Video Quality of Experience in the Presence of Accessibility and Retainability Failures

(Invited Paper)

Weiwei Li, Hamood-Ur Rehman,
Diba Kaya, 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*—Accurate Quality of Experience measurement for streaming video has become more crucial with the increase in demand for online video viewing. Quantifying video Quality of Experience is a challenging task. Significant efforts to quantify video Quality of Experience have primarily focused on the measurement of Quality of Experience for videos with network and compression related impairments. These impairments, however, may not always be the only main factors affecting Quality of Experience in an entire video viewing session. In this paper, we evaluate Quality of Experience for entire video viewing sessions, from the beginning to the end. In doing so, we evaluate videos with temporary interruptions as well as those with permanent interruptions or failures. We consider two types of failures, namely Accessibility and Retainability failures, and present the results of two subjective studies. These results indicate: (a) Accessibility and Retainability failures are rated lower compared to temporary interruption impairments; (b) Accessibility failures are rated close to the lowest value on the rating scale; and (c) the traditionally used 5-point scale to measure video Quality of Experience is not sufficient in the presence of Accessibility and Retainability failures.**

***Index Terms*—Video quality assessment, quality of experience (QoE), comparison of rating scales, subjective evaluation, accessibility, integrity, retainability.**

I. Introduction

Video traffic on the Internet has seen dramatic growth [1], [2] and the associated bandwidth requirement is stressing networks, especially wireless cellular networks. The increasing importance of video services implies that network and video service providers need to be aware of the expectations of their consumers in terms of video quality, and that they need to have the mechanisms in place to determine whether these expectations are being met. In order to understand of video quality and user expectations it is necessary to develop new models to assess video quality from the perspective of users.

Video Quality of Experience (QoE) is a measure to quantify the video quality perceived subjectively by the end-user. Quality of Experience is described by the International Telecommunication Union's Telecommunication Standardization Sector (ITU-T) as "The overall acceptability of an application or service, as perceived subjectively by the end-user [3]." The determination of QoE involves the presentation of various stimuli to users and the collection of their subjective judgment of the video quality. The assessment of video QoE is therefore laborious and expensive, and this drives a need to develop models that can be used to predict QoE.

A variety of impairments and network and system effects affect video QoE. Traditional video compression coding has focused on the effect of video coding artifacts on the viewer's subjective assessment of quality. A second focus has involved the impact of transmission errors and packet delay and losses on video quality. More recently, the proliferation of Over-the-Top (OTT) video streaming video that is delivered over TCP has placed a new focus on the impact video image freezing (that occurs when the coder playout buffer is depleted) on user assessment of QoE. For example, ITU-T Study Group 12 is currently developing models for progressive download [4], [5] and presently is working on developing models for adaptive streaming.

We are interested in the assessment of the QoE for the overall session because video streaming occurs in the context of a session that has a life cycle from invocation of the video, to video play, and then video conclusion. Traditional studies of QoE focus on the impact of specific impairments that affect the "Integrity" of a video, typically using relatively short video clips. Video streaming typically involves longer video sequences and therefore the determination of session-based QoE entails new experimental studies.

Video sessions are subject to network and system failures that can result in the inability to access a video or in the premature termination of a video session. These "Accessibility" failure and "Retainability" failure events clearly impact the user experience and their impact on QoE needs to be investigated. Baumeister et al [6] have proposed the principle that "bad is stronger than good" from their findings that bad events have more impact than good ones across a broad range of psychological phenomena. In the context of video QoE sessions, this principle suggests that it is very important to develop an understanding of the impact of Accessibility and Retainability failures.

In this paper we consider the assessment of QoE for video sessions that are subject to integrity impairments as well as Accessibility and Retainability failures. In the next section, we discuss the development of a generalized MOS

QSHINE 2014, August 18-20, RHODES, Greece

DOI 10.4108/icst.qshine.2014.256294

(gMOS) measure to assess video session QoE. We discuss the attributes of an ideal gMOS model that incorporates traditional integrity-focused MOS models in a straightforward fashion. We identify the critical role of rating scale in the development of the gMOS model. We summarize the results of a prior study (Experiment 1) indicating that the traditional 5-point scale is not sufficient to achieve our ideal gMOS model, and suggesting the use of an expanded rating scale. In Section III we present the experimental approach (Experiment 2) that we used to investigate gMOS with an expanded 6-point scale. We discuss the video clips, impairments, and failures that were used to produce the stimuli presented to users. We also discuss the number of subjects and the sample sizes that were gathered from the experiment. Section IV presents an analysis of our experimental results. Specifically we present detailed comparisons of QoE assessments of integrity impairments, and failures in Accessibility and Retainability in Experiments 1 and 2. In Section V we present our conclusion that, while the expansion to a six-point assessment scale takes us closer to the ideal gMOS model, additional future work should investigate even broader rating scales. We also conclude that a more detailed examination of Retainability failure impacts on QoE is necessary.

II. Generalized MOS Model and Prior Work

Video QoE is usually evaluated using a 5-point rating scale ranging from 1 ("Bad") to 5 ("Excellent"'). The MOS (Mean Opinion Score) for a stimuli σ is simply the mean of the responses from subjects. Typically we are interested in the MOS for a certain impairment over a class of videos, for example, the MOS for videos from a certain set with a freezing beginning at time t_0 and ending at time t_1. Thus in a study on the QoE of videos with a variety of freezing impairments we will prepare several versions of each video from a set of clips, each with a particular impairment. The QoE study then provides a mapping that specifies the estimated MOS Q(σ) for each impairment of interest σ from a set Σ.

In session-based QoE we are interested in finding the mapping for gMOS $Q(\sigma)$ for each impairment of interest σ from an expanded $\tilde{\Sigma}$ that now also include Accessibility and Retainability failures, that is, video clips that fail to play or that terminate prematurely. In the ideal case, the gMOS value for σ from Σ, e.g. an integrity impairment, would remain unchanged in an experiment that also includes Accessibility and Retainability failures. If this were the case, then it would be possible to re-use MOS models that focus on integrity only and augment the model with the gMOS values for the additional failures. In our first study (Experiment 1) we found that the gMOS values for Integrity impairments tend to shift upward with the introduction of Accessibility and Retainability failures. In Experiment 1, Accessibility and Retainability failures were judged much worse than Integrity impairments, and they also had the effect in an upward shift in gMOS values for the Integrity impairments. These results suggest that an expanded rating scale should be considered to provide room for the lower ratings for failure events without changing the ratings of the Integrity events. In the next section we present the design of Experiment 2 to explore gMOS using an expanded 6-point scale.

III. Experiment Description

This section presents the particulars of Experiment 2. The setup of Experiment 2 was similar to that of Experiment 1 described in [7]. The main difference was the choice of rating scale. In Experiment 1, we used a 5-point scale for gMOS measurement. On the other hand, in Experiment 2, we used a 6-point scale for gMOS measurement. We expanded the rating scale by adding the value 0 at the low end with the label "Terrible [8]."

A. Experimental Procedure

The second experiment consisted of three parts, Part 1, Part 2, and Part 3. In Part 1, subjects filled out two pre-questionnaires. These pre-questionnaires were used to collect information about subjects' demographics, personality, patience and video viewing habits, and general preferences with regards to video quality.

In Part 2, subjects evaluated 30 videos. Absolute Category Rating (ACR) method recommended in [9] was used. The subjects answered four questions after viewing each video. Table I shows these questions [7]. Question no. 2 in Table I is similar to the video-technical-quality question used in [10]. This question was our measure for gMOS and used a 5-point scale in Experiment 1 [7] and a 6-point scale in Experiment 2. In order to avoid fatigue, Part 2 was divided into three video evaluation sessions, Session 1, Session 2, and Session 3. In each session, 10 videos were assessed. In Session 1, subjects assessed videos that either had no impairments or had Integrity impairments only. Videos with Integrity impairments consisted of videos that had 1 to 4 temporary interruptions with each interruption lasting for 10 seconds. In Sessions 2 and 3, subjects assessed videos that were a mix of videos with Integrity impairments, videos with Accessibility failure, videos with Retainability failure, and videos with no impairments or failures. In videos with an Accessibility failure, a message indicating inability to play the video is displayed either immediately or after 10 seconds. In videos with a Retainability failure, video playback is permanently interrupted after either 20 seconds or 40 seconds after video start time. Each session comprised approximately 25 minutes of watching videos and answering questions. Session 1 and Session 2 were each followed by an optional break of about 10 minutes.

Part 3 involved answering a questionnaire, the post-questionnaire to obtain information on the subject's general preferences with regards to video quality.

B. Description of Video Database

In this section, we discuss the video content details and the impairments that were evaluated in our experiment. The video databases used in the two experiments (Experiment 1 [7] and Experiment 2) were identical. The video impairments database was created using 30 different videos. Each video

TABLE I
VIDEO QUESTIONNAIRE AND POSSIBLE ANSWERS [7]

No.	Rating Criterion/Question	Possible Ratings/Answers in Experiment 1	Possible Ratings/Answers in Experiment 2
1	Is the technical quality of this video acceptable?	Yes/No	Yes/No
2	Your overall evaluation of the technical quality in the video is:	5-point scale range: Bad (1) to Excellent (5)	6-point scale range: Terrible (0) to Excellent (5)
3	The content of the video was:	5-point scale range: Very boring (1) to Very interesting (5)	5-point scale range: Very boring (1) to Very interesting (5)
4	Your overall viewing experience (Content + Technical Quality) during the video playback was:	5-point scale range: Bad (1) to Excellent (5)	6-point scale range: Terrible (0) to Excellent (5)

had a resolution of 512x288 pixels and a frame-rate of 30 frames-per-second. The videos were between 56 seconds and 123 seconds in length. The 30 video clips consisted of 22 short movie trailers (teaser-trailers) and 8 short movies.

C. Impairment Types and Number of Subjects

For each of the 30 unimpaired videos discussed in Section III-B, additional videos with different impairments were created. Please refer to Table II, that contains information repeated from [7], for a description of these impairments and the corresponding number of impaired videos created for each impairment type. Table II shows that a total of 212 videos were created. Each subject viewed 30 different videos without repetition of content. Each viewed video appeared only once and could have one of the possible impairments shown in Table II. The number of different impairments seen in different sessions is also shown in Table II. From Table II, we can see that 3 unimpaired videos (impairment type I_0), 2 videos with impairment type I_1, 2 videos with impairment type I_2, two videos with impairment type I_3, and one video with impairment type I_4 were evaluated in Session 1. Similarly we can see that videos with Accessibility and Retainability failures were added to Sessions 2 and 3. To prevent two subjects from viewing impairments in the same order, the order in which the impairments were shown was randomized with the exception that, for each subject, the first two videos in Session 1 were I_0 and I_4 respectively, and the first 4 videos in Session 2 were I_0, R_2, I_0, and A_2 respectively. This ordering was purposely chosen to show the best and worst impairment types early in Sessions 1 and 2 [7].

The conditions of participation for each subject 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. In Experiment 2, 16 subjects participated. All subjects were aged over 18 years.

IV. RESULTS

In the discussion of results for Experiment 1 in [7], it was found that a 5-point scale to measure gMOS was not sufficient in the presence of Accessibility and Retainability failures. The use of an expanded scale was suggested. In this section, we discuss the results of using an expanded 6-point scale.

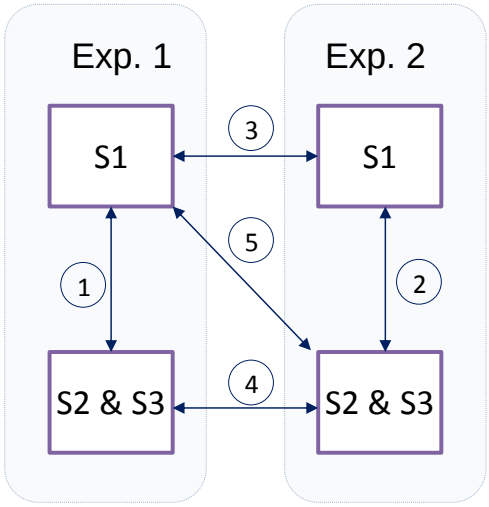

Fig. 1. Summary of Discussion of gMOS Comparisons

Fig. 1 depicts the different comparisons that we discuss in this section. In Fig. 1, "S1" refers to Session 1 and "S2 & S3" refer to Sessions 2 and 3 combined. For example, we will compare Session 1 gMOS of the two experiments. This comparison is labeled "3" in Fig. 1. Comparison "5" is of particular interest, as it involves the assessment of MOS for Integrity impairments only using a 5-point scale and the assessment of gMOS when Accessibility and Retainability failures are included and a 6-point scale is used. In an ideal model, the MOS values for the Integrity impairments would be identical for these two cases. We discuss our findings below.

TABLE II
DESCRIPTION OF VIDEO IMPAIRMENTS AND THEIR DISTRIBUTION ACROSS SESSIONS [7]

Impairment Set	No. of Videos	Description of Impairment	Total No. of Videos for Session 1	Total No. of Videos for Sessions 2 & 3
I_0	30	Unimpaired (Pristine)	3	6
I_1	30	Single temporary interruption of 10s duration happening at 40s (time after video playback start)	2	2
I_2	30	Two 10s (temporary) interruptions happening at 20s and 40s respectively	2	2
I_3	30	Three 10s (temporary) interruptions happening at 10s, 20s, and 40s respectively	2	1
I_4	30	Four 10s (temporary) interruptions happening at 10s, 20s, 30s, and 40s respectively	1	1
R_1	30	A permanent interruption happening at 20s	0	2
R_2	30	A permanent interruption happening at 40s	0	2
A_1	1	Video never starts to play. Video player displays "failure-to-play" message immediately	0	2
A_2	1	Video never starts to play. Video player displays "failure-to-play" message after 10s	0	2

A. Evaluation of Integrity gMOS shift using 5-point and 6-point Scales

As explained in Section II, the introduction of Accessibility and Retainability failures (in Experiment 1) resulted in a shift in the gMOS values for Integrity impairments on a 5-point scale. In this section, we evaluate the gMOS using a 6-point scale and compare the results with our previous results [7].

Recall from our discussion in Section III that Session 1 video set did not include any videos with Accessibility or Retainability failures. On the other hand, Sessions 2 and 3 included videos with Integrity impairments as well as those with Accessibility or Retainability failures. Figs. 2 and 3 show the mean gMOS for each of the impairments across different sessions in Experiments 1 and 2 respectively. The error bars in the bar charts indicate 95% confidence intervals about the mean. Tables III and IV show the results of one-tailed t-tests for Experiments 1 and 2 respectively. The hypothesis tested in these t-tests was whether, for Integrity impairments, Session 1 gMOS was equal to the gMOS in Sessions 2 and 3 (combined). The alternative hypothesis tested was whether Session 1 gMOS is less than gMOS in Sessions 2 and 3. This allows us to check the significance of (an upward) shift in gMOS between Session 1 and Sessions 2 and 3.

As already observed in [7], we can see from Table III that, between Session 1 and Sessions 2 and 3 in Experiment 1, which used a 5-point scale, there is (an upward) shift in gMOS values for I_0, I_2, and I_4. On the other hand, we can see from the data in Table IV, that with the use of 6-point scale in Experiment 2, gMOS values shift upward for I_0 only.

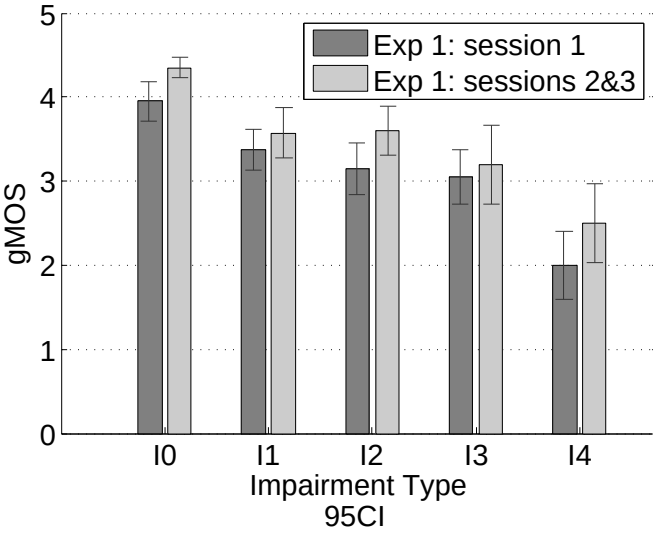

Fig. 2. Variation of gMOS across sessions in Experiment 1

There is no significant shift in the gMOS for I_1, I_3, and I_4. Furthermore, no-significant-shift in the gMOS for I_2 is weakly supported by the t-test results in Table IV. We, therefore, find that with the use of a 6-point scale, there is less (upward) shifting in gMOS values in the presence of Accessibility and Retainability failures. However, a shift is still present and suggests further research in finding an appropriate scale.

TABLE III
SIGNIFICANCE OF CHANGES IN IMPAIRMENT SCORES BETWEEN DIFFERENT SESSIONS IN EXPERIMENT 1

Impairment	I_0	I_1	I_2	I_3	I_4
P-value	0.0014	0.1497	0.0189	0.2964	0.0491

TABLE IV
SIGNIFICANCE OF CHANGES IN IMPAIRMENT SCORES BETWEEN DIFFERENT SESSIONS IN EXPERIMENT 2

Impairment	I_0	I_1	I_2	I_3	I_4
P-value	0.0281	0.1668	0.0578	0.2296	0.1938

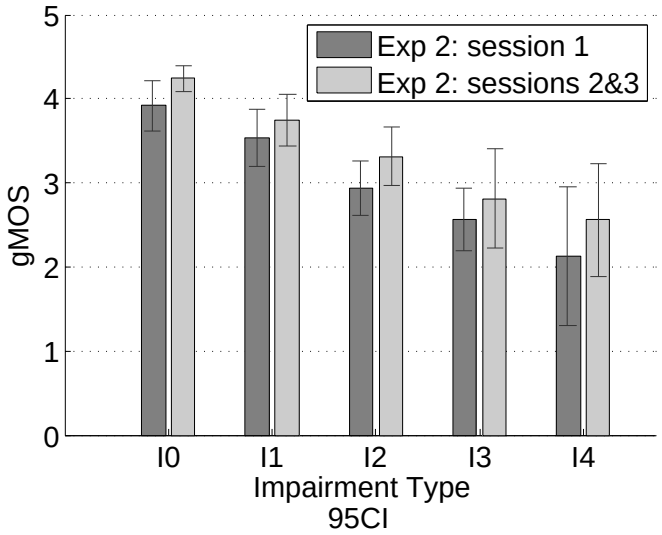

Fig. 3. Variation of gMOS across sessions in Experiment 2

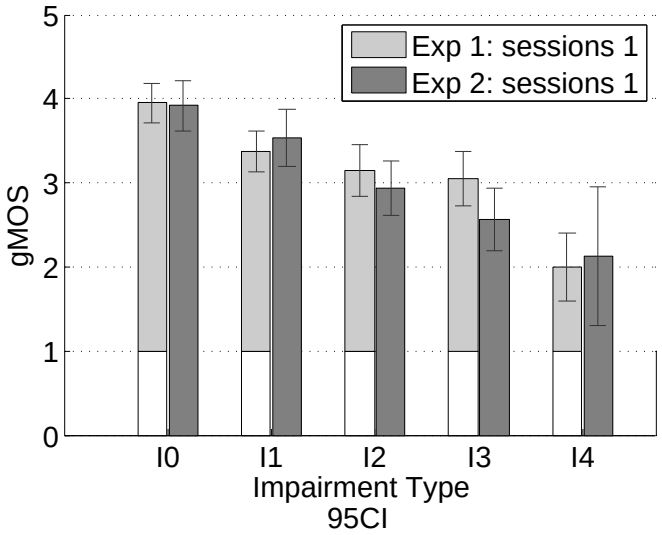

Fig. 4. gMOS for Session 1 in Experiment 1 and Experiment 2

B. Session 1 Integrity gMOS using 5-point and 6-point Scales

In this section, we compare Integrity gMOS values using 5-point and 6-point scales in Session 1 (that is, in the absence of Accessibility and Retainability failures).

Fig. 4 shows the Session 1 Integrity gMOS values (that is, the gMOS values for Integrity impairments) for Experiments 1 and 2. Table V shows the results of two-tailed t-test to find any change in Session 1 Integrity gMOS as a result of change in the rating scale. The data in Table V suggests that, in the absence of Accessibility and Retainability failures, there is no significant change in gMOS values for Integrity impairments as a result of scale change. In particular, based on the t-test results,there is no significant change in gMOS for I_0, I_1, I_2, I_3, and I_4. It should be noted that the result of no-significant-change in gMOS is weakly supported for I_3 in Table V.

C. Session 2 and 3 gMOS using 5-point and 6-point Scales

In this section, we compare the use of 5-point and 6-point scales for video QoE in the presence of Integrity impairments, and Accessibility and Retainability failures (that is, the gMOS evaluations for Sessions 2 and 3 in the two experiments). Fig. 5 shows the bar chart with 95% confidence intervals for gMOS in Sessions 2 and 3 for both experiments. Table VI shows the results of the two-tailed t-tests to check for significance of changes in gMOS between the two experiments for Sessions 2 and 3.

Table VI suggests no statistically significant change in gMOS for Integrity impairments. However, the gMOS values for Accessibility and Retainability failures are significantly different in the two experiments. In particular, as seen in Fig. 5, the gMOS values for Accessibility and Retainability failures are lower in Experiment 2 (on 6-point scale) than the corresponding values obtained in Experiment 1 (on 5-point scale).

It can be seen in Fig. 5 that regardless of the scale used, the Retainability failures are rated lower than Integrity impairments. The Accessibility failures are rated lowest, and close to the bottom of the scale. For Experiment 1, the gMOS values for both types of Accessibility failures, A_1 and A_2 are close to 1, the lowest possible value on the (1 to 5) 5-point scale, whereas for Experiment 2, they are close to 0, the lowest possible value on the (0 to 5) 6-point scale.

An important point to observe from Fig. 5 is that with the 6-point scale, the relative drop in gMOS values from I_4 to R_2 is larger than with the 5-point scale.

TABLE V
SIGNIFICANCE OF CHANGES IN IMPAIRMENT SCORES BETWEEN EXPERIMENT 1 AND EXPERIMENT 2 FOR SESSION 1

Impairment	I_0	I_1	I_2	I_3	I_4
P-value	0.8595	0.4551	0.3447	0.0526	0.7745

TABLE VI
SIGNIFICANCE OF CHANGES IN IMPAIRMENT SCORES BETWEEN EXPERIMENT 1 AND EXPERIMENT 2 FOR SESSIONS 2 AND 3

Impairment	I_0	I_1	I_2	I_3	I_4	R_2	R_1	A_1	A_2
P-value	0.2591	0.4061	0.2059	0.2860	0.8728	0.0123	0.0000	0.0000	0.0000

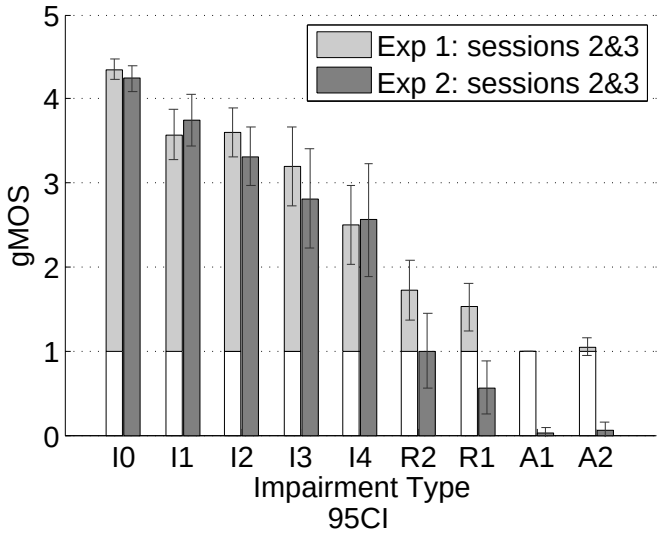

Fig. 5. gMOS for Sessions 1 and 2 in Experiment 1 and Experiment 2

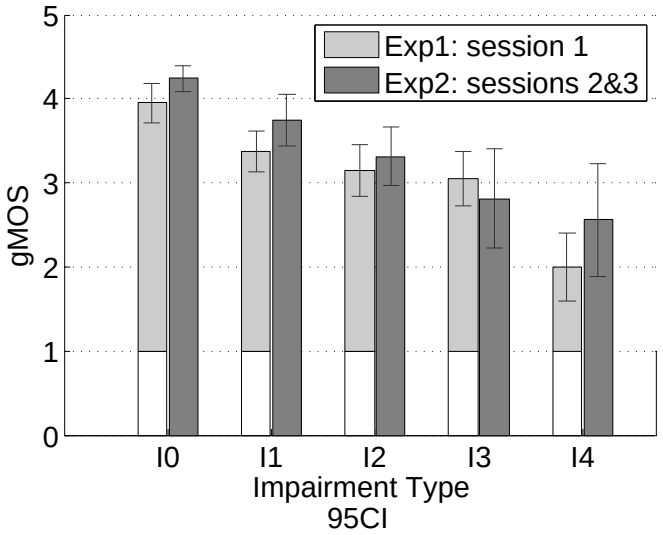

Fig. 6. gMOS for Session 1 in Experiment 1 and Sessions 2 and 3 in Experiment 2

D. Integrity gMOS in Experiment 1 Session 1 and Experiment 2 Sessions 2 and 3

In this section, we evaluate Integrity gMOS on 5-point scale in the absence of Retainability and Accessibility failures and Integrity gMOS on 6-point scale in the presence of Retainability and Accessibility failures. Fig. 6 depicts this comparison. Table VII presents two-tailed t-test results to check for changes in impairment scores across Session 1 in Experiment 1 and Sessions 2 and 3 in Experiment 2.

From the data in Table VII, we can see that the there is change in gMOS values for I_0. On the other hand, there is no significant change in the gMOS values for I_1, I_2, I_3, and I_4. We further note that the no-significant-change deduction for I_1 is weakly supported by the data in Table VII.

TABLE VII
SIGNIFICANCE OF CHANGES IN IMPAIRMENT SCORES BETWEEN EXPERIMENT 1 SESSION 1 AND EXPERIMENT 2 SESSIONS 2 AND 3

Impairment	I_0	I_1	I_2	I_3	I_4
P-value	0.0374	0.0561	0.4823	0.4673	0.1404

V. CONCLUSION

In this paper, we have investigated the impact of rating scale on the assessment of QoE in video sessions that include Integrity impairments and failures in accessibility and retainability. We found that an expanded 6-point rating scale reduces the degree of shift in the Integrity gMOS that was reported in [7]. We found that, regardless of the rating scale used, the Retainability failures are rated lower than Integrity impairments and the Accessibility failures are rated the lowest, close to the bottom of the rating scale used. We also conclude that although the 6-point rating scale results in improving the gMOS shift problem reported in [7], it does not completely eliminate the shifting of gMOS values. More research is needed to investigate the suitability of the use of broader rating scales. Furthermore, the impact of Retainability failures on video QoE needs to be examined in more detail.

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," *MultiMedia, IEEE*, vol. 20, no. 3, pp. 88–94, 2013.

[2] The global internet phenomena report: 2h 2013. Sandvine Incorporated ULC. [Online]. Available: https://www.sandvine.com/downloads/general/global-internet-phenomena/2013/2h-2013-global-internet-phenomena-report.pdf

[3] ITU-T, "Vocabulary for performance and quality of service, amendment 2: New definitions for inclusion in recommendation itu-t p.10/g.100," 2008, recommendation ITU-T P.10/G.100 (2006) Amendment 2.

[4] ——, "Parametric non-intrusive assessment of audiovisual media streaming quality," 2012, recommendation ITU-T P.1201.

[5] ——, "Parametric non-intrusive bitstream assessment of video media streaming quality," 2012, recommendation ITU-T P.1202.

[6] R. F. Baumeister, E. Bratslavsky, C. Finkenauer, and K. D. Vohs, "Bad is stronger than good." *Review of general psychology*, vol. 5, no. 4, p. 323, 2001.

[7] W. Li, D. Kaya, A. Ghayoori, M. Faghani, H.-U. Rehman, M. Chignell, A. Leon-Garcia, J. Jiang, and L. Zucherman, "Impact of accessibility and retainability impairments on the scale to assess video quality of experience," in *6th International Conference on Mobile Networks and Management*, submitted for publication.

[8] L. V. Jones and L. L. Thurstone, "The psychophysics of semantics: an experimental investigation." *Journal of Applied Psychology*, vol. 39, no. 1, p. 31, 1955.

[9] ITU-T, "Subjective video quality assessment methods for multimedia applications," 2008, rec. P. 910.

[10] 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," *Broadcasting, IEEE Transactions on*, vol. 59, no. 1, pp. 47–61, 2013.