International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.917

(01/2019)

SERIES P: TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS

Audiovisual quality in multimedia services

Subjective test methodology for assessing impact of initial loading delay on quality of experience

Recommendation ITU-T P.917

International Telecommunication Union

ITU-T P-SERIES RECOMMENDATIONS

TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS

Vocabulary and effects of transmission parameters on customer opinion of transmission quality	Series	P.10
Voice terminal characteristics	Series	P.30
		P.300
Reference systems	Series	P.40
Objective measuring apparatus	Series	P.50
		P.500
Objective electro-acoustical measurements	Series	P.60
Measurements related to speech loudness	Series	P.70
Methods for objective and subjective assessment of speech quality	Series	P.80
Methods for objective and subjective assessment of speech and video quality	Series	P.800
Audiovisual quality in multimedia services	**Series**	**P.900**
Transmission performance and QoS aspects of IP end-points	Series	P.1000
Communications involving vehicles	Series	P.1100
Models and tools for quality assessment of streamed media	Series	P.1200
Telemeeting assessment	Series	P.1300
Statistical analysis, evaluation and reporting guidelines of quality measurements	Series	P.1400
Methods for objective and subjective assessment of quality of services other than speech and video	Series	P.1500

For further details, please refer to the list of ITU-T Recommendations.

Recommendation ITU-T P.917

Subjective test methodology for assessing impact of initial loading delay on quality of experience

Summary

Recommendation ITU-T P.917 defines a procedure for conducting behavioral studies targeted at investigating video streaming performance and its relation to users' quality of experience (QoE). The studies are to be conducted in controlled environments. Subjects are exposed to different initial loading delay conditions and other quality degradations typical for video streaming, and subjects may be asked to rate audiovisual quality or their experience. Additionally, their behaviour as a response to long loading times (i.e., aborting the video playback during its loading phase) may be investigated.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T P.917	2019-01-13	12	11.1002/1000/13843

Keywords

Initial loading delay, quality of experience, video streaming.

* To access the Recommendation, type the URL http://handle.itu.int/ in the address field of your web browser, followed by the Recommendation's unique ID. For example, http://handle.itu.int/11.1002/1000/11830-en.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure, e.g., interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, ITU had not received notice of intellectual property, protected by patents, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database at http://www.itu.int/ITU-T/ipr/.

Table of Contents

Introduction

Initial loading delay during video streaming is known to impact the quality of experience (QoE) of end-users. However, the specific relationship between initial loading delay and a user's likelihood of aborting a playback – or the perceived overall quality of the initial loading delay and/or entire media session – varies for different services. Furthermore, it depends on a user's expectations. It is important for Internet service providers (ISPs) to be able to assess the impact of initial loading delay on users' QoE, as numbers quantifying these relationships are not readily available.

In this Recommendation, a procedure for conducting studies with human test subjects is described. These studies are called behavioral, since interaction with a (simulated) video streaming service is the main focus. Results from a study conducted following this Recommendation may be used for several purposes, including but not limited to developing prediction models that estimate:

– the likelihood of users abandoning a video playback during its initial loading phase; and/or

– the perceived overall quality of a video playback impacted by initial loading, stalling, and audiovisual quality changes; and/or

– the perceived quality of the initial loading experience.

The test procedures specified in this Recommendation aim to bridge the gap between traditional video quality assessment methods such as those specified in ITU-T P.910 and large-scale data collections that over-the-top (OTT) video providers may conduct.

First, background and related work on tests aimed at assessing the impact of initial loading delay are described. A general overview of factors that influence user behaviour and subsequent ratings is given. Then, this Recommendation, ITU-T P.917, describes a test procedure, which is to be conducted in a controlled environment.

Recommendation ITU-T P.917

Subjective test methodology for assessing impact of initial loading delay on quality of experience

1 Scope

This Recommendation describes a test methodology for use in controlled environments with human subjects. The scope of this Recommendation is currently limited to the following factors:

– test environment: controlled environment (see [ITU-T P.913], clause 8.1);

– video length: short videos (< 5 mins);

– simulated platform: video on demand (VOD);

– test devices: PCs or mobile phones.

The technical scope is limited to progressive download streaming and adaptive streaming type services (using reliable transport), which includes:

– over-the-top (OTT) services, as well as operator managed video services (over TCP);

– video over both mobile and fixed connections;

– the protocols HTTP/TCP/IP, RTMP/TCP/IP, HLS/HTTP/TCP/IP, and DASH/HTTP/TCP/IP. This Recommendation is agnostic to the specific network delivery method (HTTP or DASH or other), with one exception that it assumes reliable delivery (TCP/IP) that results in apparent loading times (indicated with a "buffering" symbol to the user);

– video services typically using container formats such as Flash (FLV), MP4, WebM, 3GP, and MPEG2-TS. This test method is agnostic to the type of container format.

The methods described in this Recommendation may be used in other technical contexts where users browse content and have to wait for media playback (e.g., music streaming), but it has not been validated in such contexts.

2 References

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

[ITU-T P.10] Recommendation ITU-T P.10/G.100 (2017), *Vocabulary for performance, quality of service and quality of experience*.

[ITU-T P.910] Recommendation ITU-T P.910 (2008), *Subjective video quality assessment methods for multimedia applications.*

[ITU-T P.913] Recommendation ITU-T P.913 (2016), *Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment*.

[ITU-T P.1203] Recommendation ITU-T P.1203 (2017), *Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport*.

[ITU-T P.1203.1] Recommendation ITU-T P.1203.1 (2019), *Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport – Video quality estimation module*.

3 Definitions

3.1 Terms defined elsewhere

This Recommendation uses the following terms defined elsewhere:

3.1.1 initial loading delay [ITU-T P.1203].

3.1.2 stalling [ITU-T P.1203].

3.1.3 integral quality [ITU-T P.1203].

3.1.4 quality of experience [ITU-T P.10].

3.2 Terms defined in this Recommendation

This Recommendation defines the following term:

3.2.1 overall quality: Synonym for integral quality as defined in clause 3.1.3.

3.2.2 quality of initial loading experience: The degree to which a user's expectations of the initial loading delay are fulfilled.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

ACR Absolute Category Rating

FHD Full High-Definition

HD High Definition

HRC Hypothetical Reference Circuit

ILQ Initial Loading Quality

MOS Mean Opinion Score

OTT Over-The-Top

PVS Processed Video Sequence

QHD Quad HD (2560 × 1440)

QoE Quality of Experience

SRC Source

VOD Video on Demand

5 Conventions

None.

6 Background

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 (for factors influencing quality of experience (QoE), see [b-Reiter]). These have been analyzed in various studies, such as [b-Dobrian]

and [b-Krishnan] 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 their engagement patterns. The problem with these studies is that the underlying data are not available and hence do not allow building a model. Also, mean opinion scores (MOSs) are not captured for these types of data.

In a laboratory context, there have been attempts at presenting users with patterns of video delivery degradations and studying their behavioral responses or quality ratings, such as [b-Mok] and [b-Robitza]. [b-Mok] captured quality rating responses plus user behaviour (such as seeking or pausing the video) from participants, but the paper is not detailed enough regarding the test paradigm description in order to allow reproducing the experiment.

[b-Robitza] found that a third 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 behaviour. 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.

To develop a model that predicts a subject's initial loading experience for a video session, a MOS methodology may be used. This is in line with other models that use the MOS as a simplification for quantifying the user's QoE, without giving the user any other means to indicate a bad experience. However, in practice, the experience of the user may translate to a certain corrective action (e.g., cancelling the video playback). Hence, the option of quitting the video loading process will be given to the test subjects.

7 Factors influencing user behaviour and ratings

When gathering MOS or assessing user abandonment, it is expected that users will be influenced by several factors during the test. These factors may bias the ratings or user behaviour, making ratings either more or less critical when compared to a rating that would be taken in real-life, or leading to users not behaving as they would normally do, for example at home. Table 1 lists factors influencing a user's rating.

Table 1 – Factors influencing ratings

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 the watch video and do not care about its content	Show more interesting content; do not repeat content
Extrinsic motivation to watch video	Are users being motivated to watch the video?	Might yield less critical ratings if users have no reason to watch the video	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

Table 1 – Factors influencing ratings

Factor	Comment / Explanation	Possible influence on ratings	Possible alleviation
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., VOD vs. live vs. duration of content)

The following clause describes a method to gather user ratings and observe user behaviour (e.g., quitting during video loading) in a laboratory context. Where indicated, modifications to the method are allowed. If deviations from the method are made, they must be properly documented.

7.1 Test platform

The test must be conducted on a platform that simulates a real video platform with which users may already be familiar. It should therefore use similar navigation patterns and have similar performance. This is expected to yield more realistic ratings and behaviour.

The test platform may be a PC-based or mobile phone-based platform. It must be stated that different browsing and interaction patterns are commonly used for such platforms.

The user interface of both platforms should have two key components: a navigation bar and a video grid. Through the navigation bar, subjects can browse different video categories, and the related videos will be shown on the video grid as a list of thumbnails.

When subjects select a specific video by clicking its thumbnail, the video will start to play after an initial loading period. The subject may be allowed to abort the video playback at any time during the initial loading period, through either a dedicated button in the interface (e.g., an *X* button), or a hardware button on the device (e.g., a "back" button).

7.2 Source content

7.2.1 Genres

The audiovisual material sources (SRCs) must be representative of typical video clips on major VoD websites. The SRCs may, for example, belong to any of the following genres:

– movies, movie trailers;
– TV shows;
– sports;
– music videos;
– animation;
– news;
– documentaries;
– comedies.

The test platform may expose these genres as navigation categories to enable subjects to browse through the content based on their interests.

7.2.2 Characteristics and length

The SRCs must provide well-synchronized audio and video content. Lip-synchronicity must be guaranteed for non-dubbed video material and must be good enough for dubbed video material. SRCs should also provide a natural combination of the presented audio and video signal content. Dubbed video material is allowed in countries where users typically watch dubbed video.

It is recommended to use interesting and/or engaging SRCs, such that subjects will be presented videos that they will be motivated to watch. This is expected to increase the validity of the captured behaviour and ratings.

For the scope of the presented test methodology, the duration of each video sample should be less than three mins.

7.3 Test conditions

During the test procedure, subjects may be shown different audiovisual quality and different combinations of initial loading delays and stalling with varying length and frequency of the latter. A particular instantiation of initial loading delay, stalling, and audiovisual quality profile is called a hypothetical reference circuit (HRC).

7.3.1 Types of degradations

The audiovisual quality may be intentionally degraded using different video codecs, and/or by reducing video resolution, bitrate, and/or framerate, as typically done for HTTP adaptive streaming services.

Within an HRC, audiovisual quality may change (quality switching) to simulate network bandwidth fluctuations. Typical quality switching patterns may be included in the HRC design, such as ramp-ups (switching from low to high quality after initial loading).

Initial loading delays may vary between short (0.1 seconds) and long (60 seconds).

Stalling events may vary between short (0.1 seconds) and long (60 seconds). There is no limit to the number of stalling events, but they should be kept reasonable.

7.3.2 Resolutions, codecs and codec settings

The encoding schemes of video streaming providers differ depending on the type of streamed video and the video codecs being used. The particular choice of video codec and encoding settings will therefore depend on the application under study.

Examples of common video coding resolutions and video bitrate ranges can be found in [ITU-T P.1203.1].

The lowest and highest video resolutions used in a test may differ between mobile and PC platforms, respectively. For example, a mobile phone may not be able to display 2160p video.

7.3.3 Reference conditions

Table 2 lists possible reference initial loading and video coding conditions (HRCs) to be shown in the test. These should be shown to every subject in order to produce anchors for high, medium and low quality ratings, respectively.

Table 2 – Test conditions

Anchor	Initial loading time	Quality switch pattern
High-quality reference	Very short	Constant high resolution
Medium-quality reference	Short, but noticeable	Constant medium resolution
Low-quality reference	Long	Constant low resolution

7.4 Assignment of SRCs to HRCs

To produce processed video sequences (PVSs), a SRC must be paired with an HRC. For a complete test, it is recommended that all HRCs are shown at least once to every subject.

To prevent subject fatigue or boredom, a SRC should not be shown twice during a test. Once a SRC has been chosen, it shall be removed from the list of available SRCs and shall not be played again during the same visit time of the subject. An already played SRC may be greyed out in the user interface to prevent subjects from selecting it again.

There are two general approaches for associating SRCs with HRCs:

1) When a SRC is selected by the subject, an HRC is selected randomly from the entire list of HRCs. The HRC will not be shown again (sampling without replacement).
2) The pairing of SRCs and HRCs is determined before the test, and is the same for all subjects. When an SRC is selected by the subject, the matching HRC is selected based on the pairing.

Each HRC may be shown more than once during a test, in order to increase the number of ratings and thereby also increase the reliability in determining the impact of the HRC on the measured variables. In this case, the HRC can be sampled more than once in option 1 above.

7.5 Overall test procedure

The overall test consists of several steps (see Figure 1).

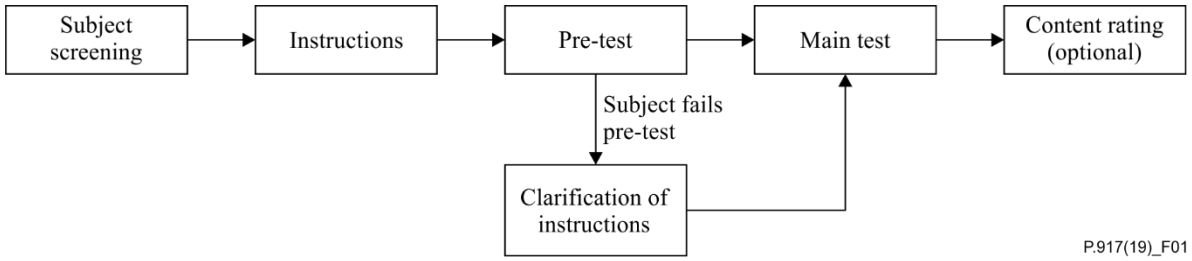

Figure 1 – Overall test procedure

7.5.1 Subject screening

Subjects may be screened before the test; see clause 7.7.2.

7.5.2 Instructions

Subjects must be given instructions (see Appendix I) in written form, which explain the overall test procedure including the training, pre-test, and main test procedure.

7.5.3 Length of test and sessions

For a single test, the time of actively viewing videos or hearing audio and voting should be limited to a maximum of 60 mins. The total visit time, including instructions, training, main test, and content rating procedure, should be limited to 1 hr. 45 mins.

Each test may be split into 2-3 sessions between which subjects have a short break (5-10 mins). A single session should last no longer than 30 mins.

7.5.4 Training procedure

The first session of each visit includes a training phase. This phase consists of showing the subject(s) how to use the test platform, which includes getting familiar with the user interface of the video platform, browsing and playing video contents, showing the rating screen, the types of video material, initial loading delay, stalling, and the quality range of the video sequences.

The training procedure consists of the same steps as the main test procedure (see below).

As an example, in the training phase, a total of four videos may be viewed by the subject. Each video may be shown with one of the following conditions:

1) short initial loading time (< 1 s), good video quality (e.g., 720p or higher);
2) medium initial loading time (< 10 s), medium video quality;
3) medium initial loading time (< 10 s), bad video quality;
4) long initial loading time (1 min), bad video quality.

Unlike the main test procedure, during training, the video playback may end after 30 s, after which subjects are shown the rating screen.

When a video with an extremely long initial loading delay is shown, if the subject does not abort the playback, it can be assumed that the subject has misunderstood the instructions or goals of the test procedure. In this case, further clarification of the instructions should be given to the subject after the training phase has been completed.

7.5.5 Main test procedure

The main test procedure consists of several steps. The user interface must alternate between showing the video overview, the video initial loading period, the video playing, and a rating screen.

7.5.5.1 Video overview/browsing

The subject browses the video platform and finds a video they would like to view, optionally by filtering different video categories. Once the subject finds a video, they can click on the video thumbnail to start the playback.

7.5.5.2 Initial loading period

The video starts to load according to the initial loading delay of the HRC chosen in the test design. During the loading period, if the subject feels the loading delay is too long, they may abort and return to the overview.

7.5.5.3 Video playing

The video starts to play with the audiovisual quality according to the HRC chosen in the test design, optionally interrupted by stalling events.

7.5.5.4 Rating schemes

There are two rating elements, from which either one or both may be shown (see details below on allowed rating questions in each element):

1) rating element 1: Directly after a video has started playing, the subject is asked to provide ratings. This captures the subject's opinion directly after a possibly good or bad loading experience.
2) rating element 2: After the video playback ends, the subject is asked to provide ratings. This captures the subject's opinion after the experience of watching the entire video.

There are three rating schemes, depending on the choice of rating elements:

1) rating scheme 1: Rating element 1 is shown.

2) rating scheme 2: Rating element 2 is shown.

3) rating scheme 3: Rating element 1 and 2 are shown.

The choice of rating scheme(s) and included rating questions depend on the overall aim of the test, and the particular kind of model that may be developed based on the subjective test results.

On one hand, rating element 1 aims at capturing a subject's experience immediately after starting the playback; it may therefore yield more critical ratings than rating element 2. The use of more critical ratings may be of importance to operators who want to optimize for a quick loading experience. On the other hand, rating element 2 resembles the remembered quality of an entire session and is therefore comparable to the subjective experiment design as used in, for example, [ITU-T P.1203]. Rating scheme 3 combines both rating elements, but may result in more effort for the subject or more attention to the ratings themselves, thereby potentially introducing bias.

For the proposed rating schemes, see Figures 2, 3 and 4.

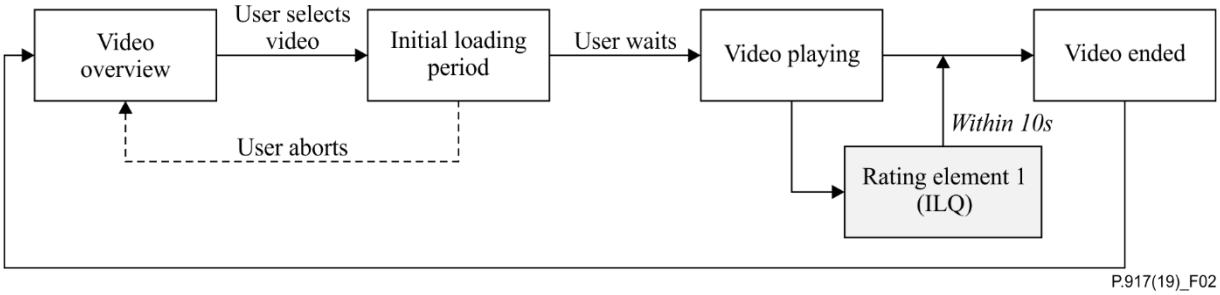

Figure 2 – Rating scheme 1 of main test procedure

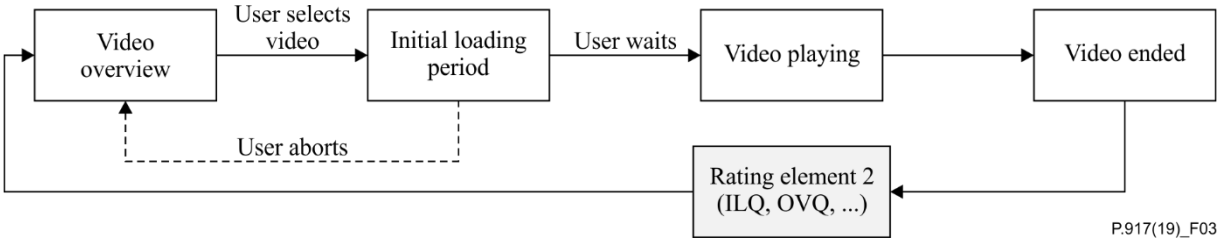

Figure 3 – Rating scheme 2 of main test procedure

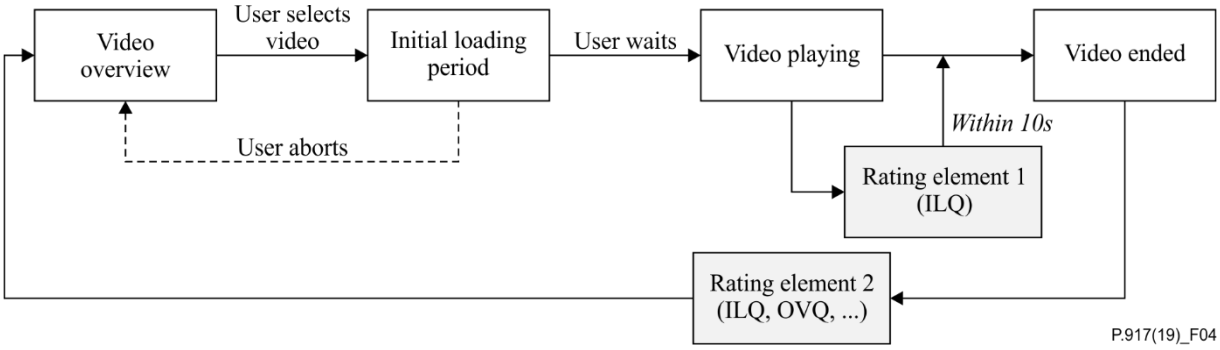

Figure 4 – Rating scheme 3 of main test procedure

7.5.5.4.1 Required rating questions

At least one of the following questions must be asked:

– Rating of initial loading quality (ILQ):

  • Applies to rating element 1 and 2.

- Question: *How was the quality of your initial loading experience?* (see definition of "quality of initial loading experience").
- Possible answers: absolute category rating (ACR) scale (1-5, including labels).
- If the question is asked in rating element 1, the playback must not be interrupted by a rating screen shown in the interface, and subjects should give their scores verbally rather than writing down the scores on a sheet of paper.

– Rating of overall quality (OVQ):
  - Applies to rating element 2.
  - Question: *What is your opinion of the overall quality?*
  - Possible answers: ACR scale (1-5, including labels).

7.5.5.4.2 Optional rating questions

The following additional (optional) questions may be asked. An example of optional questions is given below:

– Audiovisual quality (AVQ)
  - Applies to rating element 2 only.
  - Question: *What was the quality of the audio and video?*
  - Answers: ACR scale (1-5, including labels).

– Loading acceptability (LA)
  - Applies to rating elements 1 and 2.
  - Question: *Was the initial loading delay acceptable?*
  - Answers: Binary (acceptable, unacceptable).

Additional optional questions may be specified in a future appendix to this Recommendation.

An example of the ACR scale is given in Figure 5.

5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

Figure 5 – Rating scale (numerical values will be used in addition to the labels)

ACR labels "excellent" to "bad" may be translated to the main language of a country in which the test is conducted. The numbers must be shown next to the scale labels.

After rating, the subject can continue with Step 1, unless there are no more videos to be rated.

7.5.5.5 Content rating procedure

After completing the main test, subjects may be shown a list of source sequences they have seen, together with questions about the content. For example, subjects may be asked whether they liked a particular content, and may judge their opinion on a particular scale.

Examples of such questions are given below:

– *What were your expectations of the video content before watching it?*

– *How much did you like the content after having watched it?*

– Answers: ACR scale (1-5, including labels), or 5-star rating.

7.5.5.6 Optional questionnaire

An optional questionnaire may be shown to a subject, which aims at gathering background information about their typical video streaming service or Internet usage. Appendix II provides an example of such a questionnaire.

7.6 Subjective test environment and set-up

Two types of subjective tests may be carried out, either using a PC (e.g., computer playback) or mobile equipment (e.g., mobile phone).

7.6.1 Common properties

The following conditions must be met for both PC and mobile tests.

It should be ensured that:

– playback mechanism is guaranteed to play at frame rate without dropping frames;

– playback mechanism does not add visible artifacts.

The tests must be conducted indoors. See the following clauses for detailed requirements.

7.6.2 PC tests

PC tests will be conducted using a computer where test sequences are loaded from a hard disk and presented on a computer monitor.

7.6.2.1 Test environment

The test room shall conform to the requirements set out for controlled environments according to clause 8.1 in [ITU-T P.913].

7.6.2.2 Display specification and set-up

The monitor must support at least full high-definition (full-HD) resolution (1920 × 1080).

Video sequences may be played in full-screen mode, or in a smaller window, where they may be surrounded by user interface (UI) elements of the simulated video platform. If played in full-screen mode, the monitor must have an equal to or higher resolution than the highest resolution that is used in the test's HRCs.

The display shall be set up using the following procedure:

– use the auto setting to set the default values for luminance, contrast and color shade of white;

– adjust the brightness according to [ITU-T P.910], but do not adjust the contrast (it might change balance of the color temperature);

– set the gamma to 2.2;

– set the color temperature to 6500 K.

Any post-processing done by the monitor (e.g., frame interpolation, motion smoothing) must be deactivated.

7.6.2.3 Viewing distance

The instructions given to subjects will request that they maintain a specified viewing distance range from the display device, but overall maintain a comfortable viewing distance of their own choice.

The viewing distance range should correspond to about 1 min of arc. For example, this is equal to about 1.5 H for UHD displays, or 3 H for full-HD displays, where H = height of the visible part of the display.

7.6.2.4 Viewing conditions

Only one subject should be seated in front of the viewing device.

The test room will conform to requirements specified in clause 7.6.2.3.

It is recommended that subjects be seated facing the center of the video display at the specified viewing distance.

7.6.2.5 Listening conditions

Audio will be presented using headphones or speakers.

When listening is carried out with headphones, audio will be played using a diotic presentation (both ears receive the same mono signal) or binaural presentation (each ear receives one channel of a stereo signal). Headphones should be diffuse-field equalized headphones.

Artificial background noise (e.g., Hoth noise) will not be used.

Presentation (listening) level should be 73 dB sound pressure level (SPL) at both ears when using headphones.

7.6.3 Mobile equipment tests

Mobile equipment audiovisual tests will be conducted using a mobile phone where test sequences are loaded from the phone's internal memory.

The voting questions may be presented by one of the following means:

– on the phone display;

– on a paper-based questionnaire.

7.6.3.1 Test environment

The test room shall conform to the same specifications as in clause 7.6.2.1.

7.6.3.2 Device specification and set-up

The test will be carried out on mobile phones. The devices may have specifications as given in the following table.

Table 3 – Example mobile device specifications

Device feature	Device 1	Device 2
Diagonal display size	5.1"	5.7"
Display resolution	2560 × 1440	2560 × 1440
Display type	LCD	AMOLED

The phone should be charged during the session to avoid depleting the battery.

7.6.3.3 Viewing distance

The viewing distance should be between 6-8 H, where H = display height (in landscape mode), according to the subject’s preference.

7.6.3.4 Viewing conditions

Only one subject should be seated in front of the viewing device.

The device should be mounted in landscape mode on a desk or wall. This will also help avoid fatigue. Cycle mounts are useful for this and can be mounted to an adjustable boom mic stand. Car mounts can be fixed on the table as well

To minimize dirt and grease marks on the screen the screen should be wiped before each subject starts their test and during each break.

Subjects will be seated facing the device at a horizontal angle between 0 and 45°.

7.6.3.5 Device settings

Mobile devices may have default options turned on that can interfere with the test procedure. The following settings may require changes from factory defaults:

– brightness: The device display must be set to the highest brightness level, or a level that is comfortable in the test environment. Automatic brightness correction must be disabled.
– power saving: Any power or battery saving mode must be turned off.
– display lock: The display must not lock automatically. Any screensavers must be disabled.
– screen mode: If the device offers screen or color enhancement modes, these must be turned off or set to "standard". Depending on the device, such enhancement modes may be called "adaptive display", "dynamic", "professional", "photo", or "cinema".
– notifications: Notifications from applications on the phone must be disabled.

7.6.3.6 Listening conditions

See clause 7.6.2.5.

An audio extension lead may be used to avoid a heavy headphone connection directly into the phone.

7.7 Subjects and subjective test control

7.7.1 Number of subjects

At least 50 subjects should participate in each test. For a pilot-test, no fewer than 25 subjects should be used.

Subjects who have failed the pre-test will be allowed to continue, but their ratings must be checked for reliability.

It is recommended to have a 50-50 split or near 50-50 split between female and male subjects. If the parity between male and female participants cannot be achieved, then a maximum of 60-40 split is permitted.

It is recommended to include subjects from different socioeconomic backgrounds.

Only non-expert subjects should participate. The term non-expert is used in the sense that a subject's profession does not involve audio or picture quality and they are not experienced assessors. They must not have participated in a subjective quality test over a period of six months.

7.7.2 Subject screening

Prior to participation in a video test, subjects must be screened for the following:

– severe visual impairments;
– color blindness;
– failure to comprehend the written instructions (e.g., due to language deficiencies);
– hearing loss.

7.7.3 General instructions for subjects

A set of instructions that the test subject will have to read must be written down. The instructions must clearly explain why the test is being run, what the subject will see/hear, and what the subject should do. The instructions should be tested with non-experts to make sure they are clear, and must be revised as necessary.

The experimenter must ask the subject whether they have understood the instructions and clarify any remaining questions.

Detailed example instructions to subjects are provided in Appendix I.

Appendix I

Example instructions to subjects

(This appendix does not form an integral part of this Recommendation.)

These instructions must be handed to subjects in written form. They must be translated to the local language of the lab in which the tests are being conducted.

Introduction: Thanks for coming in today to participate in our study. The study is about a video-on-demand service; it is being sponsored and conducted by companies that are developing and testing new technologies to enhance consumers' online video experience. These companies are interested in your view on the overall quality.

Screening: Please indicate if you have any problems seeing (including color blindness), hearing, or understanding *(the language in which the test is performed).*

Overall goal: We are going to ask you to browse a video-on-demand platform, select the videos you are interested in, and start to watch. Afterwards, we will ask you to judge each video's loading time and quality – we will explain more below about what we mean by that. It takes time to load these videos; sometimes there may be loading problems, as you probably have experienced in real-life. In case of loading problems, if the loading feels too long for you, you can abort the video playback *(Editor's note: Insert specific instructions which button or UI element to press to abort)* and select another video.

Set-up:

– When we get started with the study, please sit at *(location).* The video-on-demand service will be displayed on the *(screen/phone).*

– Please stay close to the position indicated by the mark. This is because the videos might look a little different from different positions, and we would like everyone to judge the videos from about the same position. Feel comfortable to move around a little bit though, if necessary.

– (*Optional for mobile tests*) Please do not press any of the physical buttons on the phone. This may turn off the phone display or exit the test, making your ratings invalid.

Process:

– You will first see an overview of different video genres and thumbnails for each video that you can select.

– Please choose a video you would like to watch, then select the video by clicking on its thumbnail.

– Each video will be a few minutes long.

– Rating *(optional)*: As soon as the video has started playing, please rate the quality of your initial loading experience on the following scale: (*Picture of ACR scale).* Your task is to judge how well the delay you experienced between starting the selected video that you wanted to see, and the time that it took until you saw the video, met your expectations. *(Editor's note: Depending on the rating scheme chosen, the question can also be asked after the video has finished; the text has to be modified in this case.)*

– Rating: Once the video has finished you should rate the overall quality on the following scale: *(Picture of ACR scale).* Your task is to judge the *overall quality* of each sequence – not the content. Parts of the playback where the video is not playing and a waiting indicator is shown are part of the test. Any variation of video and audio quality is also part of the test. You should consider both of these things as part of your overall quality judgment.

– This process will be repeated until you have seen and voted on (*insert number*) sequences, then we'll take a break. Then, there will be another similar session.

NOTE – Due to limited duration of the video, some videos may end abruptly in a middle of scene. Please do not consider this abrupt ending in your judgment. There is no wrong answer in this task; just rely on your own judgment.

Practice: At the start of the test the first sequences you see will be practice sequences so you can get a feel for the set-up and how to make your ratings. After several of these practice sequences you will then be told when the test starts.

Questions: If you have questions, please ask the experiment leader.

Subject consent: *(example, may be different for each lab).* The *(name of experiment)* experiment is being conducted at the (*name of your lab*) lab. The purpose, procedure, and risks of participating in the *(name of experiment)* experiment have been explained to me. I voluntarily agree to participate in this experiment. I understand that I may ask questions, and that I have the right to withdraw from the experiment at any time. I also understand that (*name of lab*) lab may exclude me from the experiment at any time. I understand that any data I contribute to this experiment will not be identified with me personally, but will only be reported as a statistical average.

Signature of participant		Signature of experimenter
Name of participant	Date	Name of experimenter

Appendix II

Test questionnaire

(This appendix does not form an integral part of this Recommendation.)

In order to understand a subject's background on video streaming usage, the following questionnaire may be asked after the main test procedure.

Q1. How fast is your Internet connection at home? (Single select)

Less than 2 Mbps	
2 ~10 Mbps	
10 ~ 50 Mbps	
50 ~ 100 Mbps	
More than 100 Mbps	
Not sure	
Do not have internet connection	

Q2. What generation of mobile network are you currently using on your smartphone? (Single select)

2G	
3G	
4G-LTE	
Not sure	
Do not use smartphone	

Q3a. What devices do you typically use for streaming video at home (Multi-select)

Television	
Tablet	
Smartphone	
Laptop	
Personal computer	
Other, please specify_____	

Q3b. What devices do you typically use for streaming video on the go (Multi-select)

Television	
Tablet	
Smartphone	
Laptop	
Personal computer	
Other, please specify_____	

Q4a. In the past three months, how often did you watch online videos via your mobile phone/PC? (Single select)

Less than once per day	
Once per day	
2-3 times per day	
4-5 times per day	
More than 5 times per day	

Q4b. In the past three months, on average how much time did you spend in a session when you watched online videos on your mobile phone/PC? (Single select)

Less than 5 mins	
5-9 mins	
10-19 mins	
20-29 mins	
30-59 mins	
1-2 hr	
More than 2 hr	

Q5. What is your favorite type of video? (Multi-select)

Movies, movie trailers	
TV shows	
Sports	
Music video	
Animation	
News	
Documentaries	
Other, please specify_____	

Appendix III

Pilot test on the impact of initial loading delays

(This appendix does not form an integral part of this Recommendation.)

This appendix describes a pilot test performed by ITU-T on the impact of initial loading delays on user experience, and which provided input for the development of the main body of this Recommendation.

Target

The pilot test aimed at capturing a user's experience of initial loading delays during audiovisual streaming. The purpose of the pilot test includes three aspects:

1) verifying the validity of the mobile test system;
2) optimizing the subjective testing process;
3) analysing preliminary results; suggestions for the evaluation of the initial loading scores and future tests are made.

Pilot test plan

In this pilot test, to study a user's experience of initial loading of audiovisual streaming, the subjective experiment was designed with the following factors: initial loading delay, video duration, and video quality.

Experimental materials

There are 84 experimental sequences. SRCs include: landscapes, food, outdoor sports, advertising, entertainment programs, and news interviews. The SRCs have six duration levels, which are 10 s, 20 s, 40 s, 1 min, 2 min and 3 min. Video resolution includes five levels of 320p, 480p, 720p, 1080p, and 1440p. The specific settings are shown in Table III.1.

Table III.1 – Pilot test SRC settings

Resolution	Video duration/s	Num	Resolution	Video duration/s	Num
320p	10	3	720p	60	3
320p	20	3	720p	120	3
320p	40	3	720p	180	2
320p	60	3	1080p	10	3
320p	120	3	1080p	20	2
320p	180	3	1080p	40	3
480p	10	3	1080p	60	3
480p	20	3	1080p	120	2
480p	40	2	1080p	180	3
480p	60	3	1440p	10	2
480p	120	3	1440p	20	3
480p	180	3	1440p	40	3
720p	10	3	1440p	60	2
720p	20	3	1440p	120	3
720p	40	3	1440p	180	3

Pilot test platform and test environment

To assess the quality of the initial loading experience, a subjective test system based on the Android mobile phone platform was developed. The pilot test system can control the initial loading delay and provide subjects with video playbacks with different HRCs.

To ensure a close to realworld experience for the subjects, the pilot test system was made consistent with current mainstream video services. The mobile interface of the test system mainly simulates the interface of a major OTT video platform, as shown in Figure III.1.

Subjects select a video for full-screen playback and can then evaluate the quality of the video at the end. Experimental environment followed the settings in [ITU-T P.911].

Figure III.1 (a) – Main interface

Figure III.1 (b) – Play interface

Score sheet

No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Score														
No.	15	16	17	18	19	20	21	22	23	24	25	26	27	28
Score														
No.	29	30	31	32	33	34	35	36	37	38	39	40	41	42
Score														

Break

No.	43	44	45	46	47	48	49	50	51	52	53	54	55	56
Score														
No.	57	58	59	60	61	62	63	64	65	66	67	68	69	70
Score														
No.	71	72	73	74	75	76	77	78	79	80	81	82	83	84
Score														

Figure III.2 (a) – Score sheet for rating element 1

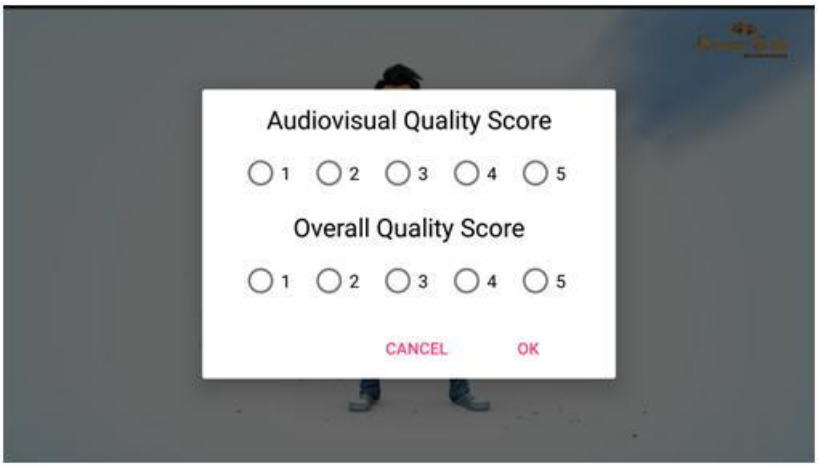

Figure III.2 (b) – Scoring interface of rating element 2 – Option 1

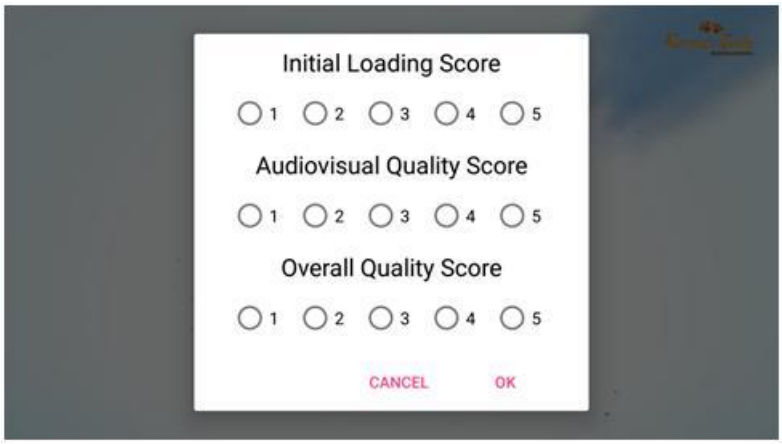

Figure III.2 (c) – Scoring interface of rating element 2 – Option 2

HRC settings:

The initial loading delay was set to 14 levels of 0.1 s, 0.2 s, 0.3 s, 0.5 s, 0.7 s, 1 s, 2 s, 4 s, 6 s, 8 s, 10 s, 15 s, 20 s, and 30 s. There are six video sequences for each initial loading delay. The video resolution and video duration are randomly assigned to a total of 84 HRCs.

Table III.2 – Experimental HRCs

HRC	Initial loading delay (s)	Video duration	Resolution	HRC	Initial loading delay (s)	Video duration	Resolution
HRC1	0.1	10 s	320P	HRC43	0.1	1 min	320P
HRC2	0.2	10 s	480P	HRC44	0.2	1 min	480P
HRC3	0.3	10 s	720P	HRC45	0.3	1 min	720P
HRC4	0.5	10 s	1080P	HRC46	0.5	1 min	1080P
HRC5	0.7	10 s	1440P	HRC47	0.7	1 min	1440P
HRC6	1	10 s	320P	HRC48	1	1 min	320P
HRC7	2	10 s	480P	HRC49	2	1 min	480P
HRC8	4	10 s	720P	HRC50	4	1 min	720P
HRC9	6	10 s	1080P	HRC51	6	1 min	1080P
HRC10	8	10 s	1440P	HRC52	8	1 min	1440P
HRC11	10	10 s	320P	HRC53	10	1 min	320P
HRC12	15	10 s	480P	HRC54	15	1 min	480P
HRC13	20	10 s	720P	HRC55	20	1 min	720P
HRC14	30	10 s	1080P	HRC56	30	1 min	1080P
HRC15	0.1	20 s	1440P	HRC57	0.1	2 min	1440P
HRC16	0.2	20 s	320P	HRC58	0.2	2 min	320P
HRC17	0.3	20 s	480P	HRC59	0.3	2 min	480P
HRC18	0.5	20 s	720P	HRC60	0.5	2 min	720P
HRC19	0.7	20 s	1080P	HRC61	0.7	2 min	1080P
HRC20	1	20 s	1440P	HRC62	1	2 min	1440P
HRC21	2	20 s	320P	HRC63	2	2 min	320P
HRC22	4	20 s	480P	HRC64	4	2 min	480P
HRC23	6	20 s	720P	HRC65	6	2 min	720P
HRC24	8	20 s	1080P	HRC66	8	2 min	1080P
HRC25	10	20 s	1440P	HRC67	10	2 min	1440P
HRC26	15	20 s	320P	HRC68	15	2 min	320P
HRC27	20	20 s	480P	HRC69	20	2 min	480P
HRC28	30	20 s	720P	HRC70	30	2 min	720P
HRC29	0.1	40 s	720P	HRC71	0.1	3 min	1080P
HRC30	0.2	40 s	1080P	HRC72	0.2	3 min	1440P
HRC31	0.3	40 s	1440P	HRC73	0.3	3 min	320P
HRC32	0.5	40 s	320P	HRC74	0.5	3 min	480P
HRC33	0.7	40 s	480P	HRC75	0.7	3 min	720P

Table III.2 – Experimental HRCs

HRC	Initial loading delay (s)	Video duration	Resolution	HRC	Initial loading delay (s)	Video duration	Resolution
HRC34	1	40 s	720P	HRC76	1	3 min	1080P
HRC35	2	40 s	1080P	HRC77	2	3 min	1440P
HRC36	4	40 s	1440P	HRC78	4	3 min	320P
HRC37	6	40 s	320P	HRC79	6	3 min	480P
HRC38	8	40 s	480P	HRC80	8	3 min	720P
HRC39	10	40 s	720P	HRC81	10	3 min	1080P
HRC40	15	40 s	1080P	HRC82	15	3 min	1440P
HRC41	20	40 s	1440P	HRC83	20	3 min	320P
HRC42	30	40 s	320P	HRC84	30	3 min	480P

Pilot test procedure

For the initial loading experience, two kinds of subjective experimental test programs were designed, in order to obtain scores for the subject's initial loading experience at different stages of viewing the video (see rating elements 1 and 2 above).

Through the analysis of experimental results, it is possible to determine which subjective test method (i.e., choice of questions in each rating element) reflect the subject's experience and provides a reasonable and effective experimental basis for subsequent proponents to conduct tests in a similar fashion.

By using the results of such subjective experiments, the relationship between the subject's quit rate and the initial loading delay, the influencing factors of the initial loading experience, and the relationship between the initial loading score and the overall quality of the viewing experience were analysed.

According to the initial research goal, the experiment included two sessions: the first session was mainly used to compare two subjective test schemes, and identify what their impact on the measured initial loading experience is. The second session was chosen to select one of the schemes from the first session, to conduct more detailed experiments to study the impact the initial loading score on the subject's overall experience.

Session 1

This session was designed with two subjective experiment schemes:

1) The first approach was to rate initial loading score immediately after the mobile started to play the video, and to rate the audiovisual quality and overall experience at the end of each video. The scoring method used the 5-point ACR scoring standard. The subject waited for the initial loading before viewing. If they felt that the initial waiting time was unacceptable, they could exit the video.

2) The second approach was to evaluate the initial loading experience, video quality, the overall experience at the end of each video. The initial loading experience, video quality, and overall experience all used the 5-point ACR scoring standard. Similar to the first approach, if the subject could not accept the initial waiting time before watching, they could quit. Acceptability of loading time was also tested by writing the 2-valued scoring on a paper at the end of each video. It was later shown to have a very strong correlation with the quitting rate.

For the experiment, each HRC from Table III.2 was selected with a video duration of 10 s, 1 min, and 3 min for pilot testing. There are 14 initial loading time values and 5 quality levels, resulting in a total of 42 HRCs. Details of the setting are shown in Table III.3. For each of the two schemes, 42 different video SRCs were used.

Table III.3 – HRC settings in session 1

HRC	Initial loading delay(s)	Resolution	Video duration	HRC	Initial loading delay(s)	Resolution	Video duration
HRC1	0.1	320P	10 s	HRC50	4	720P	1 min
HRC2	0.2	480P	10 s	HRC51	6	1080P	1 min
HRC3	0.3	720P	10 s	HRC52	8	1440P	1 min
HRC4	0.5	1080P	10 s	HRC53	10	320P	1 min
HRC5	0.7	1440P	10 s	HRC54	15	480P	1 min
HRC6	1	320P	10 s	HRC55	20	720P	1 min
HRC7	2	480P	10 s	HRC56	30	1080P	1 min
HRC8	4	720P	10 s	HRC71	0.1	1080P	3 min
HRC9	6	1080P	10 s	HRC72	0.2	1440P	3 min
HRC10	8	1440P	10 s	HRC73	0.3	320P	3 min
HRC11	10	320P	10 s	HRC74	0.5	480P	3 min
HRC12	15	480P	10 s	HRC75	0.7	720P	3 min
HRC13	20	720P	10 s	HRC76	1	1080P	3 min
HRC14	30	1080P	10 s	HRC77	2	1440P	3 min
HRC43	0.1	320P	1 min	HRC78	4	320P	3 min
HRC44	0.2	480P	1 min	HRC79	6	480P	3 min
HRC45	0.3	720P	1 min	HRC80	8	720P	3 min
HRC46	0.5	1080P	1 min	HRC81	10	1080P	3 min
HRC47	0.7	1440P	1 min	HRC82	15	1440P	3 min
HRC48	1	320P	1 min	HRC83	20	320P	3 min
HRC49	2	480P	1 min	HRC84	30	480P	3 min

Session 2

Through the analysis of the results of the first session, it was decided to focus further studies on the first scheme, as it would lead to more critical ratings of the initial loading experience, which were of main interest to the researchers.

Session 2 used the methods in the first session to conduct more detailed experiments to study the impact of initial loading experience. HRCs in Table III.2 with video durations of 20 s, 40 s, and 2 min were selected for testing. These videos included 14 initial loading delay delays and 5 quality levels, for a total of 42 HRCs. The distribution is shown in Table III.4.

Table III.4 – HRC settings in session 2

HRC	Initial loading delay (s)	Video duration	Resolution	HRC	Initial loading delay (s)	Video duration	Resolution
HRC15	0.1	20 s	1440P	HRC36	4	40 s	1440P
HRC16	0.2	20 s	320P	HRC37	6	40 s	320P
HRC17	0.3	20 s	480P	HRC38	8	40 s	480P
HRC18	0.5	20 s	720P	HRC39	10	40 s	720P
HRC19	0.7	20 s	1080P	HRC40	15	40 s	1080P
HRC20	1	20 s	1440P	HRC41	20	40 s	1440P
HRC21	2	20 s	320P	HRC42	30	40 s	320P
HRC22	4	20 s	480P	HRC57	0.1	2 min	1440P
HRC23	6	20 s	720P	HRC58	0.2	2 min	320P
HRC24	8	20 s	1080P	HRC59	0.3	2 min	480P
HRC25	10	20 s	1440P	HRC60	0.5	2 min	720P
HRC26	15	20 s	320P	HRC61	0.7	2 min	1080P
HRC27	20	20 s	480P	HRC62	1	2 min	1440P
HRC28	30	20 s	720P	HRC63	2	2 min	320P
HRC29	0.1	40 s	720P	HRC64	4	2 min	480P
HRC30	0.2	40 s	1080P	HRC65	6	2 min	720P
HRC31	0.3	40 s	1440P	HRC66	8	2 min	1080P
HRC32	0.5	40 s	320P	HRC67	10	2 min	1440P
HRC33	0.7	40 s	480P	HRC68	15	2 min	320P
HRC34	1	40 s	720P	HRC69	20	2 min	480P
HRC35	2	40 s	1080P	HRC70	30	2 min	720P

About the subjects

Twenty subjects were selected; each subject had normal vision.

Before the pilot test, the test procedure was explained to the subject. For the two pilot test schemes, three video samples with different initial loading delays and different video quality were selected to allow the subjects to be familiar with the operation of the test platform, the test process, and the scoring criteria.

After viewing the sample videos, the main test procedure was started. The distance between the subjects and screen was set to be 3H. To avoid fatigue, each session did not exceed 30 min. After each session, there was a 20 min rest. It was ensured that subjects would not be disturbed during the test.

Bibliography

[b-Dobrian] 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.

[b-Krishnan] 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.

[b-Mok] 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).

[b-Reiter] 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.

[b-Robitza] 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.

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	Tariff and accounting principles and international telecommunication/ICT economic and policy issues
Series E	Overall network operation, telephone service, service operation and human factors
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Environment and ICTs, climate change, e-waste, energy efficiency; construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	**Telephone transmission quality, telephone installations, local line networks**
Series Q	Switching and signalling, and associated measurements and tests
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks, open system communications and security
Series Y	Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of Things and smart cities
Series Z	Languages and general software aspects for telecommunication systems

Printed in Switzerland
Geneva, 2019