|  |  |  |
| --- | --- | --- |
| D:\usr\campos\TSB-Reference\Logos\ITU\sigleITU.gif | INTERNATIONAL TELECOMMUNICATION UNION**TELECOMMUNICATIONSTANDARDIZATION SECTOR**STUDY PERIOD 2017-2020 | SG12-C.n |
| **STUDY GROUP 12** |
| **Original: English** |
| **Question(s):** | Q13 | Krakow, 27-29.11.2017 |
| **CONTRIBUTION** |
| **Source:** | Deutsche Telekom AG |
| **Title:** | Overall Test Plan and Test Factors for P.QUITS |
| **Purpose:** | Discussion |
| **Contact:** | Werner RobitzaDeutsche Telekom AGGermany | Tel: +491604794390E-mail: werner.robitza@telekom.de |
| **Contact:** | Alexander RaakeDeutsche Telekom AG / TU IlmenauGermany | E-mail: alexander.raake@tu-ilmenau.de |

|  |  |
| --- | --- |
| **Keywords:** | P.QUITS |
| **Abstract:** | This document proposes a general structure for the tests conducted in P.QUITS, considering different test factors. |

#### 1. Introduction

P.QUITS aims at developing a “Subjective test methodology for assessing impact of initial loading buffering delay on user experience”. From the A.1 justification, it reads: “The impact of the initial loading delay on the user experience depends on the usage scenario. Therefore, it is important for service providers to assess the impact of initial loading delay on user experience. To do that, the following aspects will be included in this Recommendation: a) Develop subjective test methodology. b) Develop initial loading delay model.”

Another contribution to the Krakow meeting provides a test plan for the first subjective test conducted in the lab, on PC and mobile devices.

This text proposes a general structure for performing tests related to P.QUITS, including recommendations for test factors to be studied and possible biases that need to be addressed. Therefore, the other contribution could be seen as an instantiation of this test plan; this document would serve as a general guideline.

#### 3. Related Work

In the literature, it has been shown that user experience with regard to initial loading delay is impacted by several factors, including human, system, and contextual factors (Reiter 2014). These have been analyzed in various studies, e.g. Dobrian 2013 and Krishnan 2013 for the case of large-scale data from real streaming services. The studies have shown that previous experiences of users (leading to expectations about their Internet performance) heavily influence the engagement patterns. The problem with these studies is that the underlying data is not available and hence does not allow building a model. Also, Mean Opinion Scores are not captured for these types of data.

In the lab, there have been attempts at presenting users with patterns of video delivery degradations and studying their behavioral responses or quality ratings, such as Mok 2011 and Robitza 2016. Mok et al. captured quality rating responses plus user behavior (such as seeking or pausing the video) from the participants, but the paper is not detailed enough regarding the test paradigm description in order to allow reproducing the experiment.

Robitza et al. found that 1/3rd of test participants – when they do not know that initial loading or stalling will happen – will react differently than in normal life, thus not showing the expected behavior. The users had not been informed that problems will occur when attempting to load a video. This test series showed that it is possible to elicit realistic responses from video loading degradations, such as getting people to reload the browser window or to select another video, but a systematic test of different loading patterns will be impossible with this paradigm, without testing a large number of people.

It therefore important to note that in a lab setting, users expect a certain hypothesis from the test they are taking part in, and that the task given to the users will significantly impact their responses in terms of behavior. It can be hypothesized that the same holds true for quality ratings.

#### 3. Acquiring MOS Ratings for Initial Loading Times

In order to build a model that predicts the subject’s Quality of Experience for a video session, collecting Mean Opinion Scores (MOS) seems obvious. This is in line with other models that use the MOS as a simplification for quantifying the user’s experience. However, in practice – and in the technical context of P.QUITS –, the experience of the user might rather translate to a certain corrective action (e.g. reloading the website or leaving the video provider for another one, in the extreme case). In the absence of a framework to predict specific user behavior, we accept this simplification and thus study the impact of initial loading delay on MOS.

However, when gathering MOS from the same user for a series of different initial loading patterns, we expect that users will be influenced by several factors during the test procedure. These factors might bias the ratings, making them either more or less critical when compared to a rating that would be taken in real life, from just *one* occurrence of the same initial loading pattern.

#### 4. Factors Influencing User Ratings

Factors influencing the users’ rating behavior might be one of those listed in Table 1.

| **Factor** | **Comment / Explanation** | **Possible influence on ratings** | **Possible alleviation** |
| --- | --- | --- | --- |
| Intrinsic motivation to watch video | Do users want to watch the upcoming video because they are interested in it or the continuation of the test? | Might yield less critical ratings if users do not want to watch video at all, don’t care about content | Show more interesting content; do not repeat content |
| Extrinsic motivation to watch video | Are users being motivated to watch video? | Might yield less critical ratings if users have no reason to watch videos | Give users a realistic task that relates to video content; do not repeat content |
| Test hypothesis communicated to users | What do subjects think this test is about? (What do the researchers want to know?)Subjects may want to “please” experimenter | Might skew ratings depending on whether subjects want to fulfill this hypothesis | Give written instructions to subjects; be clear about hypothesis |
| Realism of the test environment | How realistic / ecologically valid is the simulated test environment? (e.g. presenting within a real browser vs. just video playback software) | Might skew ratings | Increase realism of test environment platform (software); be clear about simulated physical context (e.g. home)  |
| Assumption of usage context | What are subjects told they should *imagine* they are doing? | Might skew ratings | Be clear about usage context (e.g. Video on Demand vs. Live vs. Duration of content etc.) |

Table 1: Factors influencing user ratings.

#### 5. Test Factors

The following experimental design factors shall be studied within P.QUITS to allow the developed models to predict the impact of different initial loading delays on QoE:

| **Design Factor** | **Levels** | **Tested in P.QUITS / Comments** |
| --- | --- | --- |
| Test environment | Lab | ✓ will be studied first |
| Casual (Office, Café, …)  | May yield more ecologically valid results than lab |
| Home | Through crowd testing |
| Gathering of ratings | Rating interface through dedicated app | ✓ Priority 1 |
| Crowd sourcing from a real service | ✓ Priority 2, may serve as validation of lab results |
| Test platform | Simulated video platform (e.g. dedicated OTT provider) | ✓ |
| Real video platform | May be harder to implement, but expected to lead to more ecologically valid responses |
| Test material type | Entertaining videos including audio | ✓ |
| “Reference” (likely boring) videos without audio | Might lead to lower intrinsic motivation, thus not recommended |
| Test material resolutions | 360p up to 2K |  |
| Initial loading / stalling / quality change patterns | Different lengths of initial loading times | ✓ |
| Different combinations of initial loading plus stalling |  |
| Different combinations of initial loading plus stalling and quality changes |  |
| Device | Mobile phone | ✓ |
| PC | ✓ |
| Viewing distance | Fixed | May create an artificially limited test situation in which users do not behave naturally |
| Flexible / Comfortable | Recommended for viewing when video quality is not of concern |
| Test instructions | Guide the user towards testing / imagining a real VoD portal | Preferable for ecologic validity |
| Guide the user towards a more psychophysical lab test |  |
| Training procedure | Introduction on how to use app and website |  |
| Example clips with different loading patterns |  |
| What is rated? | Quality of initial loading  | Likely most useful for model development |
| Overall video experience | More ecologically valid, but requires more realistic test scenario in order to be valid at all; preferable for tests where quitting is allowed |
| Quality of initial loading and audiovisual content | More in line with previous video quality tests (e.g. P.NATS Phase 1), less focused on overall experience |
| Content preference | Useful in addition to capture possible motivation to watch content |
| Acceptability of the loading time |  |
| Likeliness to quit | Only useful if quitting is not allowed in test |
| Rating scale | ACR | To translate into MOS |
| Degradation-type scale (hinting at acceptability or frustration) |  |
| Binary scale | For acceptability |
| Likelihood scale (Likert-type scale) | E.g. from “very likely” to “very unlikely” |
| Time of rating | A few seconds into the video | Only when initial loading is rated separately |
| At a point defined by the user | Might interrupt video session and natural experience |
| At the end of the video playback (or after aborting) | More in line with previous video quality tests |
| Continuously (using slider) | Useful for time-continuous models |
| Are users allowed to cancel loading? | Yes | More ecologically valid responses, but requires a) implementation in software and b) when presenting long loading conditions too often, users might cancel earlier than in real life |
| No | More in line with previous video quality tests conducted; requires users to imagine how acceptable loading time was |
| If users may cancel loading, are they told that they can do it, or instructed that they should do it? | Yes | Might increase expectancy effects; users more likely to abort than in real environment. Phrasing of instructions will be important. |
| No | May lead to users not knowing that they can abort 🡪 needs to be hinted at by realistic software / testing environment (e.g., known video platform in a browser) |

Table 2: Possible test factors.

#### 6. Possible Timeline

As suggested in “Subjective Test Plan for assessing user experience of initial loading of streaming video (P.QUIT)”, a first test series may focus on the following factors:

* PC and mobile testing
* Interesting audiovisual content
* Different initial loading conditions; no quality changes
* Users are allowed to cancel initial loading
* ACR-type rating of the initial loading experience

In a next series of tests, the test paradigm itself could be varied, e.g. by changing the interactivity possibilities of the user (e.g. allowing more interactions, or not allowing any interactions at all, thus making the test more “passive”).

Concrete steps are to be discussed at the Interim meeting; this document shall then be extended with a detailed timeline. Additional factors and their possible consequences may also be added to Table 2.

#### References

Reiter, U., Brunnström, K., De Moor, K., Larabi, M.-C., Pereira, M., Pinheiro, A., … Zgank, A. (2014). Factors influencing Quality of Experience. In Quality of Experience: Advanced Concepts, Applications and Methods (pp. 55--72). Springer.

Dobrian, F., Awan, A., Joseph, D., & Ganjam, A. (2013). Understanding the impact of video quality on user engagement. Communications of the ACM, 56(3), 91–99.

Krishnan, S. S., & Sitaraman, R. K. (2013). Video stream quality impacts viewer behavior: Inferring causality using quasi-experimental designs. IEEE/ACM Transactions on Networking, 21(6), 2001–2014. <https://doi.org/10.1109/TNET.2013.2281542>

Robitza, W., & Raake, A. (2016). (Re-)Actions Speak Louder Than Words? A Novel Test Method for Tracking User Behavior in Web Video Services. In Eighth International Workshop on Quality of Multimedia Experience (QoMEX). Lisbon.

Mok, R. K. P., Chan, E. W. W., Luo, X., & Chang, R. K. C. (2011). Inferring the QoE of HTTP video streaming from user-viewing activities. In Proceedings of the first ACM SIGCOMM workshop on Measurements up the stack - W-MUST ’11 (p. 31). <https://doi.org/10.1145/2018602.2018611>