Subjective QoE Assessment on Video Service: Laboratory Controllable Approach

Petros Spachos[§], Thomas Lin[*], Weiwei Li[*], Mark Chignell[*], Alberto Leon-Garcia[*], Jie Jiang[‡], Leon Zucherman[*]
[*]University of Toronto, Toronto, ON, Canada
[§]School of Engineering, University of Guelph, Guelph, ON, Canada
[‡]Technology Strategy and Operations, TELUS communications Company, Toronto, Canada

***Abstract*—This paper introduces research that addresses the subjective assessment of Quality of Experience (QoE) during the entire life cycle of a video session. We define a video session life cycle as the time from when a user attempts to initiate playback, until such time that the video ends either from normal video conclusion or through a network-induced failure. We provide a detailed description of our assessment methodology designed to discern whether a user's QoE would be impacted by the presence of failures. To accomplish this, we carefully select various test conditions to take into consideration the rating scale used, the types of impairments and failures seen by the user, and whether impaired videos are seen together with failed videos in multi-video sessions. The selection and creation of source video sequences are also discussed, as well as the use of between-subjects and within-subjects approaches for running our experiments in a controlled laboratory setting.**

Statistical analysis was carried out to interpret our experimental results. We compared the results of the between-subjects measures and the results of the within-subjects measures, and concluded that the introduction of a scale with an extended lower bound enabled subjects to more clearly express their dissatisfaction of videos with failures when compared to the traditional ITU 5-point rating scale. In addition, we observed that videos that were simply impaired but concluded normally did not have a statistically significant difference when an extended scale was used.

I. Introduction

Quality of Experience (QoE) has received wide attention from research institutions due to the explosive growth of mobile services. The fragile wireless networks used to deliver these services need to deal with a large amount of traffic data, while also satisfying users' expectations of quality. In view of this situation, service providers value the QoE for each service delivered by their wireless networks.

The recent trend in video traffic growth highlights the importance of delivering quality video to the end-users [1], [2]. To be able to deliver video at a quality level acceptable to their subscribers, service providers need to accurately measure the quality of the videos delivered over their networks. Without accurate QoE measurement, it is hard to predict customer satisfaction, which prevents service providers from being aware of any video viewing problems until reported by their customers. On the other hand, with the ability to automatically measure video quality, a service provider can respond to any network or service problem causing poor video quality in a timely manner, possibly even before a majority of their subscribers become aware of it.

978-1-5386-2723-5/17/$31.00

As a customer-centric measurement, subjective assessment is always the first step of a QoE assessment. Moorthy et al. have generated a video quality database for video distortions in heavily-trafficked wireless network from more than 50 subjects [3]–[5]. In addition, Staelens et al. have conducted a subjective assessment with test sequences which were full-length movies. They concluded that the QoE assessment results of these movies were different from the assessment results of short video sequences, which are often used in existing subjective assessment methodologies [6].

In this paper, we present a subjective assessment to investigate the QoE in a complete video session. The subjective assessment involved 108 subjects. Our studies employed video sequences containing a continuous story from YouTube, which provided a real-life viewing experience but much shorter than full-length movies. Our assessment compares the traditional International Telecommunication Union (ITU) QoE measurement methodology with our proposed session-based QoE measurement methodology by utilizing within-subjects and between-subjects measures.

In the following sections, we will introduce the methodology used for session-based QoE measurement. The rest of the paper is organized as follows. Section II introduces the notion of a session-based QoE to understand a user's perception of a video session. Section III describes the methodology of our subjective assessment. Section IV presents the statistical analysis of the subjective assessment results, and Section V concludes this paper.

II. Research and Standards for QoE Measurement

A. The Origination of QoE

The concept of QoE originates from the subjective quality assessment of speech signals. It is a general term that is defined by the ITU as the overall acceptability of an application or service, as perceived subjectively by the end-user [7]. According to this definition, the QoE of a video service can be measured by obtaining the user's subjective opinions on video quality. However, obtaining subjective opinions from end-users is not a feasible way to measure video QoE as the process is time consuming and requires human interaction.

A feasible alternative is to use a model that can reasonably estimate QoE. The performance of such a model can be judged by how well its estimations correlate with actual subjective measurements. Obtaining subjective opinions is always the first step in building a video QoE model. It guarantees that

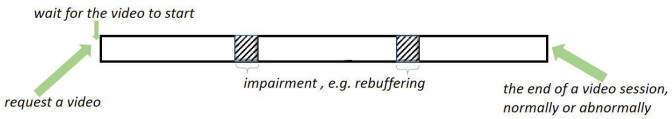

Fig. 1. Life cycle of a video session

the model can be designed to provide video QoE estimation which correlates well with subjective perceptions.

B. A Life-Cycle of Session, Impairments and Failures

With the popularity of over-the-top (OTT) video streaming, research on QoE proposed that user perception of a video service should be studied in the context of the life-cycle of a video session [8], [9]. They have found that the customer experience in a service is significantly impacted by the entirety of interactions during sessions of the service. As a type of service, the QoE of video streaming should be determined by the entire interaction during the life-cycle of the video session.

As shown in Figure 1, a video service session begins upon an attempt to play a video. This might happen, for example, when a user clicks on a video link within a web-page. During the playback of the video, the session may or may not have impairments, such as a rebuffering. The session ends when the user stops viewing the video, which may be categorized as either a normal end or an abnormal end. A normal end means that the user has successfully watched all the content he/she wanted to watch, and an abnormal end means the user cannot continue watching the video session due to some technical issue.

Concerning the QoE measurement of an entire session, three components impacting QoE have been proposed by the ITU Recommendation [10]:

- *Accessibility*: Accessibility refers to the successful start of the session. When the subject attempts to initiate the session, the session may or may not start successfully. If the session fails to start, we say that an *Accessibility* failure has occurred.
- *Retainability*: Retainability is the capability to continue the session until its completion, or until it is interrupted by the subject. If the session is permanently terminated due to a failure, this is a *Retainability* failure.
- *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.

To distinguish from Integrity impairments, we jointly refer to both Accessibility and Retainability problems as *failures*, since they both represent an abnormal end of a video service. Previous QoE assessment primarily focused on Integrity impairments [11], which did not include scenarios to examine whether a video can successfully start and/or normally end. Such scenarios should be included to examine the effects of Accessibility and Retainability failures on an end-user's QoE.

III. Subjective Assessment of Session-Based QoE

We are interested in developing a methodology for subjective assessments of session-based video QoE. Our previous studies focused on examining the impact on session-based QoE in the presence of failures [9], [12]–[17]. We have found that Retainability failures, in general, receive lower QoE values than Integrity impairments, and that Accessibility failures are always at the bottom [12]. The significant drop of subject ratings in the presence of failures motivated the experiment presented in this paper. In order to confirm the impact caused by the failures, we repeat our experiments while employing multiple rating scales and use both within-subjects/between-subjects measures. In this section, we present our session-based subjective assessment methodology.

A. Subjective Rating Method

The Absolute Category Rating (ACR) method, which was recommended in [11], was employed for our experiments. The ACR method is a quality assessment method where each subject answers a few questions based on the video he/she has just watched, and where each video (i.e. each single test condition) should be displayed only one time. Figure 2 shows the presentation pattern of ACR in a subjective video experiment, where v_i means the i-th video clip displayed in the experiment in Figure 2.

Four questions, as listed in Table I, are presented to the subjects at the end of each video clip.

1) Question 1 refers to the video acceptability. As a binary measure, the only possible answers are 'Yes' or 'No'.
2) Question 2 is related to the user's perception of the video's Technical Quality (TQ). We used a Likert scale to collect participants' responses regarding TQ, and under different groups of participants, we employed different rating scales. This is because the session-based QoE introduced Accessibility and Retainability failures, and our intent is to study the Mean Opinion Score (MOS) of failures under various rating scales. Therefore, the rating scale is considered a test condition in our experiment. More details will be introduced in Section III-C.
3) Question 3 is the evaluation of Content Quality (CQ). The full rating scale for Q3 is shown in Table II.
4) Question 4 asks the subject to evaluate their overall viewing experience, or OX (Overall eXperience). The rating scale of Q4 is always the same as the rating scale of Q2.

B. Video Sequences

Video sequences are video clips used for subjective evaluation. Source sequences for QoE experiments should be typical videos for mobile applications. The content of these sequences cover different categories including news, sports, music videos, advertisements and films (trailers or clips) [18], [19].

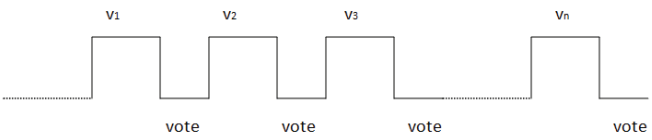

Fig. 2. Presentation pattern of videos followed by voting, with an optional break at prescribed intervals

TABLE I
QUESTIONNAIRE FOR EACH VIDEO

No.	Rating Criterion/Question	Possible Answer
1	Is the technical quality of this video acceptable?	Yes/No
2	Your evaluation of the technical quality in the video is:	A Likert Scale
3	The content of the video is:	Very interesting (5) to Very boring(1)
4	Your overall viewing experience (Content + Technical quality) during the video play back is:	A Likert Scale

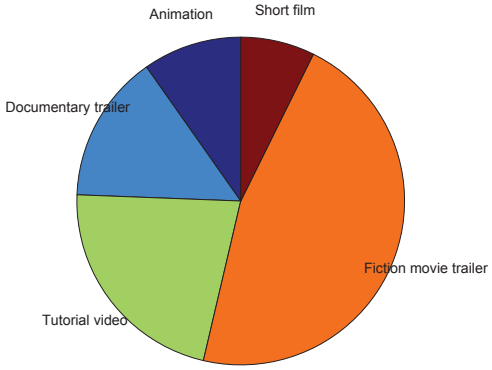

Fig. 3. The distribution of source sequence types

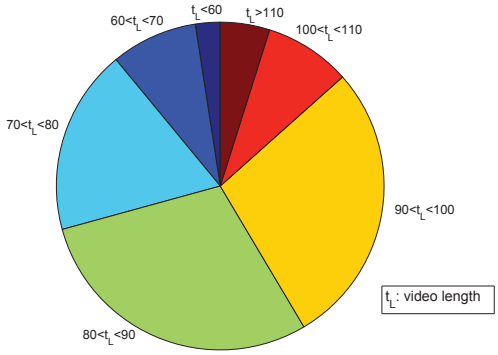

Fig. 4. The distribution of source sequence lengths

TABLE II
THE RATING SCALE OF CQ

Score	Label
5	Very interesting
4	Interesting
3	Neutral
2	Boring
1	Very Boring

Most researchers get source sequences from either public or private video libraries. These libraries provide standard sequences which can be converted based on the specific requirements of experiments. Video standards working groups, such as MPEG (Moving Picture Experts Group) and VQEG (Video Quality Experts Group), have released databases which are open to the public for testing. Another approach is for the research community itself to generate source sequences, but it requires professional equipment and well-designed content. Another video assessment project by the University of Texas at Austin collects raw sequences through a RED ONE digital cinematographic camera [3], [4], [20].

In our case, we used videos downloaded from YouTube as our source sequences. We selected videos which have a complete story or provide continuous description, since we value the user's perception when encountering Retainability failures. As shown in Figure 3, our source sequence library is composed of movie trailers (fiction and documentary), short films, explained videos (e.g. tutorials), and cartoons. The lengths of the video sequences are broken down and summarized in Figure 4. The lengths of our chosen videos generally ranged from 1 to 2 minutes, which is a balance between the consideration of video content and the feasibility of running our experiments. All sequences employed the same resolution (864×482) and a constant frame rate of 30 frames per second.

C. Test Conditions

There are two test conditions in our subjective assessment: rating scale and video quality issues. On the basis of the ITU 5-point scale, we introduced two more scales for subjective QoE assessment. We also designed multiple types of impairments and failures to evaluate users' perceptions on video quality issues.

Rating scale

Figure 5 shows the three rating scales we used in our experiments. As we mentioned before, only Q2 and Q4 in Table I employed the rating scale as a test condition. We named the 5-point scale as Scale A, the 6-point scale as Scale B, and the 7-point scale as Scale C. We use 5 to -1 to represent the corresponding score for each word label.

The characteristics of these scales are as follows:

- **Scale A**: This scale strictly follows the ITU standard, which is a 5-point rating scale.
- **Scale B**: This scale extends Scale A on the negative side by adding one more choice as the new bottom (Terrible).

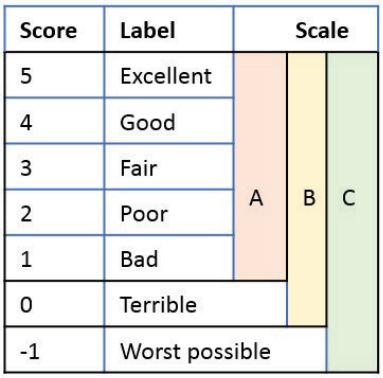

Fig. 5. Rating scales

We are interested in whether the user's evaluations tends to the negative side under the presence of failures.

- **Scale C**: This scale also extends on the basis of Scale A. However, two more negative choices are provided, i.e. "Terrible" and "Worst Possible". The goal of the designs for Scale B and Scale C is to discern whether the opinion scores of quality issues are stable when even worse opinion scores are provided as options.

Impairment and Failure Types

Each video sequence was used to represent a variety of impairment and failure issues including:

- Sequences containing Integrity impairments during playback. In our experiments, only rebuffering events were used to present Integrity impairments. A video sequence may have more than one rebuffering event during playback.
- Sequences containing Retainability failures during playback. In our experiments, a Retainability failure may happen with or without Integrity impairments. As long as the Retainability failure happened, the video session ended.
- Sequences containing Accessibility failures during playback. An Accessibility failure is a failure which occurs before any content of the video sequence is displayed.

In the experiment, we had eight types of impairment and failures. These include one type of Accessibility failure ($A1$), four types of Integrity impairments ($I0$, $I1$, $I2$, $I3$), and three types of Retainability failures ($R0$,$R1$, $R2$). Table III describes the details about these impairment/failure types. Note that $I0$ has no impairments nor failures during the playback, which we include in the table for completeness.

Block

One more concept we need to introduced in this section is *block*. A block refers to a set of videos that the user has continuously watched back-to-back, which more closely mimics a real-life user's time spent online watching videos. To study the difference between the traditional ITU QoE and session-based QoE, we have two different blocks: I-block and IF-block. Participants in an I-block only view videos with impairment types. On the other hand, participants in an IF-block view a mixture of videos with either impairment types or failure types. In our experiment, we use "I" to represent videos in an I-block, and "IF" to represent videos in an IF-block.

D. Experimental Design

The goal of this experiment is to investigate the impact of failures on QoE assessment. The general arrangement of the experiment is shown in Table IV. We divided participants into nine groups, G1 to G9. The symbol $Scale\_Block$ is used to represent the treatment of each group. For example, A_I means the subjects completed the experiment under Scale A with impairments only, and A_IF means the subjects completed the experiment under Scale A with both impairments and failures.

Each group contains twelve subjects, and each participant attended the experiment on two separate days, Day 1 and Day 2. In Day 1, they finished the first round of the experiment, i.e., they evaluated 31 videos. In Day 2, they evaluated another 31 videos, which is the second round of the experiment. The time interval between Day 1 and Day 2 was at least three days. We hypothesized that this forced delay between experimental rounds would eliminate any lingering memory effect of Day 1's experiment (e.g. if a subject was very frustrated by their experience in the first round, this would not be carried forward to the second round). This two-day experimental design was a within-subject repeated measure for session-based QoE.

The main difference between these two rounds is either the rating scale or block, or both. Across G1 to G6, the difference between the two rounds was the scale, and every two groups used the same scales with a reversed order, which was used to counter-balance any possible effect of memory regarding the previous scale within a group of subjects. Day 1 of G7 to G9 was fixed to use A_I, which follows the ITU standard (i.e. a 5-point rating scale involving only Integrity impairments). We did not arrange any reversed order to observe the traditional ITU QoE measure in our experiment, and to avoid the impact of failures on the A_I case.

Lastly, the design of this experiment contains both a between-subjects design and a within-subjects design. The between-subjects measure provides an overall comparison amongst A_I, A_IF, B_IF and C_IF between different groups of subjects. This method enabled us to study the impact of failures under various rating scales. To avoid any possible memory effect, we only employed data collected during Day 1. Table VI summarizes the details about rating scales and group arrangement. Conversely, the within-subjects measure compares A_I to one of $A/B/C\_IF$ within the same group of subjects. As shown in Table V, one factor was changed in this repeated measurement in each group.

E. Experimental Procedure

The experimental assessment was comprised of two components, a pre-questionnaire assessment and the main QoE assessment. Figure 6 shows the experimental procedure and the estimated time for each part.

In the pre-questionnaire assessment, the subjects answered questions which collected a verification of consent, as well

TABLE III
IMPAIRMENT AND FAILURE TYPES

Impairment/failure	Description
$I0$	There is no impairments and failure. Video is pristine.
$I1$	Video has a single temporary interruption of 10s duration happening at 15s.
$I2$	Video has two 10s temporary interruptions happening at 15s and 30s of the content display time.
$I3$	Video has three 10s temporary interruptions happening at 15s, 30s, and 45s of the content display time.
$R0\_70$	A permanent interruption happening at 70s of the content display time.
$R1\_30$	One 10s temporary interruptions happening at 15s; and a permanent interruption happening at 30s of the content display time.
$R2\_50$	Two 10s temporary interruptions happening at 15s and 30s; and a permanent interruption happening at 50s of the content display time.
$A10$	Video never starts to play. Video player display "failure-to-play" message immediately.

TABLE IV
THE EXPERIMENTAL DESIGN

Group ID	Subject No.	Day 1	Day 2
G1	12	A_IF	B_IF
G2	12	B_IF	A_IF
G3	12	A_IF	C_IF
G4	12	C_IF	A_IF
G5	12	B_IF	C_IF
G6	12	C_IF	B_IF
G7	12	A_I	A_IF
G8	12	A_I	B_IF
G9	12	A_I	C_IF

TABLE V
WITHIN-SUBJECTS DESIGN

Subject No.	Group ID	Symbols	
12	G7	First time	A_I
		Repeated	A_IF
12	G8	First time	A_I
		Repeated	B_IF
12	G9	First time	A_I
		Repeated	C_IF

TABLE VI
BETWEEN-SUBJECTS DESIGN

Subject No.	Group ID	Symbols
36	G7,G8,G9	A_I
24	G1, G3	A_IF
24	G2, G5	B_IF
24	G4, G6	C_IF

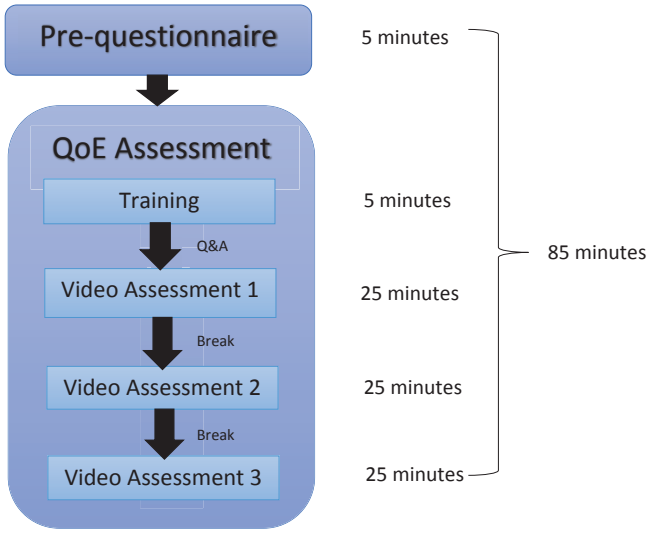

Fig. 6. Experimental procedure

as information regarding their demographics, video viewing habits, and video quality preferences. All information was anonymously recorded. The pre-questionnaire assessment needed around 5 minutes.

In the QoE assessment, the subjects first had a training phase which involved watching four video sequences. The training provided a guide of the procedure and introduced the terminology to participants. The training part was always shown before the primary video assessment, and the responses to the questionnaires here were not used for analysis. After training, we provided an optional break for subjects to ask questions in case they did not understand the questionnaire or other aspects of the experiment.

The total number of video sequences is 62, and each participant watched 31 video sequences per day/round. Each video contains a pre-defined impairment/failure type. The order of video content and impairment/failure type is randomized for each subject. However, we guaranteed that each subject watched the content of each video once, and the total number of each impairment/failure types are the same across all subjects. The distribution of impairment and failure types of the 31 videos is listed in Table VII. These videos were divided into three parts. The number of videos were almost equal in each part, except one part had 11 video sequences compared to the others' 10 video sequences. After each part, there is a 10 minute break to avoid visual and mental fatigue.

F. Laboratory Controllable Implementation

We implemented this experiment in a laboratory controllable approach, meaning that all video sequences in the experiment shown to users were loaded locally instead of from the

TABLE VII
THE DISTRIBUTION OF IMPAIRMENT AND FAILURE TYPES

Impairment/ failure	No./subject (IF block)	No./subject (I block)
$I0$	6	7
$I1$	6	8
$I2$	6	8
$I3$	6	8
$R0\_70$	2	-
$R1\_30$	2	-
$R2\_50$	2	-
$A10$	1	-

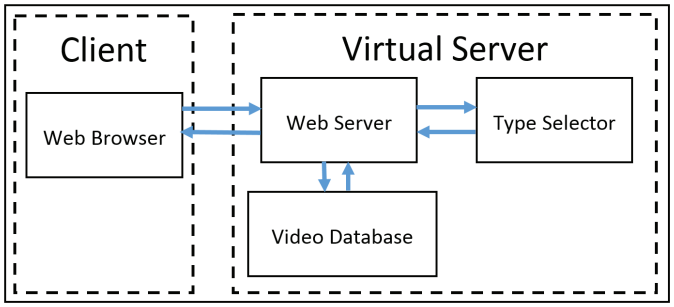

Fig. 7. The experiment implementation architecture

Internet. For each video's playback, we injected one specific impairment/failure type into the video for the user's evaluation.

The experimental implementation architecture is shown in Figure 7. When the web server receives a video request sent from a web browser, the server sends a query request to the type selector. The type selector is in charge of choosing one impairment/failure type for this video request. After the web server receives the selected type, it fetches the video with the specific impairment/failure type from the video database and returns it to the browser. The video database stores multiple copies of a video with the same content. Each copy contains a specific impairment/failure type, which we refer to as a single test sequence. For example, we have predefined seven impairment/failure types in one experiment. After we choose a pristine video as the original content, we generated six (since Accessibility failure is always the same) additional video clips which have the same content, but with different impairment/failure types injected in. The database stores these additional video clips for the QoE experiment.

The use of the video database is needed due to the essence of the laboratory controllable approach. The laboratory controllable approach requires full control over network performance, which is often difficult to achieve. Hence, we generated test sequences with hard-coded impairments and failures using an open-source video editing software, FFmpeg. To simulate a video service session in a laboratory environment, we employed Java Server Pages (JSP) technology and built a virtual server to operate a web server within the local host. For each station computer in our laboratory, we had an identical architectural setup.

The total number of participants in our experiment was 108. The subject number in each scenario satisfied the minimum

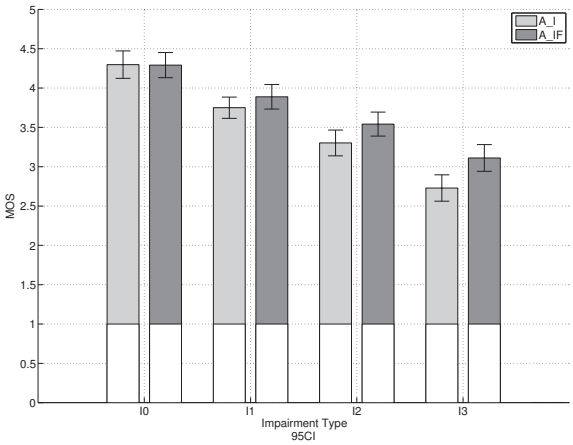

Fig. 8. The MOS of impairments, A_I vs. A_IF, G7

requirement of the ITU standard, and all subjects were over 18 years of age and had normal or corrected-to-normal vision.

IV. EVALUATION OF SUBJECTIVE ASSESSMENT

A. Within-Subjects Measurement

As mentioned in Section III, this experiment included a within-subjects measure of session-based QoE. Figure 8 plots the MOS values of impairments of G7 under A_I and A_IF, i.e. a same group of subjects took the same experiment with and without the appearance of failures on two separate days. The A_I case represents the traditional QoE measurement methodology proposed by the ITU standards, while the A_IF case measures the session-based QoE under Scale A. A shift can be visually observed, and a two-way repeated measures ANOVA, shown in Table VIII, determined that the impact of impairment types and the presence of failures are both significant factors on the MOS of TQ, which supports our observation in [12].

Figure 9 shows the MOS values of impairments of G8 under A_I and B_IF. The repeated measures ANOVA (Table IX) shows that the change of the block accompanying the employment of Scale B is no longer a significant factor to impact the MOS of TQ. However, the interaction effect between impairment types and $Scale\_Block$ becomes significant, which was unexpected, indicating that Scale B has a risk of increasing the complexity of QoE models.

The same analysis was employed on G9 (A_I vs. C_IF), with the results of the ANOVA shown in Table X. After adding two new points "Terrible" and "Worst Possible", the MOS values of TQ is significantly changed by impairment types only. From Figure 10, we can tell that the MOS values under A_I and C_IF are close to each other.

Summarizing from the above observations, the presence of failures impacts the QoE of a video session. The impact reflects not only on the QoE of the video clips with the failures, but also on the QoE of the video clips with impairments. This is because the user's perception is a subjective judgement based on multiple items he/she perceives in a given time span. To combine our experimental results from typical impairment-only QoE assessments using Scale A, we studied the possibil-

TABLE VIII
TWO-WAY REPEATED MEASURES ANOVA, A_I VS. A_IF, G7

	NumDF	denDF	F-value	p-value
(Intercept)	1	641	2398.5588	$< .0001$
A_I vs. A_IF	1	641	7.3625	0.0068
Impairment types	3	641	36.4103	$< .0001$
Block:impairment types	3	641	2.6009	0.0512

TABLE IX
TWO-WAY REPEATED MEASURES ANOVA, A_I VS. B_IF G8

	NumDF	denDF	F-value	p-value
(Intercept)	1	641	4451.021	$< .0001$
A_I vs. B_IF	1	641	0.147	0.7014
Impairment types	3	641	28.913	$< .0001$
$Scale\_Block$:impairment types	3	641	2.809	0.0388

TABLE X
TWO-WAY REPEATED MEASURES ANOVA, A_I VS. C_IF G9

	NumDF	denDF	F-value	p-value
(Intercept)	1	641	763.9484	$< .0001$
A_I vs. C_IF	1	641	1.8319	0.1764
Impairment types	3	641	23.2054	$< .0001$
$Scale\_Block$:impairment types	3	641	0.5413	0.6542

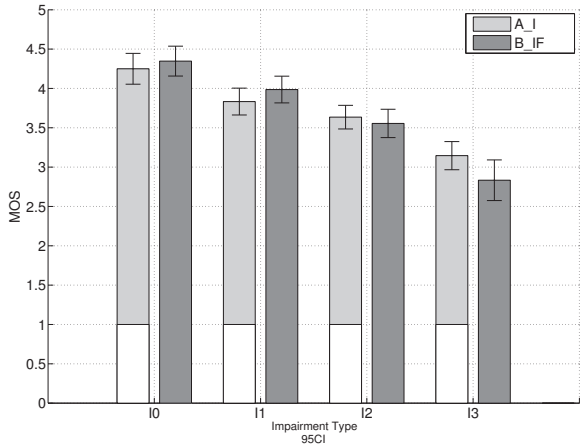

Fig. 9. The MOS of impairments, A_I vs. B_IF, G8

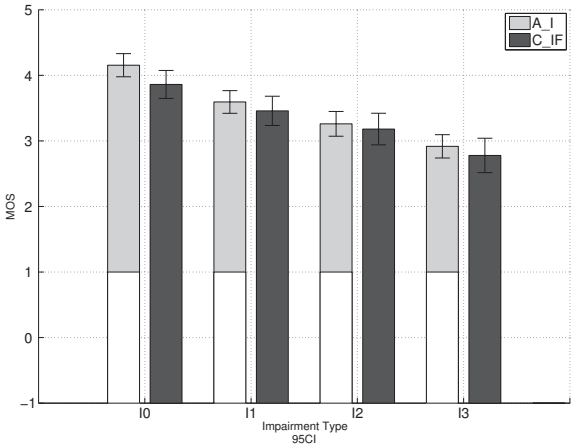

Fig. 10. The MOS of impairment, A_I vs. C_IF, G9

ity of employing a broader scale and assessing impairments and failures, i.e. the statistical analysis on the B_IF and C_IF cases. It seems that Scale C provides MOS values for impairments which are not statistically distinguishable from Scale A. This indicates that the data from QoE assessments using Scale A are compatible with QoE assessments that combines impairments and failures using Scale C.

On the other hand, we compared the MOS values, 95CI, and standard deviations of failure types under different rating scales. As shown in Table XI, we also found that using the broader scale introduced higher standard deviations as well as higher 95CIs of ratings of failures. The risk of extending the number of rating points is to increase the variability of the scale. However, we have a relatively small sample size (twenty-four samples for each case). This within-subjects measure provides a primary comparison under various rating scales and excluded the possibility of Scale B in future QoE assessments. To further understand the feasibility of Scale C, more subjective experiments should be conducted as discussed in [21].

B. Between-Subjects Measurement

Besides the within-subject measures, our subjective assessment also included sufficient data which can be used for between-subjects measures. We will compare TQ under A_I, A_IF, and C_IF between different groups of subjects as shown in Table VI.

Table XII shows the results of the two-way ANOVA to compare the MOS of TQ of impairments under A_I (G7 to G9) and A_IF (G1 and G3). We find that the addition of failures does not make a statistically significant impact on the MOS values of impairments. Figure 11 shows the MOS values and 95CI of these two cases.

TABLE XI
MOS, 95CI, AND SD OF FAILURES

Case	Failure types	MOS	95CI	Standard Deviation
A_IF, G7	$R0\_70$	2.1250	0.3367	0.7974
	$R2\_50$	1.8333	0.3871	0.9168
	$R1\_30$	1.7917	0.3731	0.8836
	$A10$	1.0833	0.1834	0.2887
B_IF, G8	$R0\_70$	2.2083	0.4819	1.1413
	$R2\_50$	1.5000	0.3736	0.8847
	$R1\_30$	1.1667	0.3871	0.9168
	$A10$	0.2500	0.3749	0.6216
C_IF, G9	$R0\_70$	1.7500	0.6618	1.5673
	$R2\_50$	0.8750	0.6508	1.5411
	$R1\_30$	0.7500	0.6618	1.5673
	$A10$	0.5000	0.7899	1.2432

TABLE XII
TWO-WAY BETWEEN ANOVA, A_I (G7, G8, G9) VS. A_IF (G1, G3)

	DF	Sum Sq	Mean Sq	F-value	p-value
A_I vs. A_IF	1	1.2	1.20	1.711	0.1910
Impairment types	3	422.9	140.96	200.428	$< 2e-16$
$Scale\_Block$:impairment types	3	4.9	1.62	2.302	0.753
Residuals	1684	1184.3	0.70		

TABLE XIII
TWO-WAY BETWEEN ANOVA, A_I (G7, G8, G9) VS. C_IF (G4, G6)

	DF	Sum Sq	Mean Sq	F-value	p-value
A_I vs. C_IF	1	0.2	0.24	0.293	0.588
Impairment types	3	424.0	141.33	173.855	$< 2e-16$
$Scale\_Block$:impairment types	3	4.7	1.57	1.929	0.123
Residuals	1684	1369.0	0.81		

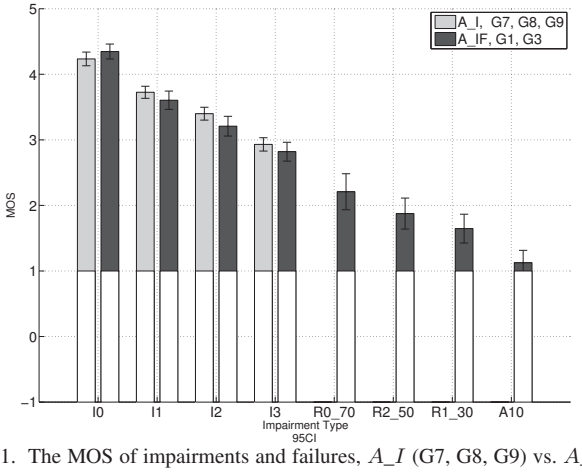

Fig. 11. The MOS of impairments and failures, A_I (G7, G8, G9) vs. A_IF (G1, G3)

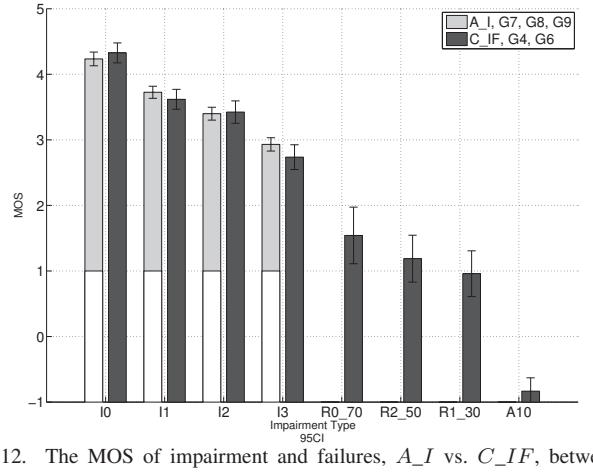

Fig. 12. The MOS of impairment and failures, A_I vs. C_IF, between-subjects

Figure 12 shows the MOS values of A_I (G7 to G9) and C_IF (G4 and G6). Table XIII shows the results of the two-way ANOVA to compare the MOS of TQ under A_I and C_IF. The results suggest that there is no statistically significant impact of scale together with the appearance of failures on the MOS of TQ. It also indicates that data from QoE experiments using only impairments with using Scale A, i.e. the traditional QoE assessment, are compatible with results combining impairments and failures using Scale C, which is valuable for further studies on session-based QoE.

One more thing we find is that the p-value in ANOVA of A_I (G7 to G9) vs. A_IF (G1 and G3) is 0.191, while the p-value in ANOVA of A_I (G7 to G9) vs. C_IF (G4 and G6) is 0.588. This suggests that there is no significant

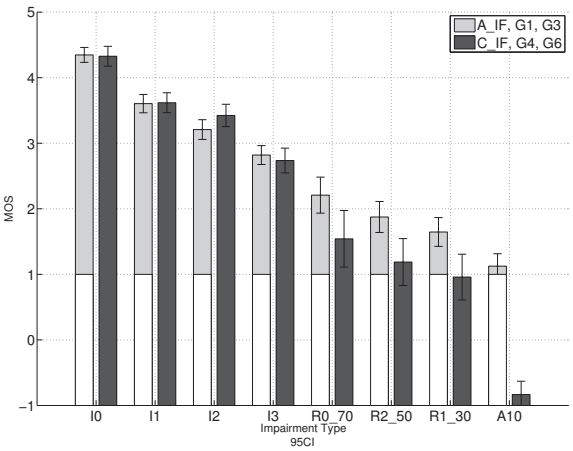

Fig. 13. The MOS of impairment and failures, A_IF vs. C_IF, between-subjects

difference between the I block and the IF block under Scale A in between-subjects measures. The results also suggests that the employment of Scale C in IF block increases the possibility that the mean values of different groups come from a same population.

Figure 13 shows the MOS values and 95CI of A_IF (G1 and G3) and C_IF (G4 and G6). The ANOVA results suggest that the rating scale has no statistically significant effect on the MOS of TQ of impairments, while the MOS values of failure types have huge differences under different rating scales.

V. Conclusion

In this paper, we provided a methodology for a session-based QoE subjective assessment of video services. We strictly followed the requirements for subjective assessment as proposed by the ITU standards while designing our own test conditions and selecting specific video sequences depending on the purpose of the experiment.

The implementation of within-subjects and between-subjects measures helped us to verify our experimental objective: confirming the impact caused by the presence of failures in session-based video services. Depending on the specific cases, two-way repeated measures ANOVA and two-way ANOVA were used to prove the existence of statistically significant differences.

The data of this subjective assessment are also used for other objective assessment. By utilizing this data, we have studied the relationship between TQ and application performance metrics [22], as well as investigated the subjective behavior on ratings for outlier detection and user classification [16].

Acknowledgment

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

References

[1] A. Pande, V. Ahuja, R. Sivaraj, E. Baik, and P. Mohapatra, "Video delivery challenges and opportunities in 4G networks," *IEEE MultiMedia*, vol. 20, no. 3, pp. 88–94, July 2013.

[2] Sandvine, "The global internet phenomena report," Sandive Incorporated. ULC, Tech. Rep., 2h 2013.

[3] 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.

[4] A. K. Moorthy, L. K. C. ande G. deVeciana, and A. C. Bovik, "Mobile video quality assessment database," *IEEE ICC Workshop on Realizing Advanced Video Optimized Wireless Networks*, 2012.

[5] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "A subjective study to evaluate video quality assessment algorithms," *Proc. SPIE*, vol. 7527, pp. 75 270H–75 270H–10, 2010. [Online]. Available: http://dx.doi.org/10.1117/12.845382

[6] N. Staelens, S. Moens, W. Van den Broeck, I. Marieï£¡n, B. Vermeulen, P. Lambert, R. Van De Walle, and P. Demeester, "Assessing quality of experience of IPTV and video on demand services in real-life environments," *Broadcasting, IEEE Transactions on*, vol. 56, no. 4, pp. 458–466, 2010.

[7] ITU-T, "Vocabulary for performance and quality of service, amendment 2: New definitions for inclusion in recommendation ITU-T p.10/g.100," *Recommendation ITU-T P.10/G.100 (2006) - Amendment 2, Telecommunication Standardization Sector of ITU*, 2008.

[8] 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.

[9] A. Leon-Garcia and L. Zucherman, "Session MOS to assess technical quality for end-to-end telecom session," in *Globecom Workshops (GC Wkshps), 2014*, Dec. 2014, pp. 681–687.

[10] ITU-T, "Overall network operation, telephone service, service operation and human factors - definitions of terms related to quality of service," *Recommendation E.800, Telecommunication standardization sector of ITU*, July, 2009.

[11] ——, "Subjective video quality assessment methods for multimedia applications," *Recommendation P.910, Telecommuniation standardization sector of ITU*, Sep, 2009.

[12] W. Li, H.-U. Rehman, D. Kaya, M. Chignell, A. Leon-Garcia, L. Zucherman, and J. Jiang, "Video quality of experience in the presence of accessibility and retainability failures," in *Heterogeneous Networking for Quality, Reliability, Security and Robustness (QShine) (invited paper), 2014 10th International Conference on*, Aug 2014, pp. 1–7.

[13] W. Li, H. Ur-Rehman, M. Chignell, A. Leon-Garcia, L. Zucherman, and J. Jiang, "Impact of retainability failures on video quality of experience," in *Signal-Image Technology and Internet-Based Systems (SITIS), 2014 Tenth International Conference on*, Nov 2014, pp. 524–531.

[14] W. Li, P. Spachos, M. Chignell, A. Leon-Garcia, L. Zucherman, and J. Jiang, "Impact of technical and content quality of overall experience of OTT video," in *IEEE Consumer Communications and Networking Conference (CCNC) 2016*, Aug 2016.

[15] M. Chignell, W. Li, A. Leon-Garcia, L. Zucherman, and J. Jiang, "Enhancing reliability through screening and segmentation: An online video subjective quality of experience case study," *Procedia Computer Science*, vol. 69, pp. 55 – 65, 2015, the 7th International Conference on Advances in Information Technology.

[16] W. Li, P. Spachos, M. Chignell, A. Leon-Garcia, J. Jiang, and L. Zucherman, "Capturing user behavior in subjective quality assessment of ott video service," in *IEEE Global Communications onference (GLOBECOM)*, Dec. 2016.

[17] P. Spachos, W. Li, M. Chignell, A. Leon-Garcia, J. Jiang, and L. Zucherman, "Acceptability and quality of experience in over the top video," in *IEEE International Conference on Communications(ICC) Workshops, 2015*, Jun. 2015.

[18] S. Winkler and R. Campos, "Video quality evaluation for internet streaming applications," *Proc. SPIE*, vol. 5007, pp. 104–115, 2003.

[19] S. Winkler and F. Dufaux, "Video quality evaluation for mobile streaming applications," *Proc. SPIE*, vol. 5150, pp. 593–603, 2003. [Online]. Available: http://dx.doi.org/10.1117/12.509910

[20] A. K. Moorthy, L. K. Choi, G. deVeciana, and A. C. Bovik, "Subjective analysis of video quality on mobile devices," *Sixth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM) (invited article)*, 2012.

[21] Q. Huynh-Thu, M.-N. Garcia, F. Speranza, P. Corriveau, and A. Raake, "Study of rating scales for subjective quality assessment of high-definition video," *Broadcasting, IEEE Transactions on*, vol. 57, no. 1, pp. 1–14, March 2011.

[22] W. Li, P. Spachos, M. Chignell, A. Leon-Garcia, L. Zucherman, and J. Jiang, "Understanding the relationships between performance metrics and QoE for over-the-top video," in *IEEE International Conference on Communications Workshops (ICC), 2016*, Jun. 2016.