Likelihood to Recommend (L2R) Prediction using Quality of Experience (QoE) Measurements: A Longitudinal Study

Amin Azad, Farzad Nejatimoharrami, Mark Chignell
Department of Mechanical and Industrial Engineering, University of Toronto,
Toronto, Ontario, Canada
Emails: amin.azadarmaki@mail.utoronto.ca, farzad.nejatimoharrami@utoronto.ca, chignell@mie.utoronto.ca

Abstract—**Models that predict satisfaction with a service over time need to consider the impact of emotions and remembered quality of experience in creating attitudes towards a service. However, prior research on subjective quality of experience has typically focused on experiments conducted in a single session or over a short period of time. Thus, there is a gap between our understanding of instantaneous quality of experience and long-term judgments, such as overall satisfaction and likelihood to recommend and likelihood to churn. The goal of the study in this paper was to carry out a longitudinal study that would provide initial insights into how experiences of service quality over time are mediated through emotions and memory and accumulated into longer term attitudes about the service. The longitudinal study was carried out over a period of roughly 4 weeks with around 3 sessions per week. A specially constructed online service was used where participants could select YouTube videos to view, and the service would randomly add impairments to the videos before playing back the videos and then asking questions relating to Quality of Experience, Technical Quality and overall frustration and satisfaction. In this paper, we report on the results obtained from the first 8 sessions of data.**

Keywords–Video quality assessment; Quality of Experience (QoE); Comparison of rating scales; subjective evaluation; accessibility; retainability; longitudinal study; consumer satisfaction.

I. INTRODUCTION

In this paper, we focus on the quality of experience associated with streaming online video. While it seems possible that the results obtained here may also apply to other online services our discussion will focus on implications for judgments that are driven by online video experience, since consumption of online video is currently a dominant component of Internet services in terms of bandwidth utilization.

Internet Service Providers (ISPs) are typically private corporations that need to maximize profits by keeping costs to a minimum while preserving, and preferably increasing, their number of customers. However, in a competitive market just maintaining the current customer base can be challenging [1]. Internet service is a commodity and thus quality of service is a key competitive differentiator between ISPs. In this competitive environment, ISPs face a trade-off between the need for high quality service that will attract and retain customers, and the need for efficient use of resources to keep costs from getting out of control [2].

Since it is generally more costly to recruit new customers than to retain existing customers [1], it is believed that the best strategy for ISPs is to try to retain existing customers by heightening customer loyalty and customer value [3].

Quality of Experience (QoE) is an important contributor to the formation of attitudes relating to likelihood to recommend (L2R) or likelihood to churn (L2C) [4][5]. Although there are many different definitions on QoE, Callet et al. [6] explain it as follows: the degree of delight or annoyance of the user of an application or service.

Attitude towards a brand or service accumulates over time and is impacted by the interactions and experiences that a person has with the service. Specific experiences may get priority in terms of the amount of attention or resources that are dedicated to them based on emotional arousal or interest, resulting in stronger memories of those experience. Thus, we should not expect overall attitudes to a service to be based on a simple average of the quality of experience for all the videos viewed, or for all the viewing sessions. Instead, memories of viewing experience may be biased based on the psychology of how people remember experiences.

Since construction of psychological models of how memories of experience are accumulated into overall attitudes is challenging, it is not surprising that marketers typically dispense with this approach and simply ask customers overall questions about how satisfied or unsatisfied they are with their service. However, these judgments may not faithfully represent the true opinion of users [7]. Additionally, after the fact measures of overall attitude are not predictive and do not help ISPs in making decisions in a timely fashion [8]. In an ideal world IPSs might be able to redirect bandwidth to customers who were in danger of forming bad attitudes to the service, but currently ISPs tend to not have the capability to do this type of dynamic reallocation. However, knowing that a customer is probably becoming unhappy with a service might trigger several interventions (e.g., discounts on the monthly bill or other benefits to compensate for problems in service quality) and may also guide the ISP in where to invest in greater bandwidth capacity.

In this research paper, we report on a longitudinal study that looks at how different patterns of service quality affect cumulative experiences and attitudes. Earlier research showed that perceived QoE is affected by the sequence of good and poor videos that are seen within a single viewing session [9]. In the research reported below, we extend this analysis to patterns and variations in QoE occurring over extended periods of time (weeks as against the one-hour duration of a typical experiment).

In Section II, we provide a background on the types of service failures studied, the metrics used to measure the technical quality of the service, and the effects of emotions on memory encoding. In Section III, we discuss the methodology of the experiment, detailing the steps each participant had to go through. Next, in Section IV, we describe the groupings of the participants and the different questionnaire types used within the study. In the results section (Section V), we evaluate the three main hypotheses we have proposed in this research. Lastly, in the discussion section (Section VI), we provide

 ISBN: 978-1-61208-693-4

arguments on why these hypotheses are true and provide recommendations on future research.

II. BACKGROUND

In the context of online video streaming, QoE is influenced by key criteria such as video quality, audio quality, speed of service access, and frequency of service interruption [10]. Two main dimensions of QoE are the Technical Quality (TQ) and Content Quality (CQ) [11].

The only manageable quality from the ISPs perspective is the TQ. ISPs do not have any input into the content that users choose and are not be able to estimate the importance or value of every instance of their service from the point of view of the customer. Thus, the only quality measure they can influence is the TQ [12].

There are two main types of TQ failures when it comes to assessing QoE in a session-oriented setting: Accessibility and Retainability. Furthermore, there is a third TQ problem, referred to as Impairments.

1. Accessibility is the successful starting of the session.
2. Retainability is the capability to continue the session (with or without impairments) until its completion, or until it is interrupted by user action [13].
3. Impairments refer to the degree to which the session unfolds without excessive impairments. In this study, Impairments have multiple levels, suggesting different number and duration for the impairments within the videos watched [14].

The Mean Opinion Score (MOS) is a commonly used metric to measure subjective video technical quality. It was published by the International Telecommunication Union Telecommunication Standardization Section (ITU-T) (2008b) [15]. MOS has a unique rating model which starts from 1 (bad) to 5 (excellent) and since its introduction in 1969, it has been used for the purposes of evaluating radio quality of compressed speech in telecommunications world [16]. In this study, we use MOS scores as a metric to measure customer satisfaction, frustration, content quality and TQ. Two types of MOS scores are gathered in this study: expected MOS and perceived MOS scores which are reflected in terms of TQ.

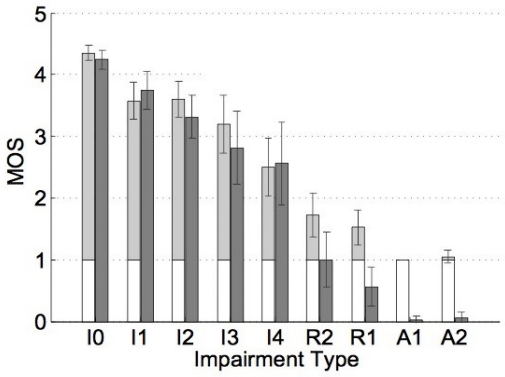

Figure 1. Trends in expected MOS scores measured in a single session setting. Results are calculated based on Li et al. [11].

The expected MOS scores are results based on experimental values gathered from the previous studies conducted by Li et al. [11]. In these studies, the same experiment was set up for only one session and the data gathered demonstrates the TQ levels based on different frequency and types of impairments. Figure 1 shows how individuals reacted to different types of impairments and what was their measures of perceived TQ. Amongst all the factors shown, only I5 was newly introduced, for which we have calculated its expected TQ based on the trend the expected MOS had in Figure 1. The MOS scores used to predict expected MOS score for this experiment are demonstrated in Table I. I0 starts from a 4.5 value instead of 5 since users are hardly giving a perfect score to a service. Also, the MOS goes as low as 1 since this is a 5-point scale model starting at 1 and ending up at 5.

TABLE I. DIFFERENT MOS SCORE USED ACCORDING TO DIFFERENT TYPE OF FREQUENCY OF IMPAIRMENTS AND FAILURES

I0	4.5
I1	3.8
I2	3.5
I3	3.2
I4	2.6
I5	2.2
NR	1.5
NA	1

1) Emotions and Memory Encoding: While emotions are typically of short duration (lasting a few seconds), there is also a cumulative emotional sense which leads to affective memories [17]. Emotional events often attain a privileged status in memory. The more negative an event, the more likely it is to be stored and remembered, and the more details will tend to be retrieved for the event [18]. Slovic described the affect heuristic where people consult their emotions in order to make judgments [19]. Individuals replaced the question of "What do I think about it?" with the question "How do I feel about it?" and answer that latter question since it is easier to recognize the feeling than to apply reason [20].

Aside from judgment based on only emotions, Greenwald [18] researched the possibility of cognitive systems that have two distinct representational stores: implicit and explicit attitude systems. The implicit attitude system is rapid and unconscious while the explicit evaluation is slow and conscious, generating more detailed evaluations [21]. More recently developed models have introduced the idea of multiphase computation of emotions. In this type of computing implicit and explicit appraisals are combined to create a final emotional appraisal that is grounded in previous emotional experiences. Thus, after the affective and cognitive appraisals, retrospective appraisal takes place and affective memories are generated, become available for retrieval and use in future situations [17].

III. METHOD

In designing the current experiment, we were concerned that the data gathered from previous studies did not replicate the real-world usage of mobile devices. Instead, participants were typically given tasks in lab environments that were significantly different from representative real-world scenarios. In order to create a somewhat more realistic experimental context, online software was created where participants could log in from any device and carry out the

 ISBN: 978-1-61208-693-4

experimental tasks at their own convenience. The users were required to complete a video viewing session three times a week for a period of 4 weeks, simulating more closely the experience of browsing videos through different short video streaming platforms, while still maintaining some degree of experimental control. Sessions in this experiment were approximately 30 minutes long. During a session, users were able to search for short videos and the search results were then displayed on the screen. Once a video was chosen from the list of results (the videos shown in the list were filtered to be between 4 to 5 minutes long), experimentally assigned interruption or impairments would be applied to the video and then the video would play on the screen.

After the video was watched, users would be taken to a page where they were asked a few questions based on the video and its technical and content quality. This cycle would repeat until the session was over.

For the purposes of this paper, only data collected up to the end of the eighth session was included in the analyses.

IV. EXPERIMENT DESIGN

Users were divided into 4 main groups. The four groups represented four different sequences of good and bad experiences as demonstrated in Figure 2. According to figure 2, groups 1 and 3 end up on a rising trend of predicted MOS scores as of session 8, while groups 2 and 4 ends up with predicted MOS scores that are trending lower. It can also be seen that groups 1 and 4 start on a decreasing trend whereas groups 2 and 3 start on an increasing trend. Each group had 6-10 participants and the data for each participant was collected for 12 sessions over a period of 4 weeks, however, in this study, we only look at the data from the first 8 sessions. Each participant was asked to log in to the software a minimum of 3 times a week and with a 24-hour gap between sessions with reminders being sent to enforce this requirement. Participants received reminders every other day and those participants who were not able to complete three sessions a week were dropped from the study. At the end of each week, participants were asked specific descriptive questions about how they felt overall about their experience with the service in the past week or past 3 sessions (if fewer than 3 sessions had been carried out within the past week).

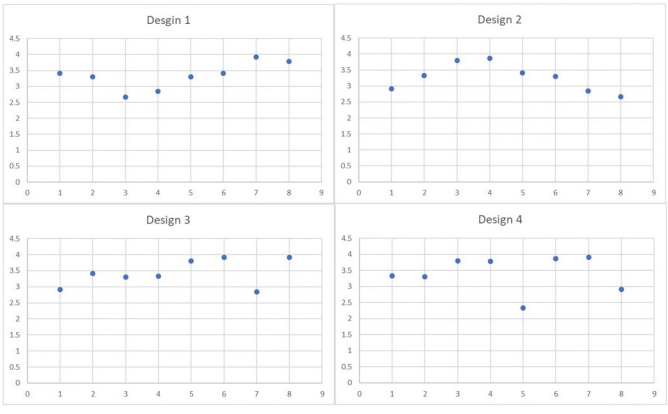

Figure 2. Different trends of experiences participants are exposed to during the 4-week long study.

Figure 2 gives a visual on how the experiment went across all four groups of participants. Each dot represents the average expected TQ rating of the session for that specific group.

Some of the other components of the study include sudden drops in quality in the service (from sessions 6 to 7 for group 3 and from sessions 4 to 5 for group 4, as shown in Figure 2) from one session to the next.

There were four questionnaires administered during the eight sessions. The first questionnaire asked users specific questions on how they felt about a video they had just watched, in terms of their level of satisfaction, frustration and acceptability of the service.

The second questionnaire checked whether the user was paying full attention or not by asking questions concerning details of the contents of the video. It was administered after a randomly selected video (but not including the first video in the selection) in each session. A third questionnaire was administered at the end of each session and asked users about their general feeling about the session they just finished, asking them in detail about their current Likelihood to Recommend (L2R) and Likelihood to Churn (L2C), pricing, devices used and overall satisfaction or frustration level at the end of the session. The fourth and final set of questions asked about how well users remember their experience from the previous session. Items in this final questionnaire also check whether users remember any of the videos and asks for a rating of their overall feeling about their previous session. Using the questionnaire ratings, we analyzed the memory and sequencing effects and compared them to the pricing decisions and the L2R and L2C ratings.

V1	Q1	V2	Q2	V3	Q3	V4	Q2	V5	Q2	V6	Q3	Q4

Figure 3. A visualization of Participant 1, Session 1 demonstrating video and questionnaire types within it.

Figure 3 shows a representation of the videos and questions that a participant was exposed to in a session. The order of the questionnaires was also altered based on the group participants fell into. More specifically, figure 3 illustrates the questionnaire and video orders for session 2 and for group 2 participants. In this model, Q stands for Questionnaire Types and V stands for Videos within a session. Q1 asks about the initial feedback of the user from the video they have just watched and evaluates Technical Quality (TQ), Frustration Level and level of satisfaction on the content of the video (CQ). Q2 is the most common questionnaire and further asks the questions on TQ, CQ and Frustration level while comparing it to the previous video(s) they have watched. Q3 is a surprise questionnaire which asks questions about the content of the video to distract the experimenter from the regular questions they are answering and bring their attention back to the experiment. This questionnaire is asked at different stages across different sessions to make sure the user does not find a pattern for answering it and could always evaluate the users' responsiveness and level of attention according to the answers. Finally, Q4 asks about the overall sessions' acceptability, TQ, frustration level and pricing of services.

 ISBN: 978-1-61208-693-4

V. RESULTS

A. Relationships between ratings

Table II shows the matrix of Pearson correlations between TQ, satisfaction, frustration, content quality, expected MOS, and TQDiff (the difference between the rated TQ and the expected MOS). The content (quality) score was a user rating of how interesting the content of the video was perceived to be. Significant correlations (p¡.05, two-tailed) were found for all possible pairs of TQ, Satisfaction, Frustration, and Content, which is consistent with earlier findings [21].

TABLE II. CORRELATION MATRIX FOR ALL THE VALUES MEASURED WITHIN THE FIRST SURVEY

Correlations

		TQ	Satisf	Frust	Content	Expected MOS	TQ Difference
TQ	Pearson Correlation	1	.920**	.869**	.476**	.376**	.692**
Satisf	Pearson Correlation	.920**	1	.904**	.537**	.310*	.666**
Frust	Pearson Correlation	.869**	.904**	1	.536**	.279*	.639**
Content	Pearson Correlation	.476**	.537**	.536**	1	.047	.433**
Expect MOS	Pearson Correlation	.376**	.310*	.279*	.047	1	-.667**
TQ Diff	Pearson Correlation	.692**	.666**	.639**	.433**	-.667**	1

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Based on the analysis made on differences within TQ amongst all four groups, it was observed that certain groups such as group 1 had a more positive behaviour towards the experiment while group 3 having a more negative impression towards the experiment overall.

This result might be an effect of the trend of the data within the experiment [22]. Groups 2 and 3 are exposed to a more upward trend in data while group 1 and 4 participants are first exposed to a poor TQ then gradually moving towards a higher TQ rating. Based on this observation, it could be interpreted that initial exposure to a service determines a great significance on the followed perceive impressions of its failures. In particular, the better the initial interaction, the more sensitive to the participants are to poor quality and the greater the impacts of trends towards the negative.

B. Carryover effect of good and bad quality

The carryover effect is visible between different sessions of the experiment. For instance, Figure 4 shows that in Design 4, it could be observed that cumulative Expected and Perceived session MOS (SMOS) scores differ from each other, due to the difference in quality of experience that the person had over preceding sessions [23]. This result is very prominent in session 5. The SMOS is relatively low however, the general average of answers is about 2 ratings above the predicted average.

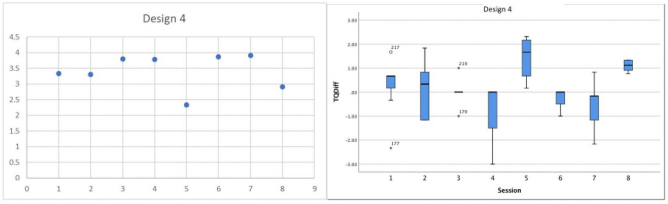

Figure 4. Carry over effect of Design 4.

It could also be observed from Design 4 that the carry-over effect is positive for a positive session and negative for a poor experience. In other words, the slight reduction of SMOS score from session one to session two in the design 4 experiment led to a more biased range of answers but more towards the negative side, however, a slightly more positive experience within session 4 compared to session 3, does not lead to an increase in likely to perceive TQ. Additionally, having the same TQ in two proceeding sessions does not mean that the perceived TQ would be the same. It could be seen from the data that most of the participants have more negative biases, perceiving the TQ as the same or worst if the TQ does not change within two consecutive sessions. Design 4 is one of the four results in the experiment, chosen at random.

C. General Group Behaviour

A significance change in TQ also has its effects, both towards good or bad perception. Figure 5 demonstrates this case by highlighting the differences between sessions 4 and 5 within design 4. The significant drop in TQ in session 5 compared to session 4 is towards the negative but the carry over effect impacts the perceived quality and participants mostly have a more positive impression of the service and are more forgiving of its failures. The same case happens between session 6 and 5 whereas here, session 6 is a significantly more positive experience compared to session 5 but the perceived quality is about the same or less due to session 5s carryover effect.

		Subset	
Group	N	1	2
3	5	-.3683	
2	8	-.2530	
4	8	.1711	.1711
1	8		.4062
Sig.		.069	.680

Figure 5. Difference in TQ of each subgroup within the study by looking at the Homogeneous Subsets of TQ differences.

VI. DISCUSSION

The results obtained from the preliminary analysis of the data from the Introduction Questionnaire (Q1), demonstrate three main findings. First, video Content is a determinant of video quality and could not be neglected. Second, expected and perceived TQ are different in value when we are looking at longer term interactions between service providers and users. Third, in usage of a service, initial interactions and trends of impairments are crucial to the perceived quality.

It has been previously proven that Content, is one of the main influencers of the perceived TQ. Here, we once again prove this relationship in a longitudinal study, demonstrating that aside from the real TQ and the number of impairments in a video, frustration and satisfaction levels are also heavily correlated with the content of the video watched.

In the previous studies, we were able to predict the perceived TQ based on the expected TQ [10]. However, in this study, our results show that the expected TQ to be having a lower correlation with perceived TQ compared to the single session study. One of the primary reasons for these results is possibly the effect of longitudinal study on the attitude towards the service, which is carried over a few sessions. Here, we have the carryover effect of the previous sessions, affecting the perception and attitudes of users towards the service even before starting a new experiment. A participant exposed to a poor service quality in the prior session, is more

 ISBN: 978-1-61208-693-4

likely to perceived and report a poorer TQ for a session compared to a person exposed to a better prior service quality.

Lastly, the sequence that videos appear in is an important influencer of users' ratings of the quality of a service. How participants remember and evaluate the experience as a whole is what will influence their attitudes towards that service. It could be observed that the number of bad videos alone is not sufficient to explain how people retrospectively evaluate their experiences as a whole. The order in which the videos happen, as well are important factors of contributing towards how one perceives the quality of the video. The later the negative effect happens within a session, the more participants would tend to rate the session as a poor session. This result could be also shown from comparing the actual SMOS with the graphed SMOS throughout different sessions across all four groups of the experiment.

In the future steps of this study, the carry-over effect would be measured over different weeks, helping us determine the duration of the positive and negative effects. Moreover, the results of this study could be used to build models that intentionally increase the service quality for some unsatisfied users at the times at which it is seen and predicted that these users are going to make impactful decisions about their service. For instance, increasing the quality of the service for users that are close to cut their service for keeping them as a customer.

ACKNOWLEDGMENT

This research was supported a grant from TELUS and by funding from the NSERC CRD program.

REFERENCES

[1] K. Moon-Koo, M. Park, and D. Jeong, "The Effects of Customer Satisfaction and Switching Barrier on Customer Loyalty in Korean Mobile Telecommunication Services," Telecommunications Policy, vol. 28, no. 2, pp. 145159, 2004.

[2] W. Li, H. Ur-Rehman, M. Chignell, L., Zucherman, and J. Jiang, "Frustration in Response to Impairments and Failures in Online Services, and Resulting Impact on Customer Attitudes," Tenth International Conference on Signal-Image Technology and Internet-Based Systems, pp. 1-7, 2014.

[3] F. Reichheld, "The One Number You Need to Grow," Harvard Business Review, vol. 81, no. 12, pp. 46-54, 2003.

[4] E. Paksoy, A. McCree, and V. Viswanathan, "A Variable-Rate Multimodal Speech Coder with Gain Matched AnalysisBy-Synthesis," IEEE International Conference on Cognitive Informatics, vol. 2, pp. 751754, 1997.

[5] S. Moller and A. Raake, "Quality of experience: Advanced concepts applications and methods," Heidelberg: Springer, pp. 1-8, 2014.

[6] P. L. Callet, S. Moller, A. Perkis, and Qualinet, "White paper on definitions of quality of experience," European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), vol. 1, no. 2, pp. 1-7, 2013.

[7] B. Stauss, M. Schmidt, and A. Schoeler, "Customer frustration in loyalty programs," International Journal of Service Industry Management, vol. 16, no. 3, pp. 229-252, 2005.

[8] M. Chignell, D. Kaya, L. Zucherman, and J. Jiang, "Assessment of Technical Quality of Online Video Using Visualization in Place of Experience," Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 60(1), pp. 2014-2018, 2016.

[9] K. Brunnstrm et al., "Qualinet. White Paper on Definitions of Quality of Experience. Qualinet White Paper on Definitions of Quality of Experience Output,"The fifth Qualinet meetings, 2013.

[10] C. DeGuzman, M. Chignell, J. Jiang, and L. Zucherman, "Testing the effects of peak, end, and linear trend on evaluations of online video quality of experience," In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 2015, vol. 61, no. 1, pp. 813-817, doi: 10.1177/1541931213601696

[11] W. Li, P. Spachos, M. Chignell, A. Leon-Garcia, L. Zucherman, and J. Jiang, "Impact of technical and content quality on overall experience of ott video," 13th IEEE Annual Consumer Communications Networking Conference (CCNC), vol. 13, pp. 930-935, 2016.

[12] T. Hoßfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz, "Quantification of YouTube QoE via Crowdsourcing," IEEE International Symposium on Multimedia, vol. 1, pp. 494-499, 2011.

[13] 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," IEEE International Conference on Communications (ICC), pp. 1-6, 2016.

[14] G. Schroeder, International Telecommunication Union Telecommunications Standardization Sector, vol i-iv, pp. 1-42, 1995.

[15] R. N. Bolton, and J. H. Drew, "A Longitudinal Analysis of the Impact of Service Changes on Customer Attitudes," Journal of Marketing, vol. 55, no. 1, 1991.

[16] Wilson, K. Forbus, and M. McLure, "Am I Really Scared? A Multi-Phase Computational Model of Emotions," Proceedings of the Second Annual Conference on Advances in Cognitive Systems, 2013.

[17] S. K. Labar, C. Roberto, Cognitive neuroscience of emotional memory, Nature Reviews Neuroscience, vol. 7, no. 54, 2006.

[18] E. Kensinger, "Remembering Emotional Experiences: The Contribution of Valence and Arousal," Reviews in the Neurosciences, vol. 15, no. 4, 2004.

[19] P. Slovic, "Trust, Emotion, Sex, Politics, and Science: Surveying the Risk Assessment Battlefield," University of Chicago Legal Forum, pp. 59-100, 1997.

[20] S. G. Bharadwaj, R. Varadarajan, and J. Fahy, "Sustainable Competitive Advantage in Service Industries: A Conceptual Model and Research Propositions," Journal of Marketing, vol. 57, no. 4, pp. 83-99, 1993.

[21] D. Kahneman, "Thinking, fast and slow," New York: Farrar, Straus and Giroux, 2011.

[22] W. Li et al., "Video quality of experience in the presence of accessibility and retainability failures," 10th International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness, Rhodes, pp. 1-7, 2014.

[23] W. Cunningham, and P. D. Zelazo, "Attitudes and evaluations: A social cognitive neuroscience perspective," Trends in Cognitive Sciences, vol. 11, no.3, pp. 97-104, 2007.

 ISBN: 978-1-61208-693-4