u(d_i)$ in MOS range plotted against the MOS of the stimuli. Horizontal axis represents the MOS whereas the vertical axis represents the Q_{di} calculated as in Equation (1).

MOS can be represented as $\log(u(q_i))$. In our case where experiment were repeated under home and laboratory settings, these settings also have their attributes defined as $u(d_{Home})$ and $u(d_{Lab})$. Therefore, the measured QoE in each experiment can be represented as $\log(u(q_i)+u(d_i))$ where d_i is either $u(d_{Home})$ or $u(d_{Lab})$ depending on the experiment.

By using the matlab function proposed in [14], we can solve the maximum likelihood estimates (MLE) of our EBA model for parameters $u(q_i)$, $u(d_{Home})$, and $u(d_{Lab})$.

After solving the MLE, we can obtain the influence of experiment settings with varying quality range as follows:

$$Q_i = \log(u(q_i) + u(d_i)) - \log(u_{q_i}) \quad (1)$$

By following the equation above, we can than analyze the impact of experiment condition over all stimuli. Figure 4 presents the results as a scatter plot between the MOS values of each stimulus and the influence of experiment condition in the MOS scale. It can be observed that the influence of AtHome experiment environments is slightly higher for stimuli with lower MOS. Note that the overall impact of the experiment environments are estimated to be low. This further confirms our observations in the previous analyses.

4 CONCLUSION AND FUTURE WORK

The results of our preliminary experiments can be summarized as follows: 1) The video quality scores annotated in the two environments (InLab and AtHome) are highly correlated, but the confidence interval (CI) of the mean opinion scores (MOS) of AtHome is higher than that of InLab. 2) The analysis of variance (ANOVA) and t-test indicate that there is no significant difference in video quality between the two viewing environments. 3) The Eliminated-by-Aspects (EBA) model reveals that the impact of the InLab and AtHome environments on the video quality is small, which is consistent with the ANOVA results. Additionally, we observed that people tend to be slightly more critical of low-quality videos in the InLab environment compared to the AtHome environment. These conclusions help bridge the gap between the subjective experimental environment and the realistic viewing environment for end-users.

The subjective experiment can be utilized to explore the impact of various devices on the quality of experience for HD videos in future research. Furthermore, these results can be leveraged to develop device-agnostic objective quality metrics or frameworks that can be adapted to different devices and viewing conditions to enhance the accuracy of objective quality assessment.

REFERENCES

[1] Ronald Aylmer Fisher. 1992. *Statistical methods for research workers.* Springer.
[2] ITU. 2008. Subjective video quality assessment methods for multimedia applications.
[3] ITU-R. 2019. Methodology for the Subjective Assessment of the Quality of Television Pictures. ITU-R Recommendation BT.500-14.
[4] ITU-R. 2021. 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.
[5] ITU-R. 2022. Subjective video quality assessment methods for multimedia applications. ITU-R Recommendation Recommendation P.910.
[6] ITU-R BT.1769. 2006. Parameter values for an expanded hierarchy of LSDI image formats for production and international programme exchange. *Int'l*

Telecommunication Union (2006).

[7] ITU-R BT.2013-1. 2013. A reference viewing environment for evaluation of HDTV program material or completed programmes. *Int'l Telecommunication Union* (2013).

[8] Satu Jumisko-Pyykkö and Miska M. Hannuksela. 2008. Does Context Matter in Quality Evaluation of Mobile Television? *(MobileHCI '08)*. Association for Computing Machinery, New York, NY, USA, 63–72. https://doi.org/10.1145/1409240.1409248

[9] Jing Li, Lukáš Krasula, Patrick Le Callet, Zhi Li, and Yoann Baveye. 2018. Quantifying the Influence of Devices on Quality of Experience for Video Streaming. In *2018 Picture Coding Symposium (PCS)*. 308–312. https://doi.org/10.1109/PCS.2018.8456304

[10] Suiyi Ling, Yoann Baveye, Deepthi Nandakumar, Sriram Sethuraman, and Patrick Le Callet. 2020. Towards better quality assessment of high-quality videos. In *Proceedings of the 1st Workshop on Quality of Experience (QoE) in Visual Multimedia Applications*. 3–9.

[11] Margaret H. Pinson, Marcus Barkowsky, and Patrick Le Callet. 2013. Selecting scenes for 2D and 3D subjective video quality tests. *EURASIP Journal on Image and Video Processing* 2013, 1 (Aug. 2013), 50–61.

[12] Amos Tversky. 1972. Elimination by Aspects: A Theory of Choice. *Psychological Review* 79, 4 (1972), 281–299. https://doi.org/10.1037/h0032955

[13] Haiqiang Wang, Ioannis Katsavounidis, Jiantong Zhou, Jeonghoon Park, Shawmin Lei, Xin Zhou, Man-On Pun, Xin Jin, Ronggang Wang, Xu Wang, et al. 2017. VideoSet: A large-scale compressed video quality dataset based on JND measurement. *Journal of Visual Communication and Image Representation* 46 (2017), 292–302.

[14] Florian Wickelmaier and Christian Schmid. 2004. A Matlab function to estimate choice model parameters from paired-comparison data. *Behavior Research Methods, Instruments, & Computers* 36 (2004), 29–40.

[15] S William. 1908. The probable error of a mean. *Biometrika* 6, 1 (1908), 1–25.

[16] Jingwen Zhu, Ali Ak, Patrick Le Callet, Sriram Sethuraman, and Kumar Rahul. 2023. ZREC : robust recovery of mean and percentile opinion scores. (March 2023). https://hal.science/hal-04017583 working paper or preprint.

[17] Jingwen Zhu, Suiyi Ling, Yoann Baveye, and Patrick Le Callet. 2022. A framework to map vmaf with the probability of just noticeable difference between video encoding recipes. In *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*. IEEE, 1–5.