Acceptability and Quality of Experience in Over The Top Video

Petros Spachos, Weiwei Li, Mark Chignell, and Alberto Leon-Garcia
University of Toronto
Toronto, Canada

Leon Zucherman and Jie Jiang
Technology Strategy and Operations
TELUS Communications Company
Toronto, Canada

***Abstract*—Consumer acceptance is of great interest in the adoption of novel multimedia products and services. A number of factors can greatly influence the customer experience during a video session, impacting the acceptability of the product or service. Factors such as the Technical Quality (TQ) which covers the technical aspects of the signal quality can be controlled from the network provider. On the other hand, the network provider has no control on the subject's level of interest in a video, the Content Quality (CQ). Together TQ and CQ influence the Overall eXperience (OX) of the subject and, in the case of Over The Top (OTT) video, in a video session. In this paper, we present results from a user subjective study in which the impact of TQ and OX on the acceptability was investigated for OTT video sessions. To minimize the CQ impact on OX, the videos were carefully selected to have relatively neutral content. We assess the impact of TQ and OX on acceptability for videos containing impairment and failure events during lifecycle, thus affecting *Accessibility, Retainability* and *Integrity*. Our experimental results indicate that TQ and OX have a strong impact on acceptability.**

I. Introduction

Telecommunications have become a platform for the delivery of a broad range of existing and future services and applications. A crucial requirement is for networks to support this diversity of services and applications so that a customer's expectation in terms of Quality of Experience (QoE) is met. QoE is a subjective measure of a customer's experience with a Telecom service. Unlike the traditional Quality of Service (QoS), which is concerned primarily with network performance, QoE had not been systematically investigated and developed. In response, several international organizations from industry and academia have initiated efforts to research QoE in order to drive innovation in telecom services.

In every QoE system, the subject has the most crucial role. The definition of QoE is based on the notion of subject acceptability of a service. However, most systems replace acceptability by measuring subject satisfaction and degree of liking. While these systems can provide useful feedback regarding the subject opinion, they do not address the fundamental question: "Is a particular service acceptable by the subject or not?"

To answer this question, a definition of acceptability for video transmission needs to be provided. In the general sense, acceptability refers to the subject decision to either accept or reject a product or a service. In one approach, acceptability is defined by the concept of willingness to pay for a product or service [1] and hence, it is the outcome of a simple economic decision process. In another approach [2], in the Technology Acceptance Model (TAM), the acceptability is a multidimensional phenomenon and does not have one distinct definition. There are a number of factors that can predict the intention to use information systems. However, this concept targets acceptability before actual usage.

In QoE research, acceptability is treated as a whole offer – including price, cost and system – and relied on direct querying of subjects regarding the acceptability of the quality level experienced. In [3], the authors have defined acceptability in the context of mobile video QoE as *"binary measure to locate the threshold of minimum acceptable quality that fulfills subject quality expectations and needs for a certain application or system"*. In this work, we follow this definition for video transmission.

At the same time, acceptability, and QoE consequently, can be subjected to a wide range of complex and strongly interrelated factors [4]. In the context of video transmission, Technical Quality (TQ) in the delivery of the service as well as the level of the subject's interest in the content, the Content Quality (CQ), and consequently the Overall eXperience (OX), as discussed in [5], are among the most important influence factors, especially for the service provider. However, a direct connection between these factors and the acceptability is not known. This is of high importance to mobile network operators and providers. The acceptability of their offer is directly related to gaining and retaining or losing customers [6].

In this work, we address this issue by examining the influence of TQ and OX on the acceptability for OTT video sessions for videos with neutral CQ, i.e. neither too interesting nor too boring. Experiments with various types of impairments and failures were conducted. The goal is to answer the question: *How do TQ or OX impact the probability P_{Acc} that a video is judged acceptable?*

The rest of this paper is organized as follows: In Section II, the related work is reviewed. The session-oriented QoE along with the different impairment and failures are briefly described in III. Section IV gives a description of the experiments followed by a discussion on the results in Section V. Our conclusions are in Section VI.

II. Related Work

Recent research have been very active in QoE evaluation, especially in developing QoE models for OTT video applications [7]–[9]. An early indication of the need to assess QoE for an entire session is in [10]. Human factors on QoE, such as context, human memory and attention effects, are investigated in [7], [10]. The majority of quality measurement methods is based on Mean Opinion Score (MOS) [11]. Although the definition of QoE is based on the notion of acceptability, it has not been practically used as the metric of QoE in video delivery.

In recent years, a number of methods and models for acceptability have been proposed. In [3], the authors examine research methods for assessing acceptance of quality in subjective quality evaluation methods. Based on the evaluation results, a bidimensional research method is proposed, which combines acceptance and satisfaction in consumer oriented experiments. In [12], [13], a number of QoE models are proposed for videos with six different content types (sports, news, music, animation, comedy and movie) and with the use of different viewing devices (mobile phones, laptops and PDA). These models assume that the video content types are known and do not consider any other factors such as integrity impairments. In [14], they investigate the concept of acceptability for interactive data services. Their experiments demonstrate there is a consistent mapping between subjects' binary acceptance and ordinal satisfaction rating. A number of acceptability-based QoE models are proposed in [15]. The models are based on the results of comprehensive user studies and try to predict users acceptability. They also compare the results with three other well-known objective Video Quality Assessment metrics: PSNR, SSIM and VQM. In [16], the parameters video quality, network type, user watching behaviour and transport protocol are used to build a decision tree of audiovisual quality acceptance. In [17], the influence of interruptions during video playback has been examined.

In this work, we examine the impact of three factors – TQ, CQ and OX – on acceptability. The factors also discussed in [5], however, only integrity impairments are examined. We further extend the MOS concept to that of session MOS to include not only impairments but also failures.

III. Session-Oriented QoE and assessment

In this section, the session-oriented QoE concept that was proposed in [18] is briefly described, followed by the QoE assessments.

The customer experience in a service is determined by the entirety of interactions during the session of a customer with a service. Most QoE models have used network related factors such as encoding and frame rate as predictors to estimate the subject's MOS of received video quality [19], [20]. However, the traditional MOS is concerned with the "integrity" or degree of impairment in signal quality. In [18], we propose to extend the concept of MOS to that of a session MOS (sMOS) that encompasses *Integrity* impairments as well as failures in *Accessibility* and *Retainability*. We also demonstrate the need to extend the classical MOS scale (typically from 1 to 5) to take into account the negativity bias created by *Non-Accessibility*. Consequently, the sMOS scale had been extended, and it covers the range from 0 to 5. Hence, the measurement of the customers experience cover the entire session. In this paper, we follow this approach.

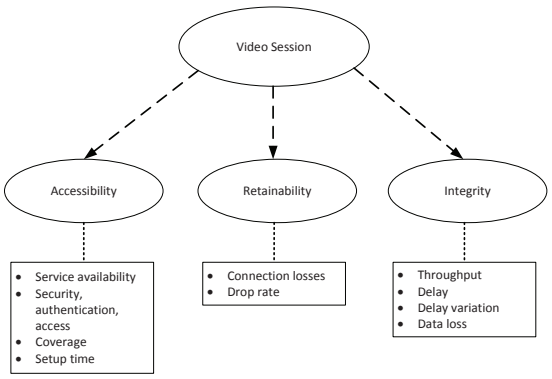

Fig. 1. Important components of a typical video session.

A session typically includes a customer who tries to start the service. A number of impairments and/ or failures can happen at any point of this process – at the beginning, during, and at the end of the session. A typical session is concerned with the following:

- *Accessibility*. Accessibility refers to the successful start of the session. It is the subject who first attempts to initiate the session. The session may start successfully. However, if the session fails to start, we say that an *"Accessibility"* failure has occurred. Service accessibility may be related to availability, security (authentication), access, coverage etc.
- *Retainability.* Retainability is the capability to continue the session until its completion, or until it is interrupted by subject action. If the session is terminated permanently due to a failure, this is a "*Retainability*" failure. In general, *Retainability* characterizes connection losses.
- *Integrity.* Integrity indicates the degree to which a session unfolds without excessive impairments. Even when a session does not experience any of the previous two failures, there are a number of service-specific impairments that may impact the QoE of the service. For instance throughput, delay, delay variation (or jitter) can impact the perceived quality of the service.

Figure 1 shows the different types of impairments/ failures along with related examples of each type that can occur during a session.

IV. Experiment Design and Methodology

In this section, our experimental approach and setup is described, followed by the procedure which was followed and the design of the impairments and failures introduced in our experiments.

TABLE I
VIDEO QUESTIONNAIRE AND POSSIBLE ANSWERS

Question	Possible Answers	Score
Is the technical quality of this video acceptable?	Yes/ No	1 or 0
Your evaluation of the technical quality in the video is:	Excellent/ Good/ Fair/ Poor/ Bad/ Terrible	5 to 0
The content of the video is:	Very interesting/ Interesting/ Neutral/ Boring/ Very boring	5 to 1
Your overall viewing experience (Content + Technical quality) during the video play back is:	Excellent/ Good/ Fair/ Poor/ Bad/ Terrible	5 to 0

A. Experimental Setup

The experiment took place in a controlled environment at the University of Toronto. Thirty subjects participated in the experiments. All the subjects finished the experiment. The conditions of participation were to have normal or corrected-to-normal vision and to not have participated in a video quality assessment experiment in the six months prior to the date of the experiment. All subjects were aged over 18 years.

All the subjects used the same computer with the same configuration. Each subject evaluated 30 video sessions in total, which lasted around 90 minutes. The videos are displayed in random order to control possible effects. Each video used in the experiment had a resolution of 512×288 pixels and a frame-rate of 30 frames-per second (fps). The complete videos were between 73 and 123 seconds in length with an average of 94.1 seconds. The video sessions consisted of 22 short movie trailers (teaser-trailers) and 8 short movies.

For video evaluation, we used the Absolute Category Rating (ACR) method [21]. After each video, each subject answered 4 questions based on their viewing experience. We used an extended 6-point scale for technical quality related questions, but kept the 5-point scale for the content evaluation question as in [22]. The questions and the corresponding scales are shown in Table I. The first question addresses whether the subject accepted the technical quality. The other three questions are regarding the TQ, CQ and OX. Videos were selected to be neutral, neither too interesting nor too boring.

B. Experiment Procedure

The experiment included the following four parts:

1) First, subjects answered pre-questionnaires which collected their demographics, video viewing habits, and video quality preferences.

2) Next, a training session was used to get them familiar with the video evaluation. They viewed 5 videos and answered the questionnaire used for video evaluation. These 5 videos included either *Integrity* impairments and *Retainability* failures or *Accessibility* failure in a pre determined order. However, the responses to the questionnaire were not used in any analysis.

3) In the video evaluation part, all 30 videos were divided into three sessions, and 10 videos were assessed in each session. Each video had either an *Accessibility* failure or two randomly assigned types of *Integrity* impairments/ *Retainability* failures, as will be described in the following section. There was a 10 minute break after each session. After each video, the subjects answered the four questions in Table I. The first question refers to the video acceptability. The second question is related to subject's perception of TQ, the third question is the evaluation of CQ and the last question asked subject's overall viewing experience. Subjects responded to assessment questions using the word labels. The numerical values corresponding to word labels shown in the third column of Table I. The Score values were used in the calculation of sMOS scores.The subjects watched videos with different types of impairments and failures.

4) In the last part, subjects answered a post-questionnaire for the viewing preference.

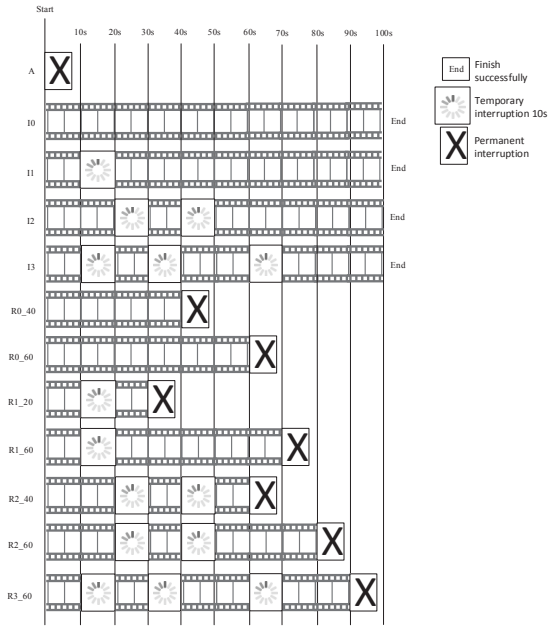

Fig. 2. Impairments and Failures used in the experiment.

C. Design Impairments and Failures

The video set was generated based on 30 unimpaired videos and the three types of impairments/ failures are introduced. A graphical representation of the different types of impairment/ failure that was used in the experiments is shown in Fig. 2.

A video with *Accessibility* failure (i.e. *Non- Accessibility*) is prefixed with an "A". This video corresponds to an *Accessibility* failure. When an *Accessibility* failure occurred, the video failed to start and the subject moved to the next video.

A video with *Integrity* disruptions is prefixed with an "I" followed by the number of the temporary interruptions. These videos have 0 to 3 temporary interruptions during the course of playback. In our experiments, we have I0, I1, I2 and I3.

A video with *Retainability* disruptions (impairments and failures) is prefixed with an "R" followed by the number of the temporary interruptions and the video time of the permanent interruption. These videos have zero to three temporary impairments and a *Retainability* failure. In our experiments

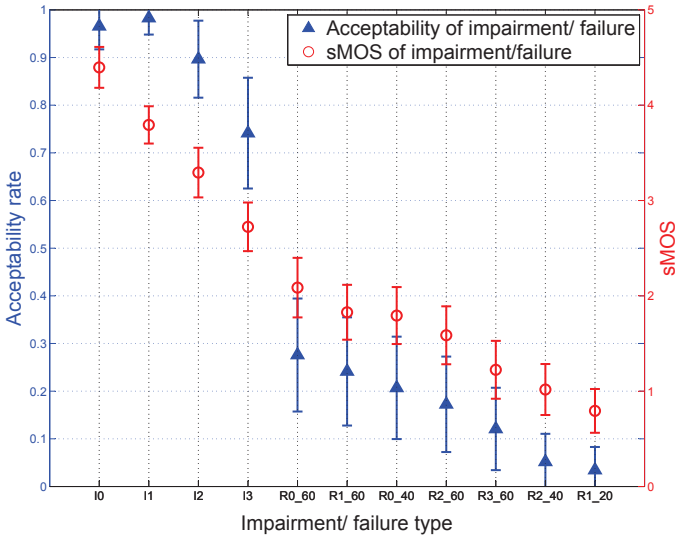

Fig. 3. Acceptability rate under Impairment/Failures

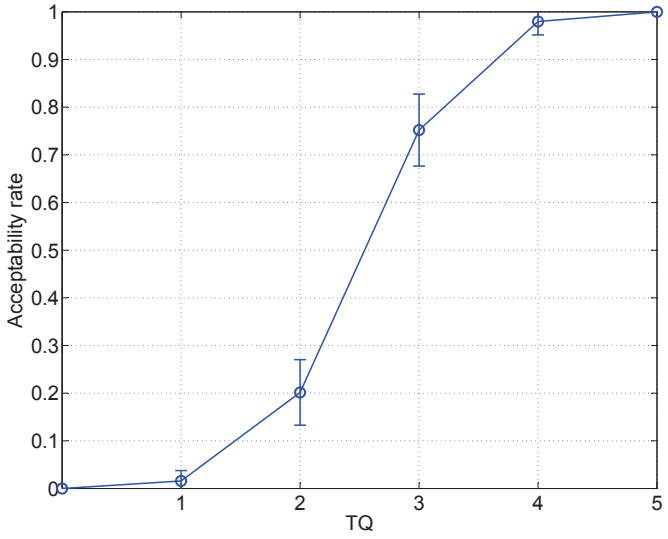

Fig. 4. Acceptability rate versus TQ

we have R0_40, R0_60, R1_20, R1_60, R2_40, R2_60 and R3_60.

As mentioned in the previous section, each video had zero, one or two *Integrity* impairment(s) possibly followed by a *Retainability* failure.

V. Results and Analysis

We employ the subject screening method proposed from Video Quality Experts Group (VQEG) High Definition Television (HDTV) Annex I. We calculate the Pearson correlations between each subject and the sMOS values for both TQ and CQ and the mean acceptability, $\bar{P}_{Acc}$. The Pearson correlation shows the linear relationship between two sets, i.e the linear relationship between TQ and acceptability and the linear relationship between CQ and acceptability. One subject is excluded because of low correlation on acceptability.

A. Objective Parameters vs. Acceptability Rate

Figure 3 shows the acceptability rate and the sMOS value for TQ under various impairments and failures. We found I0 - I2 show high acceptability (> 0.9), and the acceptability of I3 is above 0.7. However, mean acceptability of all the *Retainability* failures are less than 0.3. Clearly *Retainability* failures have a more detrimental impact on acceptability than *Integrity* impairments.

Note that acceptability shows a huge drop from 0.7414 in I3 to 0.2759 in R0_60, while TQ drops only from 2.72 to 2.0862. We conclude that when TQ is below a threshold, acceptability drops quickly even if change in TQ value is less than 1. As a binary scale, acceptability is able to provide a clear guidance to resource allocation and management.

Our experiments have shown that impairments and failures (objective parameters) have great impact on acceptability rate.

B. Technical Quality vs Acceptability

Figure 4 shows acceptability rate versus TQ. Acceptability rate is more than 70% when TQ≥ 3. It is interesting that the increase of acceptability rate is not significant from TQ$= 4$ to TQ$= 5$. It is also worth noting that the standard error for acceptability is low for small TQ and large TQ, and that it is somewhat larger in between.

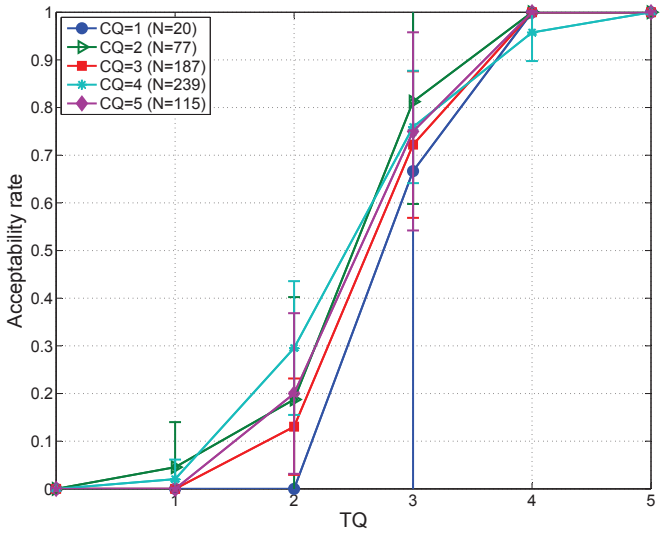

Fig. 5. Acceptability rate versus TQ under different CQ

Figure 5 gives the acceptability rates of all the N$= 638$ answers, under specific scores of CQ. We note that for the CQ level listed (for which the number of samples is sufficient), the acceptability versus TQ graph follows similar S-curves.

We further examine the CQ level under different TQ, in Fig. 6. As the TQ increases, the subjects tend to accept the video at any CQ level. We note that the CQ$= 1$ graph has very large standard error due to a very small sample size, and hence is not reliable. The results indicate that when CQ is relatively constant, the subject decides the acceptability of a video based mainly on TQ.

In an everyday scenario, the users decide to watch a video mainly when the content is relatively interesting, otherwise

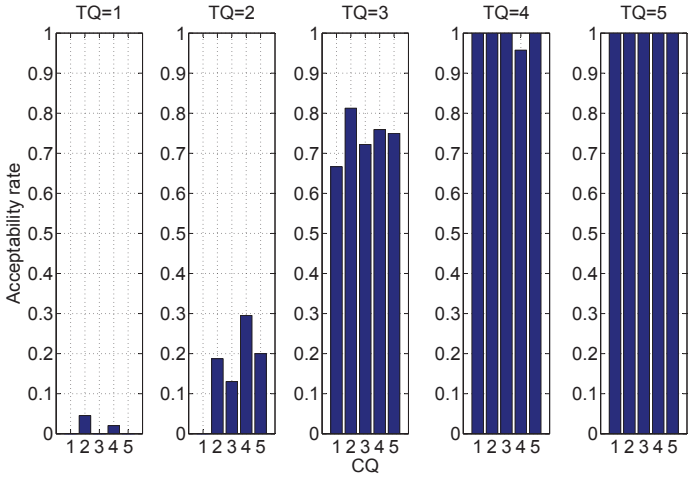

Fig. 6. Acceptability probability for different observed TQ levels. For each TQ level, the distribution of the CQ levels is shown.

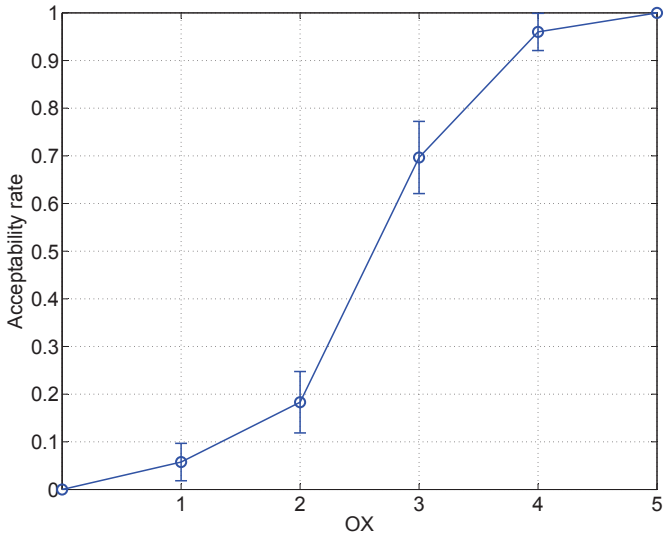

Fig. 7. Acceptability rate versus OX

they will voluntarily terminate the video. Our experimental design tried to approach a real life scenario for more realistic findings. This is useful towards modeling QoE.

We generate the regression model of acceptability (with raw regression weights) in order to fit the acceptability rate versus TQ curve. The regression model is:

$$\bar{P_{Acc}} = -0.1443 + 0.2559 \cdot TQ, \tag{1}$$

with $R^2 = 0.612$. The high value of R^2 indicates that when there are multiple impairments/ failures, TQ represents the level of acceptability to some extent.

We further applied, a piecewise quadratic regression model using the method of least squares. TQ is the independent variable, and P_{Acc} is the dependent variable. We have:

$$\bar{P_{Acc}} = \begin{cases} a_{11} \cdot TQ^2 + a_{12} \cdot TQ & TQ = 0, 1, 2 \\ a_{21} \cdot TQ^2 + a_{22} \cdot TQ + a_{23} & TQ = 3, 4, 5 \end{cases} \tag{2}$$

where $a_{11} = 0.0850$, $a_{12} = -0.0693$, $a_{21} = -0.1038$, $a_{22} = 0.9546$, and $a_{23} = -1.1776$. The goodness of fit for this model was $R^2 = 0.6824$. Comparing to the linear model, it improves the goodness of fit.

We also used the parameter values for each impairment/ failures as predictors for acceptability, we have:

$$\bar{P_{Acc}} = -0.4179 - 0.0788 \cdot B + 1.3460 \cdot VR, \tag{3}$$

where B (Buffering is the total video freezing time in seconds) is the number of temporary impairment (buffering) happened during each video, and VR (Viewing Ratio) is the viewing ratio for each video with/without a *Retainability* failure. The goodness of fit is $R^2 = 0.476$. Although B and VR are measurable objective parameters, they cannot reflect the acceptability to a service as precisely as TQ.

C. Overall Experience vs. Acceptability

Figure 7 gives the acceptability rates for the different OX levels. It is clear that an increase in OX is reflected in the acceptability. Moreover, OX is similar to TQ in its impact on acceptability.

To examine the fit of a linear model for the acceptability rate versus OX curve, we also generate the regression model:

$$\bar{P_{Acc}} = -0.1608 + 0.256 \cdot OX, \tag{4}$$

with $R^2 = 0.539$.

When entering OX as the predictor, the piecewise quadratic regression model shown below:

$$\bar{P_{Acc}} = \begin{cases} b_{11} \cdot TQ^2 + b_{12} \cdot TQ & TQ = 0, 1, 2 \\ b_{21} \cdot TQ^2 + b_{22} \cdot TQ + b_{23} & TQ = 3, 4, 5 \end{cases} \tag{5}$$

had an $R^2 = 0.5945$, where $b_{11} = 0.0340$, $b_{12} = 0.0236$, $b_{21} = -0.1117$, $b_{22} = 1.0455$, and $b_{23} = -1.4345$. It indicates that the nonlinear model improve the goodness of fit. But still, the value of R^2 when using OX as the predictor is lower than the value of R^2 when the predictor is TQ.

D. Content Quality vs. Acceptability

To further prove that CQ in our experiments has trivial impact on the acceptability of technical quality, as designed, we show the impact of CQ on acceptability in Fig. 8. It is evident that different CQ levels have little or no impact on the acceptability of technical quality. Under different CQ levels the acceptability varies between 0.3 and 0.5.

E. Pearson Correlation

We also investigated the correlation between acceptability and scores of TQ, CQ and OX. The first column in Table II shows the Pearson correlation coefficient between acceptability and CQ. We confirm that that the impact of CQ is negligible, as designed.

Our previous research has shown that although OX reflects the overall experience based on TQ and CQ, TQ is the main determinant in most cases. That's why p-values in the second column and the third column in Table II are close to each other.

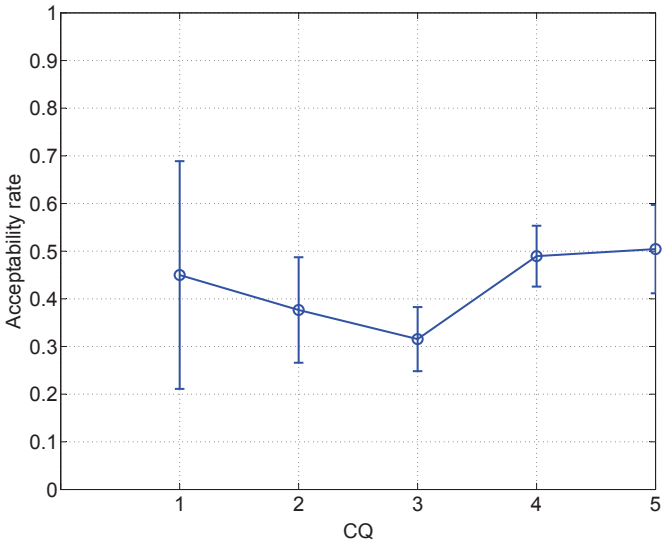

Fig. 8. Acceptability rate versus CQ

TABLE II
P-VALUE OF PEARSON CORRELATION COEFFICIENT FOR ACCEPTABILITY

		CQ	TQ	OX
I0-I3 and R failures	Acceptability	0.1118	0.7824	0.7339
I0	Acceptability	-0.0356	0.6774	0.4632
I1-I3	Acceptability	0.0294	0.5692	0.4895
R failures	Acceptability	0.0283	0.5987	0.5239

Hence, we conclude that TQ is the chief factor to determine the acceptability, when the content is neither too boring nor too interesting.

VI. Conclusions

In this paper, we examined the impacts of technical quality and overall experience on acceptability, with neutral video content. Experiments were conducted under different impairments and failures.

Following the experimental results, technical quality has a great impact on the acceptability. Moreover, subjects shown high acceptability rate when TQ is above the average. Another important outcome is that as TQ increases, the acceptability also increases.

We further examine the impact of OX on the acceptability. An increase in OX is reflected in the acceptability while CQ had little impact on it, as designed. We verify the correlation between the three factors and acceptability with the use of Pearson correlation function. We conclude that TQ and OX have significant impact on the acceptability.

Acknowledgment

This research was supported by a grant from TELUS and a matching grant from NSERC/CRD.

References

[1] A. Molnar and C. Muntean, "Cost-oriented adaptive multimedia delivery," *IEEE Transactions on Broadcasting*, vol. 59, no. 3, pp. 484–499, Sept. 2013.

[2] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view." *MIS Quarterly*, vol. 27, no. 3, pp. 425–478, 2003.

[3] S. Jumisko-Pyykkö, V. K. Malamal Vadakital, and M. M. Hannuksela, "Acceptance Threshold: A Bidimensional Research Method for User-Oriented Quality Evaluation Studies," *International Journal of Digital Multimedia Broadcasting*, vol. 2008, pp. 1–21, 2008.

[4] S. Möller and A. Raake, *Quality of Experience Advanced Concepts, Applications and Methods*. Springer, 2014.

[5] T. De Pessemier, K. De Moor, W. Joseph, L. De Marez, and L. Martens, "Quantifying the influence of rebuffering interruptions on the user's quality of experience during mobile video watching," *IEEE Transactions on Broadcasting*, vol. 59, no. 1, pp. 47–61, March 2013.

[6] P. Reichl and F. Hammer, "Charging for Quality-of-Experience - a New Paradigm for Pricing IP-based Services," in *Proc. 2nd ISCA Tutorial and Research Workshop on Perceptual Quality of Systems*, Sept. 2006.

[7] F. Dobrian, V. Sekar, A. Awan, I. Stoica, D. Joseph, A. Ganjam, J. Zhan, and H. Zhang, "Understanding the impact of video quality on user engagement," *SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 4, pp. 362–373, Aug. 2011.

[8] R. Mok, E. Chan, and R. Chang, "Measuring the quality of experience of http video streaming," *2011 IFIP/IEEE International Symposium on Integrated Network Management (IM)*, pp. 485–492, May 2011.

[9] A. Moorthy, L. K. Choi, A. Bovik, and G. de Veciana, "Video quality assessment on mobile devices: Subjective, behavioral and objective studies," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 652–671, Oct. 2012.

[10] M. Söderlund, "Behind the satisfaction façade: An exploration of customer frustration," *32nd European Marketing Academy Conference*, 2003.

[11] "Methods for subjective determination of transmission quality," *ITU-T Recommendation P.800*, Aug. 1996.

[12] F. Agboma and A. Liotta, "Quality of experience management in mobile content delivery systems," *Telecommunication Systems*, vol. 49, no. 1, pp. 85–98, 2012.

[13] V. Menkovski, A. Oredope, A. Liotta, and A. C. Sánchez, "Predicting quality of experience in multimedia streaming," *Proceedings of the 7th International Conference on Advances in Mobile Computing and Multimedia*, pp. 52–59, 2009.

[14] R. Schatz, S. Egger, and A. Platzer, "Poor, good enough or even better? bridging the gap between acceptability and QoE of mobile broadband data services," *2011 IEEE International Conference on Communications (ICC)*, pp. 1–6, June 2011.

[15] W. Song and D. Tjondronegoro, "Acceptability-based QoE models for mobile video," *IEEE Transactions on Multimedia*, vol. 16, no. 3, pp. 738–750, April 2014.

[16] T. De Pessemier, K. De Moor, A. Verdejo, D. Van Deursen, W. Joseph, L. De Marez, L. Martens, and R. Van de Walle, "Exploring the acceptability of the audiovisual quality for a mobile video session based on objectively measured parameters," *2011 Third International Workshop on Quality of Multimedia Experience*, pp. 125–130, Sept. 2011.

[17] A. Sackl, S. Egger, and R. Schatz, "Where's the music? comparing the QoE impact of temporal impairments between music and video streaming," *2013 Fifth International Workshop on Quality of Multimedia Experience*, pp. 64–69, July 2013.

[18] A. Leon-Garcia and L. Zucherman, "Session MOS to assess technical quality for end-to-end telecom session," *2012 IEEE Globecom Workshops (GC Wkshps)*, Dec. 2014.

[19] A. Khan, L. Sun, E. Jammeh, and E. Ifeachor, "Quality of experience-driven adaptation scheme for video applications over wireless networks," *IET Communications*, vol. 4, no. 11, pp. 1337–1347, July 2010.

[20] T. Zinner, O. Hohlfeld, O. Abboud, and T. Hossfeld, "Impact of frame rate and resolution on objective QoE metrics," *2010 Second International Workshop on Quality of Multimedia Experience*, pp. 29–34, June 2010.

[21] ITU-T, "Subjective video quality assessment methods for multimedia applications," *Recommendation P.910, Telecommunication standardization sector of ITU*, Sep, 2009.

[22] W. Li, H.-U. Rehman, M. Chignell, A. Leon-Garcia, J. Jiang, and L. Zucherman, "Video quality of experience in the presence of accessibility and retainability failures," in *10th International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness*, Greece, Aug. 2014.