Video Consumption in Context: Influence of Data Plan Consumption on QoE

Ali Ak, Anne Flore Perrin, Denise Noyes, Ioannis Katsavounidis, Patrick Le Callet

▶ **To cite this version:**

Ali Ak, Anne Flore Perrin, Denise Noyes, Ioannis Katsavounidis, Patrick Le Callet. Video Consumption in Context: Influence of Data Plan Consumption on QoE. IMX '23: ACM International Conference on Interactive Media Experiences, Jun 2023, Nantes, France. pp.320-324, 10.1145/3573381.3596474 . hal-04264748

HAL Id: hal-04264748
https://hal.science/hal-04264748v1

Submitted on 31 Oct 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Video Consumption in Context: Influence of Data Plan Consumption on QoE

ALI AK, Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, France
ANNE FLORE PERRIN, Nantes Université, École Centrale Nantes, CAPACITÉS SAS, France
DENISE NOYES, Meta, USA
IOANNIS KATSAVOUNIDIS, Meta, USA
PATRICK LE CALLET, Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, France

User expectations are one of the main factors on providing satisfactory QoE for streaming service providers. Measuring acceptability and annoyance of video content, therefore, provide a valuable insight when measured under a given context. In this ongoing work, we measure video QoE in terms of acceptability and annoyance for the remaining data in a mobile data plan context.. We show that simple logos can be used during the experiment to prompt the context to subjects and the different context levels may impact the user expectations and consequently their satisfactions. Finally, we show that objective metrics can be used to determine the acceptability and annoyance thresholds for a given context.

CCS Concepts: • **Human-centered computing**;

Additional Key Words and Phrases: "Acceptability and Annoyance; Quality of Experience; Eliminated-By-Aspects"

ACM Reference Format:
Ali Ak, Anne Flore Perrin, Denise Noyes, Ioannis Katsavounidis, and Patrick Le Callet. 2023. Video Consumption in Context: Influence of Data Plan Consumption on QoE. In *ACM International Conference on Interactive Media Experiences (IMX '23), June 12–15, 2023, Nantes, France.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3573381.3596474

1 INTRODUCTION

Quality of Experience (QoE) in video streaming defines the observer's level of satisfaction and how well their expectations are met while viewing the video content. Several factors may impact the QoE and they are categorized as system, context, and human in the Qualinet white paper[7]. Along with the essential video quality, other vital considerations in measuring QoE include but are not limited to fidelity, cost, ecological impact, and display device specifications.

Recent advancements provide a plethora of metrics and methodologies to assess QoE for video content. However, from the point of view of the streaming service provider [15], knowing the exact quality of the video content is not always priority and might not be enough to understand whether the delivered content satisfies the user expectations. In this regard, Acceptability and Annoyance (AccAnn) scale has been introduced and frequently used in recent years [2, 6, 8, 12]. AccAnn scale often contains three categories as "not acceptable (1)", "acceptable but annoying (2)", and "not annoying (3)". Mapping of quality measurements from Absolute Category Rating (ACR) or Degradation Category Rating (DCR) scale to AccAnn scale has also been investigated[3, 8, 11].

AccAnn ratings are classically collected via a multi-step evaluation procedure. After the stimulus is presented to subjects, a first evaluation screen is prompted, asking whether the stimulus is acceptable or not. If the subject answers no, the stimulus is rated with lowest rating as "not acceptable" and if the subject answers yes a second question is

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
© 2023 Copyright held by the owner/author(s).
Manuscript submitted to ACM

prompted asking whether the stimulus is annoying or not. If subject answers the second question as yes, the stimulus is rated as "acceptable but annoying" and if the answer is no the stimulus is rated with highest quality as "not annoying". Li et al. proposed a a single-step procedure [8] where the three ratings are presented at once with color coding to guide the subjects.

Among others, context plays an important role in satisfaction of the user expectations[5]. Depending on the use-case, context can contain various factors such as subscription level of a streaming service (premium plan vs basic plan), signal strength of the display device (4g vs 3g), remaining power of the device (low vs full battery), etc. Previous studies take context into account for acceptability and annoyance by either collecting pre/post-experiment surveys [5] or setting the context for subjects with set of instructions prior to the experiment[8].

Many mobile phone users subscribe to monthly data plans that allow them a certain volume to be used over a month, for example, 10GBytes, while others buy data that can be used over 24 hours, or 1-week. As such, users are aware of the concept of "how much data I have left in my quota" and tent to adjust their usage so as to maximize that, without exceeding it. In this ongoing work, we focus on remaining data plan context and its influence on the user expectations and acceptance and annoyance of the video content.

2 SUBJECTIVE EXPERIMENTS

Several subjective study were conducted as part of this work. Initially, an ACR experiment was conducted to collect Mean Opinion Scores (MOS) and afterwards three AccAnn experiments with the same context and three context levels were conducted. Details for each experiment are given below.

2.1 Content

3 source video (SRC) with 1080p resolution and horizontal orientation were used in the experiment. Each source video is 5 seconds long with 30 fps. To generate the processed video sequences (PVS), each SRC was compressed with VP9 coding algorithm at 6 different levels. Generated PVS with SRC were used in all experiments.

2.2 ACR Experiment

In order to collect the mean opinion scores (refered as ACR-MOS for the rest of the paper), an absolute category rating with hidden reference (ACR-HR) experiment was conducted on a 5 category quality scale. The experiment was conducted in Nantes University IPI laboratories with 25 subjects from the in-house participant panel. All subjects were checked for visual acuity and compensated for their participation. The experiment room was set according to ITU recommendations[4].

ACR experiment was conducted without any context provided to observers in order to collect traditional MOS values corresponding to each video. This will allow us to analyze the relation between MOS values and the acceptability annoyance MOS (AccAnnMOS) collected in the AccAnn experiments.

2.3 AccAnn Experiments

We followed the experiment design proposed in [8] as color coded acceptability annoyance test with a single-step evaluation procedure. The evaluation screen with three scales is shown in Figure 1-a which is prompted to subjects after each video stimulus. For the analysis, we also use the numerical representations of the AccAnn results as 1, 2, 3 for "Not Acceptable", "Annoying but Acceptable" and "Not Annoying" respectively. Similar to the ACR experiment, the

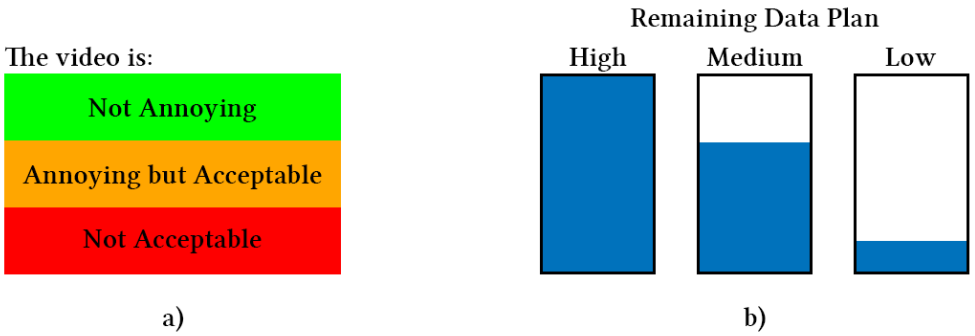

Fig. 1. (a) AccAnn experiment color coded scale as shown in the evaluation screen, and (b) presentation of different levels of the remaining data quota.

experiment room was arranged according to ITU Recommendations[4]. The experiments were conducted in Nantes University IPI laboratories with 60 subjects from the in-house participant panel.

In order to understand the impact of context on the acceptability and annoyance of the video stimuli, AccAnn experiment was conducted in three sessions where in each session subjects were assigned by a profile. The context used in the experiment was remaining data and in each session a different context level was used. Prior and during the experiment, the logos shown in Figure 1-b were used to provide the context to subjects. By collecting AccAnn-MOS with different context levels, we seek to reveal the change in expectations of the subjects.

3 MAPPING OF ACCANN-MOS TO ACR-MOS

Initially, AccAnn-MOS values to ACR-MOS values were mapped to understand the acceptability and annoyance thresholds. For each context level ("Low", "Medium", "High" data quota levels), we plotted AccAnn-MOS values against the ACR-MOS values in Figure 2. Furthermore, a 4 parameters logarithmic function was fitted. We can observe that the AccAnn-MOS values for the "Low" are particularly higher for the same content in low to mid quality range.

The thresholds where contents start to be unacceptable are 1.86, 2.21, 2.25 in ACR-MOS scale for "Low", "Medium" and "High" context levels, respectively. This indicates that the subjects have a lower expectation in terms of quality for an acceptable content when they have "Low" amount of data quota in their data plans. Meanwhile, we observe no statistically significant difference between "Medium" and "High" context levels.

On another front, the thresholds where contents start to be annoying are calculated as 3.38, 3.56, 3.59 in ACR-MOS scale for "Low", "Medium" and "High" context levels, respectively. It can be seen that the differences between context levels in annoyance thresholds are not as pronounced as unacceptability threshold.

4 EBA ANALYSIS

In this section, we use the EBA model explained in Section 4.1 to further investigate our initial observations (see Section 3) by comparing the AccAnn-MOS values to ACR-MOS values with different context levels. By utilizing EBA model, we can quantify the influence of the context level as a function of measured QoE. Results of this analysis are presented in Section 4.2.

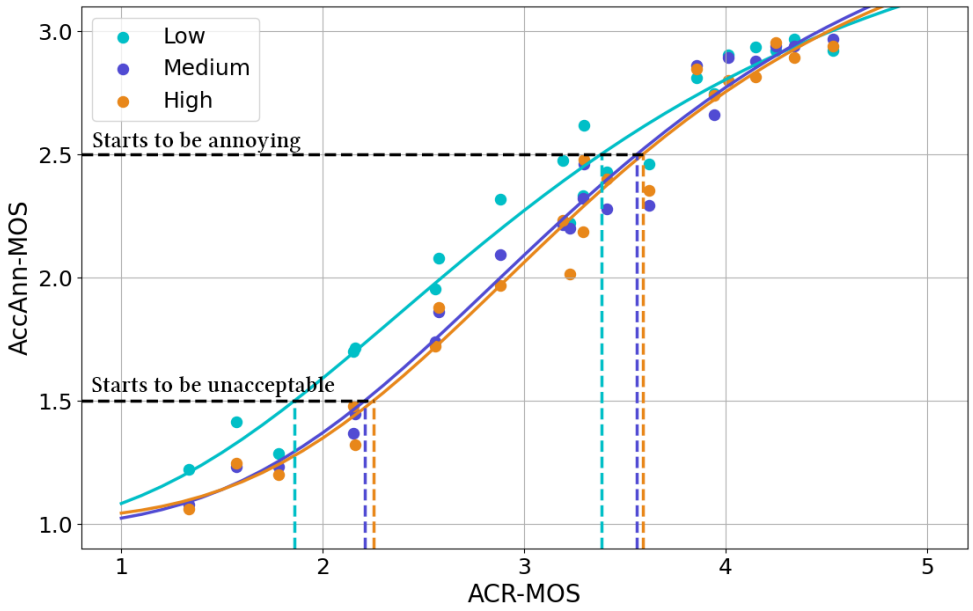

Fig. 2. Mapping of AccAnn-MOS to ACR-MOS for the three AccAnn experiment sessions with different context levels. For each context level, a 4 parameters logarithmic function is fitted. Thresholds where the stimuli start to be annoying (below 2.5 AccAnn-MOS) and start to be unacceptable (below 1.5 AccAnn-MOS) are shown with dashed lines for each context level.

4.1 Model

Tversky has proposed a set of models to study and analyze pairwise comparison data [13] as a generalization to the Bradley-Terry-Luce (BTL) model. According to EBA model, a subject prefers certain stimulus over an alternative due to presence of set of attributes one has over the other. In EBA model, stimuli may contain several attributes and all of them can impact the choice of the subject. On the other hand, BTL, is a specific case of EBA where each stimulus is defined by a unique attribute.

In QoE domain, EBA model can be used to study the effect of parameters that ultimately define the measured quality in a subjective test. By conducting a set of subjective studies with same stimuli and varying context levels (e.g. Low, Medium, High Data Quota), we can define the set of attributes that affects each measurement (AccAnn opinion scores) as the visual quality of each stimulus and the context level. In a previous study [9], Li et al. studied the influence of subscription levels and display device on the QoE for video streaming services.

Formally, we can define the i_{th} video sequence with its visual quality attribute defined as $u(q_i)$ and its visual quality as the logarithmic of the attribute, *i.e.,* $\log(u(q_i))$. Furthermore, we can define the attributes of each context level as $u(d_L)$, $u(d_M)$ and $u(d_H)$ for "Low", "Medium" and "High" data quota, respectively. Consequently, we can represent the measured QoE *(i.e.,* AccAnnMOS) in each AccAnn experiment as $\log(u(q_i) + u(d_i))$ where d_i is either $u(d_L)$, $u(d_M)$ or $u(d_H)$ depending on the context level. Therefore, following the EBA model, the probability that a subject prefers video i over video j based on the influence of context level and the visual quality of the videos can be defined as:

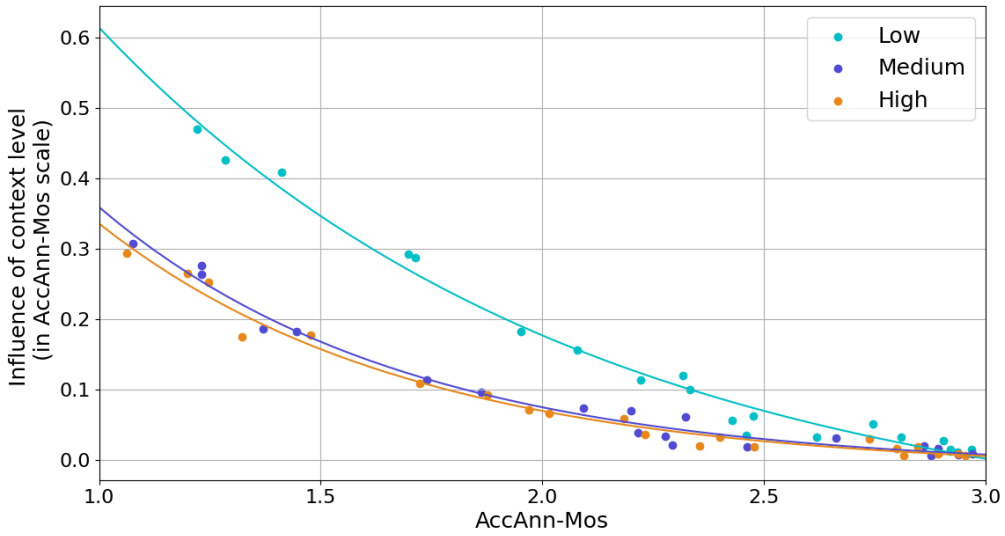

Fig. 3. Plot of influence of context level Q_{di} in AccAnn-MOS range and the measured AccAnn-MOS of each corresponding context level. Horizontal axis represents the AccAnn-MOS and the vertical axis represents the Q_{di} calculated as Equation 3.

$$P_{ij} = \frac{u(q_i) + u(d_i)}{u(q_i) + u(d_i) + u(q_j) + u(d_j)} \tag{1}$$

We rely on the Matlab implementation of EBA model proposed in [14]. In order to so, AccAnn results need to be represented as a pairwise comparison matrix (PCM) M_{ij}. i and j are in the range [1, 63], since we have 21 stimuli in the dataset and each observed with three context levels ("Low", "Medium", "High"). When converting AccAnn results into a PCM, we assign M_{ij} with 1 if the video i has higher rating than the video j, and 0 for the opposite case. In the cases where the AccAnn scores are equal, we randomly sample 0 or 1 from a uniform distribution and assign to M_{ij}.

Finally, we can calculate the maximum likelihood estimates of the EBA model attributes ($u(qi)$ and $u(d_i)$) by the following likelihood function:

$$L = \prod_{i<j} p_{ij}^{M_{ij}} (1 - p_{ij})^{M_{ji}} \tag{2}$$

where the p_{ij} is calculated as in Equation 1 and M_{ij} are the corresponding entries of the PCM.

4.2 Results

After solving the EBA model described above, we can obtain the parameters $u(qi)$ and $u(d_i)$. As described, we can acquire the visual quality of each of the 21 video sequences by $log(u_{q_{[i]}})$ and its measured QoE with the influence of the context as $log(u_{q_{[i]}} + u(d_i))$. Then, we can obtain the influence of context levels (defined as Q_{d_i}) for each video (i) with the following equation:

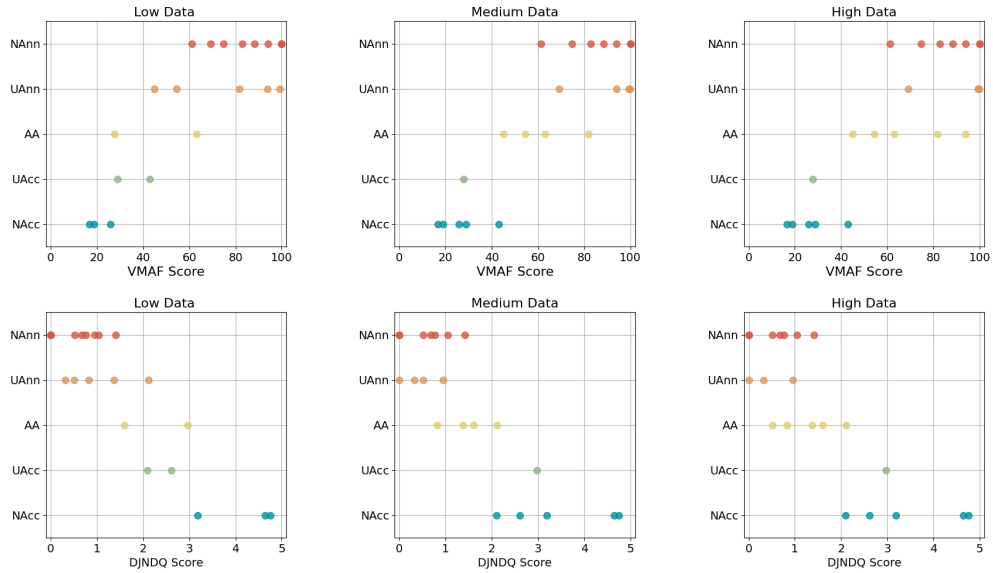

Fig. 4. Metric (VMAF and DJNDQ) score predictions for each content in the dataset. Content are ordered based on their categories indicated on the vertical axis of each plot(also color-coded). Due to categorical differences of the stimuli between the experiment, each experiment is plotted seperately.

$$Q_{d_i} = log(u_{q_i} + u(d_i)) - log(u_{q_i}) \tag{3}$$

Figure 3 presents the results of the EBA analysis. We observe a greater influence to the acceptability of the "Low" context level (*i.e.,* Low remaining data). On the other hand, we don't observe a significant difference between "Medium" and "High" context levels (*i.e.,* Medium and High remaining data). Furthermore, we observe that the impact is much greater in low quality range. These results verify our observations and conclusions derived in Section 3.

5 ACCEPTABILITY ANNOYANCE THRESHOLDS OF OBJECTIVE QUALITY METRICS

In this section, we use the following abbreviations "NAnn", "UAnn", "AA", "UAcc", "NAcc" for "Not Annoying", "Unsure about Annoyance but sure about the Acceptability", "Annoying but Acceptable", "Unsure about Acceptability but sure about the Annoyance", "Not Acceptable", respectively. The 21 PVSs in each experiment were assigned to one of these categories as proposed in Algorithm 2 in [8]. Differently from the proposed approach, we rely on Barnard's exact test instead of Fisher's exact test. "UAnn" category can be seen as the threshold for the videos to start to be annoying whereas "UAcc" is the threshold for the videos to start to be unacceptable.

After assigning each video in each experiment to one of these categories, we can analyze the objective quality metric predictions for different categories. Note that the quality metric predictions don't vary between the experiments since the content are the same and the only difference is the provided context level ("Low", "Medium" and "High" remaining Data). Two metrics (VMAF [10], DJNDQ [1]) were selected for this analysis and the results are presented in Figure 4. VMAF scores range from 0 to 100 with higher numerical values indicating a higher quality, whereas DJNDQ ranges from 0 to 5 (for this dataset, otherwise no theoretical upper bound) with lower numerical values indicating a higher quality. Note that, none of these metrics are designed to measure Acceptability and Annoyance. Although, by exploring

Table 1. Metric thresholds for Acceptability of Annoyance as the average of the predicted scores of all content in corresponding categories estimated in each metric's own scale.

	Low Data		Medium Data		High Data	
	UAcc	UAnn	UAcc	UAnn	UAcc	UAnn
VMAF	35.88	74.85	27.72	90.51	27.72	89.38
DJNDQ	2.35	1.03	2.97	0.45	2.97	0.42

the metric predictions, we can estimate a numerical threshold in metric score range to determine the Acceptability and Annoyance of a video content for a given context. It can be observed from the figure that metrics provide a relatively good distinction between the three main categories ("NAnn", "AA", "NAcc"), while showing difficulties in identifying content on the thresholds ("UAnn", "UAcc").

Moreover, we can determine a metric threshold by simply averaging the metric score predictions of the content in the thresholds ("UAnn", "UAcc"). Table 1 presents the result of this analysis. Note that the results may not be generalized due to low number of content available in the experiment. For each context, we can estimate a threshold in metric score range where the content start to become annoying (UAcc) and where content start to become unacceptable (UAnn). In accordance with the EBA analysis results, we don't see a significant difference in metric score thresholds ("UAnn", "UAcc") for "Medium" and "High" remaining Data scenarios. Again, similar to EBA Analysis results, the subjects' expectations in quality for acceptability and annoyance are lower for metric scores in "Low" remaining Data scenario

6 CONCLUSION AND PERSPECTIVE

In this ongoing work, we provided an analysis on the influence of context over the Acceptability and Annoyance of video content. We showed that the subjects are capable to incorporate the provided context in their evaluation. Moreover, we conducted a set of detailed analysis to quantify the influence of context on the subject expectations and consequently the acceptability and annoyance of the video content. We showed that when the QoE of the video content is low, the context has a higher impact on the user expectations and consequently their satisfaction. On another front, we provided preliminary results on the metric thresholds for predicting acceptability and annoyance of a video content.

Although due to low number of samples in the experiments the results might not be generalized reliably, we still believe that these results provide crucial insights for the streaming service providers. We believe that, this work can lead to new avenues to explore for streaming service providers in regard the satisfying user expectations. Streaming service providers can utilize the preliminary findings and the analysis scheme to adjust their video encoding recipes to provide a similar satisfaction with lower bandwidth in certain contexts. Moreover, we believe that the preliminary results may inspire and lead the community to explore different context and their impact on the user satisfaction.

REFERENCES

[1] Ali Ak, Andreas Pastor, and Patrick Le Callet. 2022. From Just Noticeable Differences to Image Quality. In *Proceedings of the 2nd Workshop on Quality of Experience in Visual Multimedia Applications* (Lisboa, Portugal) *(QoEVMA '22)*. Association for Computing Machinery, New York, NY, USA, 23–28.

https://doi.org/10.1145/3552469.3555712

[2] M Angela and Hendrik Knoche. 2006. Quality in Context-an ecological approach to assessing QoS for mobile TV. (01 2006).

[3] Toon De Pessemier, Katrien De Moor, Wout Joseph, Lieven De Marez, and Luc Martens. 2012. Quantifying Subjective Quality Evaluations for Mobile Video Watching in a Semi-Living Lab Context. *IEEE Transactions on Broadcasting* 58, 4 (2012), 580–589. https://doi.org/10.1109/TBC.2012.2199590

[4] ITU-R. 2019. Methodology for the Subjective Assessment of the Quality of Television Pictures. ITU-R Recommendation BT.500-14.

[5] Satu Jumisko-Pyykkö and Miska M. Hannuksela. 2008. 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* (Amsterdam, The Netherlands) *(MobileHCI '08)*. Association for Computing Machinery, New York, NY, USA, 63–72. https://doi.org/10.1145/1409240.1409248

[6] H. Knoche and M. A. Sasse. 2009. The Big Picture on Small Screens Delivering Acceptable Video Quality in Mobile TV. *ACM Trans. Multimedia Comput. Commun. Appl.* 5, 3, Article 20 (aug 2009), 27 pages. https://doi.org/10.1145/1556134.1556137

[7] Patrick Le Callet, Sebastian Möller, Andrew Perkis, et al. 2012. Qualinet white paper on definitions of quality of experience. *European network on quality of experience in multimedia systems and services (COST Action IC 1003)* 3, 2012 (2012).

[8] Jing Li, Lukáš Krasula, Yoann Baveye, Zhi Li, and Patrick Le Callet. 2019. AccAnn: A New Subjective Assessment Methodology for Measuring Acceptability and Annoyance of Quality of Experience. *IEEE Transactions on Multimedia* 21, 10 (2019), 2589–2602. https://doi.org/10.1109/TMM.2019.2903722

[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] Zhi Li, Anne Aaron, Ioannis Katsavounidis, Anush Moorthy, and Megha Manohara. 2016. Toward a practical perceptual video quality metric. *The Netflix Tech Blog* 6, 2 (2016), 2.

[11] Anne Oeldorf-Hirsch, Jonathan Donner, and Ed Cutrell. 2012. How Bad is Good Enough? Exploring Mobile Video Quality Trade-offs for Bandwidth-Constrained Consumers. (10 2012). https://doi.org/10.1145/2399016.2399025

[12] Wei Song and Dian W. Tjondronegoro. 2014. Acceptability-Based QoE Models for Mobile Video. *IEEE Transactions on Multimedia* 16, 3 (2014), 738–750. https://doi.org/10.1109/TMM.2014.2298217

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

[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] V. Zeithaml, A. Parasuraman, and Leonard L. Berry. 1990. Delivering quality service : balancing customer perceptions and expectations.