Quality and User Experience (2021) 6:1
https://doi.org/10.1007/s41233-020-00041-2

RESEARCH ARTICLE

In vitro vs in vivo: does the study's interface design influence crowdsourced video QoE?

Kathrin Borchert[1] · Anika Seufert[1] · Edwin Gamboa[2] · Matthias Hirth[2] · Tobias Hoßfeld[1]

Received: 1 December 2019 / Published online: 2 November 2020

Abstract

Evaluating the Quality of Experience (QoE) of video streaming and its influence factors has become paramount for streaming providers, as they want to maintain high satisfaction for their customers. In this context, crowdsourced user studies became a valuable tool to evaluate different factors which can affect the perceived user experience on a large scale. In general, most of these crowdsourcing studies either use, what we refer to, as an *in vivo* or an *in vitro* interface design. *In vivo* design means that the study participant has to rate the QoE of a video that is embedded in an application similar to a real streaming service, e.g., YouTube or Netflix. *In vitro* design refers to a setting, in which the video stream is separated from a specific service and thus, the video plays on a plain background. Although these interface designs vary widely, the results are often compared and generalized. In this work, we use a crowdsourcing study to investigate the influence of three interface design alternatives, an *in vitro* and two *in vivo* designs with different levels of interactiveness, on the perceived video QoE. Contrary to our expectations, the results indicate that there is no significant influence of the study's interface design in general on the video experience. Furthermore, we found that the *in vivo* design does not reduce the test takers' attentiveness. However, we observed that participants who interacted with the test interface reported a higher video QoE than other groups.

Keywords Video QoE · Crowdsourcing · Study design · User study · Distraction

Introduction

Video streaming is currently the most dominant Internet application, accounting for 75% of the global data traffic and is estimated to increase even more up to 82% in the next years [3]. Since video streaming grows in popularity, it is essential for Internet Service Providers (ISPs) to ensure high user satisfaction. Consequently, ISPs show increasing interest in metrics to quantify the experience and satisfaction of their customers. These efforts are in line with the general concept of QoE, which focuses on the subjective experience of a user using a service or application, for example, Voice over IP calls or YouTube video streaming. The QoE indicates the degree of delight or annoyance of a user of an application as perceived subjectively [22], and depends on various factors in the network, the application, but also on the user's context and expectations.

Subjective user studies are one of the standard tools in multimedia research to evaluate possible QoE influence factors and are typically conducted in laboratories. However, as they are expensive and time-consuming, crowdsourcing became a widely used alternative to collect subjective user ratings. In crowdsourcing, simple tasks or work processes are assigned to the broad mass of Internet users, which enables a fast collection of ratings from a diverse set of users at a fast pace and low costs [7, 21]. Therefore, this method represents a promising opportunity to enrich the toolset for subjective QoE experiments.

✉ Anika Seufert
anika.seufert@informatik.uni-wuerzburg.de

Kathrin Borchert
kathrin.borchert@informatik.uni-wuerzburg.de

Edwin Gamboa
edwin.gamboa@tu-ilmenau.de

Matthias Hirth
matthias.hirth@tu-ilmenau.de

Tobias Hoßfeld
hossfeld@informatik.uni-wuerzburg.de

[1] Institute of Computer Science, University of Würzburg, Würzburg, Germany

[2] Institute of Media Technology, TU Ilmenau, Ilmenau, Germany

Many crowdsourcing and laboratory studies focus on specific stimuli and try to keep the number of potential influence factors as low as possible. Consequently, the video streaming experience is often decoupled from real streaming services, which means the video under test is not embedded into a realistic streaming service environment like YouTube or Netflix. This becomes even more evident for the highly standardized test environments for visual media quality assessments in general [16]. We refer to this widely used study interface design as *in vitro*, which means literally, "in a glass," or "in a test tube". In contrast to this, it is also possible to design studies in a more natural setting. We denote these studies as *in vivo*, which means literally, "in the living". Here, the study is conducted in conditions that precisely mirror those existing in real life, e.g., video streaming embedded in a real service like YouTube. With both interface designs - *in vivo*, *in vitro*, and also multiple nuances in between - in place, the question arises if and to which extent the interface design, influences the test participants' QoE.

In vitro interfaces are usually stripped down to a minimum to direct the viewer's attention to the stimuli under test and thus, raters might be more sensitive towards impairments compared to a real-live setting with possible distractions. However, if an *in vivo* interface reassembles a live setting or a real streaming service, users might be biased by their current and previous experience with this service. Consequently, the subjective ratings gathered in this setting are no longer only affected by the stimulus but also the user's expectations. Further, a feature-rich *in vivo* interface can distract the participants' attention such that the testers overlook service impairments and perceive a better streaming experience. Additionally, it is also unclear what is needed to create the illusion of a *real* service, i.e., whether it is sufficient to provide a mockup with limited functionalities, or whether a re-implementation of the full functionalities with additional means for manipulating the stimuli is necessary.

Answering all these questions lies out of scope for a single research paper. Therefore, in this work, we focus on the influence of the interface design on the perceived video QoE and the acceptance of the video streaming quality. To address these questions, we conduct multiple large-scale crowdsourcing studies and compare the QoE ratings of participants who watch video stimuli on a gray background (*in vitro*) to the results from another test group that views the same stimuli on a YouTube-like web site with different levels of interactiveness (*in vivo* and *in vivo light*). The *in vitro* setting is motivated by the current video quality evaluation standards [16], the *in vivo* and *in vivo light* settings are first steps towards video quality assessments in real-service settings, but with limited functionalities. We limit the type of impairment to stalling events, as they showed a high effect on streaming video QoE in previous studies [26]. Similar to standardized evaluation recommendations in the field of video quality assessments [16], we play the video content without an audio stream.

A reduced version of this work is published in [1]. In comparison to the previously published version, this work additionally addresses different levels of interaction possibilities in the *in vivo* setting, the effect of the interface design on the perceived annoyance of stalling events, the influence of the number of interactions on perceived streaming experience, and an in-depth analysis of the focus of the test participants in relation to their interactions with the test environments. Therefore, an additional large-scale crowdsourcing study with more than 800 participants enlarges the dataset used for the evaluation.

The remainder of this work is structured as follows. Section 2 provides background information on crowdsourcing and discusses related work on influence factors for video QoE and crowdsourced study design. The methodology and design of our study are addressed in Sect. 3. The collected data and filters applied to identify and remove unreliable workers are explained in Sect. 4. Section 5 describes the evaluation of the influence of the study's interface design on the QoE results. Finally, Sect. 6 summarizes this work and gives an outlook on future work in this field.

Background and related work

Many factors influence a user's perception of a streamed video. On the network and application side, factors like initial delay, number and frequency of stalling events, the used quality layer, as well as the time on the quality layers, influence the user's QoE [25–27]. Besides these technical factors, there is also a large set of context factors affecting the QoE of video streaming, which includes, for example, the social context, gender, age, or the participant's interest in the video stimulus' content [34].

The large number and diversity of QoE influence factors open a vast parameter space for subjective studies. This directly calls for a fast and cost-effective alternative to laboratory studies, which was found in the concept of crowdsourced user studies. Over the past years, the research community developed several general guidelines on quality attributes, assessment techniques, and assurance action to correctly conduct crowdsourcing studies [4, 8, 13, 14, 29] and also several best practices explicitly for subjective assessments via crowdsourcing [6, 15]. Most crowdsourced studies in the QoE research focus only on evaluating a single or only a few specific stimuli. Therefore, most of the studies are being carried out *in vitro*, which means that the stimuli are displayed without context and without embedding them, e.g., into a real streaming service environment like YouTube or Netflix. For research on video streaming, it is well established to provide simple web interfaces including the video

to be rated on a plain background, like in [19, 28, 33]. Here, unlike a real streaming portal, no video recommendations or posted comments are visible. On the other side, there are also crowdsourced studies which analyze the perceived user experience *in vivo*, directly out of a real video streaming service, like using the tools YouSlow [24] and YoMo [31]. YouSlow is a Chrome extension which monitors YouTube re-buffering events while users watch YouTube videos on their client. For YoMo, participants watch a video directly on YouTube and rate their experience afterward. Also, *in vivo* designs can be employed as part of "Living Labs" environments, for instance, in [20] a video streaming platform was developed to analyze QoE in the context of mobile video streaming. Both study designs, *in vivo* and *in vitro*, are frequently in use, but to the best of our knowledge, there still exists no clear analysis on the comparability of the obtained results, even if works on related research topics exist.

A method of contextualized subjective quality experiments is discussed in [2]. Here, the participants were allowed to choose their preferred device, e.g., TV or computer, and their preferred social context, e.g., alone, with a partner, or with the entire family. The results showed clearly that social and emotional aspects have a substantial influence on the QoE evaluation. Another important factor regarding the perceived QoE is the environment the study is conducted in. In [30], the authors investigated the influence of the environment on the impairment visibility and acceptability. They compared a lab study to the natural setting in the living room of the participants. They showed that quality ratings obtained with one of the standardized subjective quality assessment methodologies do not always match the case of real-life QoE assessment. In [32], the authors demonstrated that context factors such as display size, viewing distance, ambient luminance, and user movements strongly correlate with QoE of mobile video. They found that the results obtained using standardized experiments significantly differ from real-life QoE assessment. For example, impairments are less visible during real-life QoE assessment than in standardized lab studies. Furthermore, a comparison of the video quality assessments on mobile devices in the field and the lab was made in [17]. The authors found that the negative effect of packet loss on the perceived QoE was higher in the lab than in the field. A comparison of web browsing in real-world and employed laboratory tests on single or multiple page views was made in [10]. The authors showed that QoE ratings for web browsing are not affected by the considered contexts or distractions.

As previous studies only focus on the external environment of the participants, it is not known if the perceived QoE from *in vitro* crowdsourcing studies is comparable with *in vivo* studies, conducted in a real streaming service environment. Thus, in this work, we investigate the influence of the study's interface design on crowdsourced video QoE and acceptance. Besides, we evaluated and compared the attentiveness of the participants in the *in vivo* and the *in vitro* interface designs.

Study description

We formulate following hypotheses to investigate the influence of the interface design (*in vivo* or *in vitro*) on crowdsourced video QoE assessments:

H1: The perceived video QoE is influenced by the study's interface design.
H2: The acceptance of the streaming quality is influenced by the interface design.
H3: The degree of annoyance of stalling events is influenced by the interface design.
H4: The participants' focus on the stimulus is influenced by the interface design.
H5: The streaming experience is influenced by the participant's interactions with the web page.

To test these hypotheses, we conducted a crowdsourced user study, which is divided into four different sub-studies. While designing the study, we considered the best practices for quality assurance in crowdsourcing [4] as well as best practices especially for QoE crowdtesting [6, 15]. The overall test design and the differences between the individual sub-studies are shown schematically in Fig. 1.

Introduction First, the participants have to read a short task description and an introduction about video stalling. Their task is explained as follows: "watch one short video and rate the quality. Please note that the video has no audio. [...] In this task you are a video quality tester. Please stay focused on the test during video playback and rate the quality immediately after the video!". To introduce the term stalling, the participants are shown a screenshot of a video with a stalling symbol with the following explanation: "During the playback of an online video, the video sometimes has to pause in order to load more data. These video stops are called video stalling events. YouTube, for example, displays in case of stalling the loading circle illustrated in the picture below.". Afterwards, the participants have to complete a short pre-task, as described in [9], for testing the contrast of the screen and the honesty of the participant. Here, equal shapes of a cat with different contrast are displayed and users have to select all visible shapes. In addition, one cat has the same color as the background and thus, cannot be seen but just found by random clicking. The results of this task are later used as a reliability check.

Demographic questionnaire In the next step, the participants have to provide demographic information including their age, gender, the continent on which they live, and

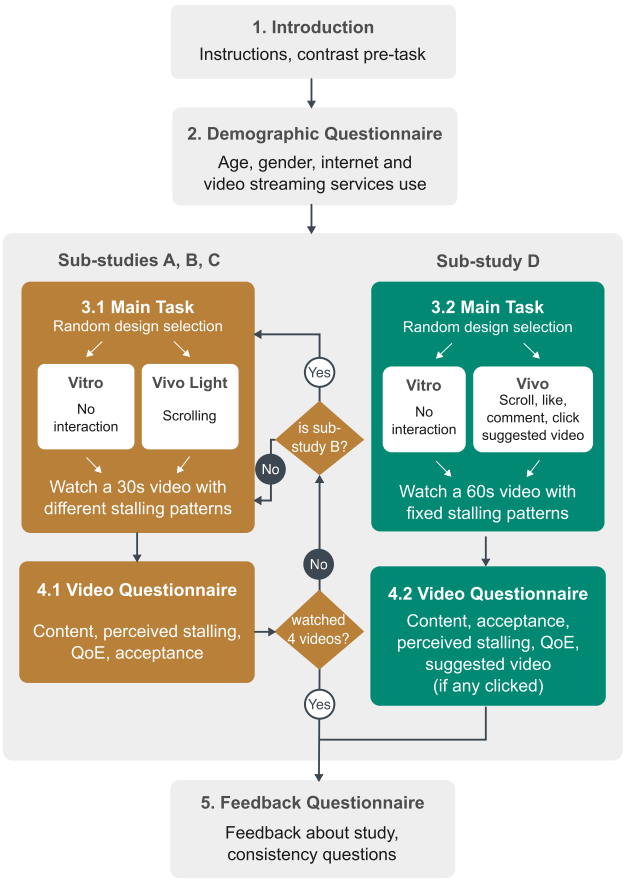


Fig. 1 Structure of the test design for all studies

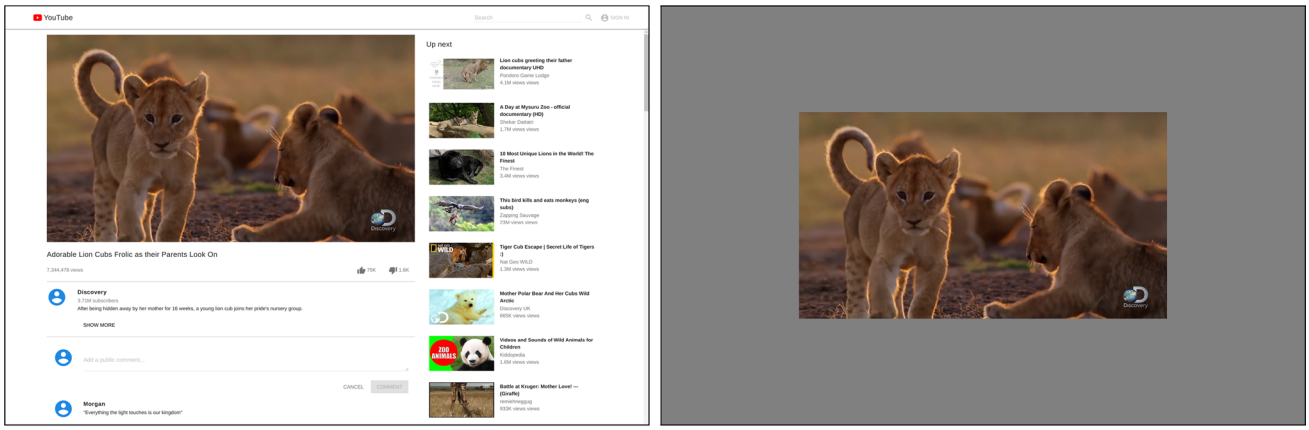
information about the frequency of surfing the Internet and using video streaming services.

Main task and video questionnaire Afterward, depending on the actual sub-study, which we will explain later in Sect. 3.2, one or four randomly selected videos are shown either on a plain gray background (*in vitro*) or in a YouTube-like web page (*in vivo*). The video can be one of four different videos, which cover a wide range of characteristics: A soccer match (fast motion), a talk (almost no motion), an animal documentary (slow motion), and a pop concert (motion, amateur recording). Since services like YouTube are among other things also platforms for amateur videos, the concert video is included to cover this aspect. The videos are provided without audio to reduce additional factors influencing the perceived quality. Furthermore, the videos are shown in high-quality (1080p) using the AVC1 codec without adaptations, playing at least 25 frames per second and have a length of 30s (sub-study *A* to *C*) or 60s (sub-study *D*). Each video is followed by questions about the content, whether stalling events were noticed, how often and when stalling events were noticed, the perceived annoyance of the stalling events, the perceived QoE of the streaming, and the acceptance of a fictional streaming service that exhibits the streaming quality observed in the test. Additionally, the participants are instructed to select the option that stalling was not annoying at all, if they did not notice any stalling. During the entire user study, it is monitored whether and for how long the study browser tab is in focus as well as the current mouse position. The annoyance rating, the content question, the questions about stalling presence, and the tracked focus time of the browser tab while playing the video are later used for reliability checking and are referred to as video dependent checks.

Feedback questionnaire After watching the video, a final questionnaire is presented to the test takers. Besides the possibility to provide feedback about the study, the users have to report the frequency of using the Internet and streaming services again. Furthermore, they have to select their country of residence. These answers, in combination with the screen contrast test in the pre-task, are used as a consistency check, further referred to as video independent checks. The users are not allowed to participate more than once.

In vivo and in vitro design

For investigating the impact of the *in vivo* or *in vitro* interface design, the selected video is shown either on a YouTube-like web site or a plain gray background similar to [28]. The video to be displayed as well as the design setting is randomly selected before a participant accesses a test. Besides the background, there are additional differences between the *in vivo* and *in vitro* setting, such as the position of the video player within the web page and the total height of the web page. These differences are inevitable because of the study design constraints. In particular, the video player for the *in vitro* design is placed in the center of the web page, whereas the video player for the *in vivo* design is placed in the same position as in the original YouTube web site. Due to the overall height of the web page, participants watching the video sequences on the gray background are not able to scroll. As the YouTube-like design includes the description field of the author who uploaded the video, comments of other users, and previews of suggested videos, the test takers can scroll and can become distracted by other parts of the web page. The mouse position relative to the web page is tracked every 500 ms for sub-study *A* to *C* and every 100ms for sub-study *D* while watching the video, to monitor the interaction behavior of participants. Also, the video position is tracked every 100 ms in the *in vivo* design. This monitoring technique, in conjunction with the screen size, allows us to check if the page is scrolled and if the video is still in the visible range. The *in vivo* design in sub-studies *A* to *C* is further referred to as *in vivo light* since it only offers scrolling as interaction possibility. Meanwhile, in sub-study *D*, test takers can interact with the web page by liking or disliking the video, displaying or hiding the whole video



(a) *In vivo* design – YouTube-like web page.

(b) *In vitro* design – Web page with video player on gray background.

Fig. 2 Screenshots of the *in vivo* and *in vitro* interface designs

Table 1 Settings of the study design and stalling pattern

Sub-study	# Different videos	# Shown videos	Video length	Design per User	# Stalling events	Stalling length	Stalling position	Research question
A	4	4	30 s	Fixed	0 or 1	6 s	At 15 s	**H1–H5**
B	4	4	3 0s	Variable	1	6 s	At 15 s	Influence of study design
C	4	4	30 s	Fixed	0 or 1	6 s	At 5 s	**H4**
D	3	1	60 s	Fixed	2	6 s	At 5 s & 50 s	**H1–H5**

description, adding a new comment, cancelling the addition of a new comment, clicking one of the suggested videos, and changing the window size as the design is responsive. After clicking on a suggested video, the test taker is requested to answer the study questionnaire including a question to indicate the reason for this behavior. During video playback, we track all the interactions for later analysis since these may lead test takers to lose focus, i.e., the test takers may not notice the stalling patterns due to interaction possibilities.

An example of the realization of the *in vivo* and *in vitro* setting is shown in Fig. 2. Figure 2a shows the *in vivo* interface for the documentary, while Fig. 2b presents the *in vitro* design, displaying the documentary on a gray background. The *in vivo light* and *in vitro* versions are implemented using the open source mockup tool ERWIN [35] which easily allows building lookalike versions of existing applications and services. We enhanced the tool for supporting video content and stalling patterns. Meanwhile, the *in vivo* design for the sub-study *D* is implemented using JavaScript and Material Design Lite[1].

[1] https://getmdl.io (Last accessed Jun. 2020).

Sub-study design

We conducted four sub-studies using different stalling pattern and interface design settings to address the five hypotheses **H1** to **H5**. The details of the studies are summarized in Table 1. For sub-study *A* to *C*, the test takers had to watch four videos with a length of 30s each, for sub-study *D* they only had to watch one video with a length of 60s. We use videos with stalling to evaluate the impact of the interface design on the perceived quality. To reduce the parameter space, we only use one stalling setting per sub-study and remove the audio track.

In sub-study *A*, the interface design is fixed for each participant, i.e., a participant watches four videos either in the *in vivo light* or *in vitro* setting. The setting for the participant is randomly chosen before the user accesses the test. Moreover, the video stimulus either plays out without any stalling event or exhibits exactly one stalling event after 15s of a fixed length of 6s. For each video the participant watches, it is randomly chosen whether a stalling event occurs or not.

In sub-study *B*, the interface design is selected randomly for each video and for all participants. Again, the participants watch 4 videos, but all videos stall for 6s after a

playback time of 15s. With this setup, we evaluate whether the study design used in sub-study *A*, i.e., showing all videos in the same interface, influences our results.

In the *in vivo light* design, we assume that some users start watching the video sequence and after a few seconds they start scrolling down to explore the parts of the web site which are not in the visible range of the screen. For investigating the inattentiveness of the test takers, we use another point in the playtime for the stalling. A stalling which appears at the beginning of the video might be missed while exploring the web site. Thus, for answering this research question, sub-study *C* uses a similar configuration as sub-study *A*, except, that if a stalling occurs, it occurs 5s after starting the playback.

In a relatively short video length of 30 s the workers might not get bored that fast. Thus, in sub-study *D*, we expand the previous study by prolonging the video duration to 60 s. In addition, to further investigate the influence of the stalling position, we add a second stalling after 50 s of playtime. Doing so, we are able to evaluate if a stalling which appears at the end of a video might be missed more frequently. Moreover, to investigate the influence of interaction with the web page, we added further interaction possibilities to the *in vivo* setting having, for example, clickable video recommendation links and like buttons as explained previously. We also increased the tracking of the mouse position (from every 500ms to every 100 ms) for a closer monitoring of the interaction behavior.

We use sub-study *A* and *D* to investigate the influence of the *in vivo light* / *in vivo* or *in vitro* interface design on the QoE (hypothesis **H1**) and the acceptance of the streaming quality (hypothesis **H2**). The participants' degree of annoyance of stalling events depending on the design (hypothesis **H3**) is also investigated by sub-study *A* and *D*. Hypothesis **H4**, the interface design influences the participants' focus on the stimulus, as well as hypothesis **H5**, the participant's interactions with the web page influences the streaming experience, are investigated by comparing the results obtained in sub-study *A* and *C* as well as evaluating the results of sub-study *D* with the *in vivo* design. Furthermore, to investigate influences of the study design on the participants' perception, we used sub-studies *A* and *B*.

Dataset description

To collect a large number of ratings, we conducted all sub-studies on crowdsourcing platforms. Sub-study *A* to *C* were posted only on the Microworkers[2] crowdsourcing platform, while sub-study *D* was additionally posted on Amazon Mechanical Turk (MTurk)[3]. Independent of the platform, the participants received a reward of 0.15 USD with an estimated time to complete the task of less than 7 min. No further restrictions, like country or skill filters were applied to limit the workers' access to the task. In the following, we will first explain our reliability checks, and, afterwards, we give an overview of the number of workers and their demographics.

Filters

To distinguish reliable users from users who did not perform the test properly, several checks were conducted. In the following, we explain our mechanisms to detect unreliable workers.

Video independent reliability checks To identify workers who did not read the instructions carefully or did not understand them, we included video independent reliability checks. The first video independent reliability check is placed at the introduction page of the study website. The page displays two low-contrast images showing pictograms of cats on a black and a white background. The participants were instructed to click on all cats that are visible for them. In particular, they were pointed out that they should not click randomly. Thus, we used the number of misclicks as our first reliability check (*R*1). Next, on the demographic questionnaire, we ask the workers on which continent they live, how often they surf the Internet, and how often they have watched video clips/streams on the Internet during the last month. All questions, except the continent of residence use a 7-point Likert scale. The answers to these questions are later compared to the answers from the feedback questionnaire at the end of the study. Here again, the participants have to state how often they use the Internet and streaming services, and in which country they live. Thus, the consistency of both answers given to the same question can be used as reliability check. In detail, the match of the answers about the Internet usage and about the usage of video streaming services are each used as second and third reliability check (*R*2, *R*3), where deviations by one point on the rating scale are accepted. The fit together of the answers about the place of residence is used as fourth video independent reliability check (*R*4).

Video dependent reliability checks In addition to the video independent checks, we also included video dependent reliability checks to verify that the workers watched the video carefully. Therefore, in the main task, we tracked the focus time of the browser tab while playing the video and used it as reliability check (*R*5). As the participants were told to stay focused on the video, we expect them to watch

[2] https://microworkers.com (Last accessed Jun. 2020).

[3] https://www.mturk.com/ (Last accessed Jun. 2020).

at least 70% of the video length (including the stalling time). Furthermore, we included four more reliability checks in the video questionnaire. For the first reliability check (*R*6), the test takers had to answer a question about the video content, for example, if they have seen a climbing scene, a car race, or a pop concert. Next, by comparing the answers of the question whether they noticed any stop during the playback and the number of stops they noticed, we were able to calculate the consistency of the answers of the user (*R*7). Additionally, we check the consistency of the stated number of noticed stalling events and position of the stalling, i.e., if the workers notice a stop in the first half, the middle, the second half, or if they did not notice a stop at all (*R*8). As the last video dependent reliability check (*R*9), we use the fit of the annoyance rating to the answer if they have noticed any stalling, as if they did not notice any stalling they were instructed to rate "not annoying at all".

Technical reliability checks We added additional checks to exclude ratings of users who had technical problems since the study environment on the participant's end-user device is not controllable. Participants who have seen less than two stalling events, e.g., due to tab changes, are filtered out. Furthermore, participants who indicated that they have seen more than two stalling events where excluded from the evaluation. As the interface in the *in vivo light* studies is not responsive, the screen width was added as technical reliability check. Here, only participants with a screen width greater than 1000px were taken into account. In addition, for sub-study *A* and *C*, we excluded all participants who saw a stalling event even though no stalling event was included.

For all sub-studies, *R*1, as well as all technical reliability checks, were used as mandatory checks, which filter out every user who failed one of these checks. For sub-studies *A* to *C*, the maximum tolerated value for *R*1 was set to 3 misclicks. As we increased the number of other reliability checks in sub-study *D*, we increased the accepted number of misclicks to 10. Focusing on the video independent and dependent checks, it was required to pass all checks from *R*3 to *R*6 as well as *R*9 for sub-study *A* to *C*. For sub-study *D*, we added additional filters and, thus, the participants have to pass 6 out of 8 checks of *R*2 to *R*9 (75%) to be taken into account in the evaluation.

Remaining workers and ratings

After collecting the data, the ratings were filtered based on the above mentioned reliability checks. The number of workers which completed the studies as wells as the number of workers who passed all reliability checks are listed in Table 2.

Sub-study *A* was conducted from January 03–14, 2019. Overall, 747 workers started to work on our study and 497 completed the final questionnaire. Of these, 278 participants passed the video independent, dependent, and the technical reliability checks, which remain providing 1021 ratings in total.

Table 2 Number of Workers per sub-study

Sub-study	Date	# Completed workers	# Workers passed all checks	# Remaining ratings
A	January 2019	497	278	1021
B	January 2019	90	58	216
C	January 2019	109	64	235
D	November 2019	822	451	451

We repeated sub-study *A* with changed settings in sub-study *B*, to evaluate influences caused by the study design, i.e., showing each video with the same interface design within the test. Sub-study *B* was available from January 17-21, 2019. Here, 99 workers started the study of whom 90 finished. Of those workers, 58 passed all checks previously mentioned, providing 216 ratings.

The third sub-study *C* was run from January 14–16, 2019. Overall, 109 workers out of 116 workers finished the study. In total, 64 workers providing 235 ratings passed all reliability checks.

Sub-study *D*, which was started by 1279 participants, was conducted from November 21–27, 2019. In this time, 822 workers completed the final questionnaire. Based on our reliability checks we excluded 45.13% of the workers and used the ratings of 451 workers for our evaluation.

Demographics

Having a closer look at the countries of residence, 20–30% of the participants originated from the top two countries of Microworkers India and Serbia throughout all studies. For sub-study *D*, which was also conducted on MTurk, one of the three top countries besides India and Serbia was the United States. Concerning gender, the majority of the participants are male (62.5–72.7%). While the distribution of the countries of residence for Microworkers differs from previous studies [12], the United States and India are reported as the most active workers countries on MTurk [5]. The share of male participants is similar to previously reported values from other studies for MTurk and Microworkers [5, 11, 23].

The analysis of the self-reported age of the participants shows a median age of 30 years for sub-study *A* and *C*, while the participants of sub-study *B* are little younger with a median of 27 years, and the participants of sub-study *D* are little older with a median of 31 years. Again, the age distribution is similar to the results for other studies [5, 11, 23]. As the age may have an impact on the accuracy of the workers' outcome [18], we statistically analyze the age

distribution for the subgroups, i.e., the *in vitro* and *in vivo light / in vivo* design, with the Kruskal-Wallis test. The null hypothesis can not be rejected with $p > 0.05$, which indicates that the samples originate from the same population.

Results

In the following, we used the collected data of all sub-studies to test our five hypotheses **H1** to **H5**. We analyze the impact of the *in vitro* and *in vivo light / in vivo* design on the perceived QoE and the acceptability, the degree of annoyance, as well as the influence of the interface design and interactiveness on the attentiveness and degree of annoyance.

Influence on QoE and acceptance

First, we compare the ratings collected under the same conditions in sub-study *A* and *B* to evaluate whether the study design, i.e., no variation of the interface design within the test per participant in sub-study *A* and *C*, influences the perception of the participants.

First, we evaluate whether the study design, i.e., no variation (like in sub-study *A* and *C*) in contrast to variation of the interface design (like in sub-study *B*) within the test per participant influences the perception of the participants. Therefore, we compare ratings collected under the same conditions from sub-study *A* and *B*. In sub-study *B*, for each video the interface design was selected randomly. Again, only ratings which pass all quality checks are considered. Due to a low number of workers passing the filters for the *in vitro* version of the video sequence showing the talk in sub-study *B*, we exclude this video from our analysis. Using Mann-Whitney U tests to compare ratings from sub-study *A* and *B* for each of the remaining videos, we do not see a significant effect between the study design and the ratings with all $p > 0.05$. Thus, we can conclude that using the same interface design (like in sub-study *A* and *C*) or changing the interface design during a test (sub-study *B*) does not affect the user's ratings.

After we have excluded influences of the test design, we analyze the impact of the *in vitro* and *in vivo light / in vivo* design on the perceived streaming quality and acceptance. As described in Sect. 4, we use the data collected in sub-study *A* and *B*, having a video length of 30s with no or one stalling event, as well as the data from sub-study *D*, having a video length of 60s with two stalling events.

Figure 3 shows the Mean Opinion Score (MOS) values with 95% confidence intervals for the streaming quality perceived by the participants of sub-study *A*. The stalling lengths are presented in different colors, with a length of 0s indicating the absence of stalling. The effect of the video content on the ratings collected under the same test

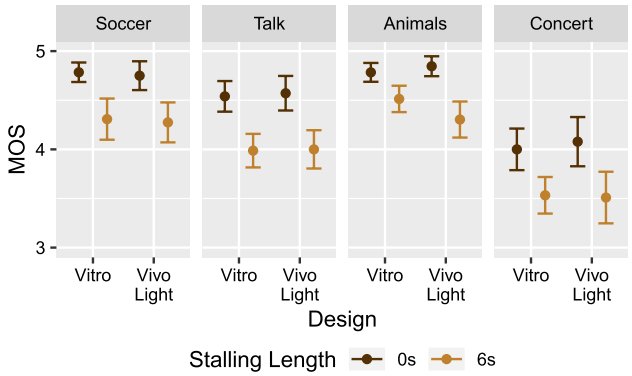
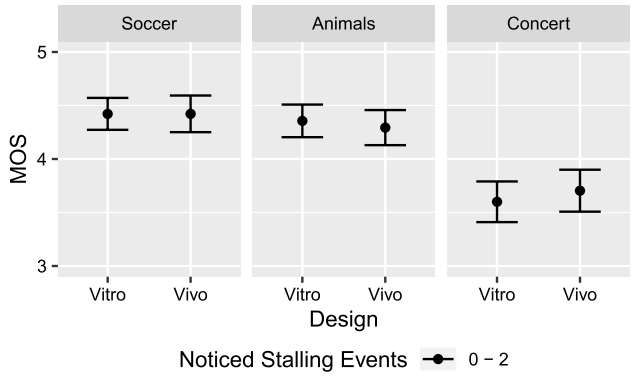


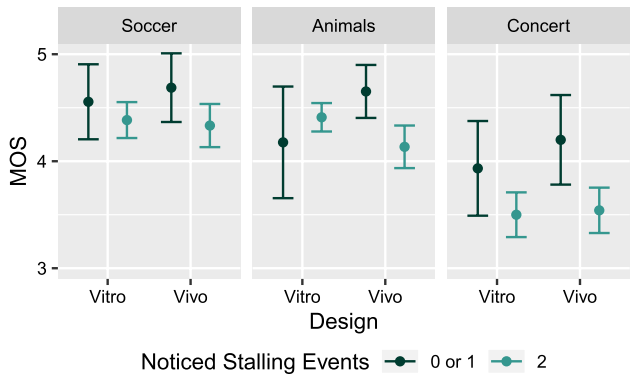
Fig. 3 MOS of perceived streaming quality with 95% confidence intervals rated by participants of sub-study *A* who watched 30s video sequences with no or one stalling event

conditions, i.e., *in vitro* and *in vivo light* with the same stalling patterns, is analyzed using Skillings-Mack tests for unbalanced block design. The tests result in the rejection of the null-hypothesis, meaning that there is a significant effect between the content and the ratings (*in vitro*: $p < 0.001$, *in vivo light*: $p < 0.01$). The pairwise comparison using the method by Conover with Bonferroni correction reveals significant differences only between the concert video and the other videos ($p < 0.001$), regardless of the used design and the number of stalling events. For the concert video, the perceived quality is significantly lower than the ratings of the other videos. This might be caused by a lower video quality due to the amateur recording. Additionally, the self-reported enjoyment factor while watching the video is lower for the concert compared to the other videos which might also affect the ratings negatively. Further, as expected, the occurrence of a stalling event has a negative effect on the perceived quality. This observation is supported by Mann-Whitney U tests for each video (all $p < 0.001$). The analysis of the impact of the *in vitro* or *in vivo light* design, again using the Mann-Whitney U test, shows that there is no significant difference between the ratings in both settings, with and without stalling with $p > 0.05$.

When looking at the results of sub-study *D* having longer videos (60s) and two stalling events in Fig. 4, we have to distinguish between the objective number of stalling events and the subjective, i.e., recognized number of stalling events. Due to the interactive design, participants might be distracted and miss a stalling event or scroll down the page such that the video is no longer visible. Thus, the objective case includes the ratings of all participants independent of the number of noticed stalling events, as long as the participants passed the before mentioned filters and indicated that they have seen no more than two stalling events. In contrast, the subjective case evaluates the ratings with respect to the number of recognized stalling events. Having a look at the



(a) Objective number of stalling events.



(b) Subjective number of stalling events.

Fig. 4 MOS of perceived streaming quality with 95% confidence intervals rated by participants of sub-study *D* who watched 60s video sequences with two 6s stalling events

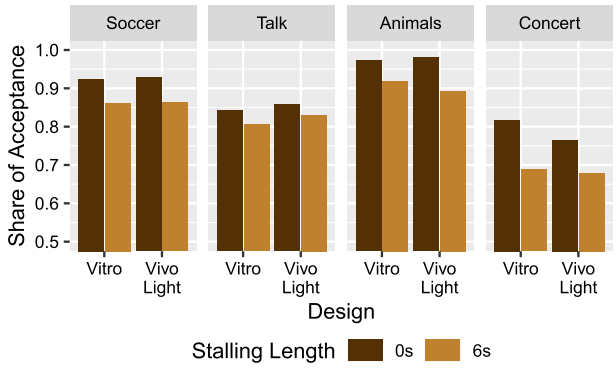
objective number, Fig. 4a shows the MOS for the *in vitro* and the *in vivo* design. Like in the results of sub-study *A*, no significant differences of the ratings regarding the used design are visible, established by a Mann-Whitney U test per video (all $p > 0.05$). Similar trends for the MOS, as observed in sub-study *A*, are visible, despite the fact that the video duration in sub-study *D* was twice as long as in sub-study *A* and two stalling events occurred. Regarding video content, again, there is a significant difference suggested by Kruskal-Wallis tests for both designs ($p < 0.001$). Conover's pairwise comparison with Bonferroni correction reveals, that only the rating values for the amateur video are significantly lower than those for the other videos ($p < 0.001$).

In Fig. 4b, the results are aggregated based on the subjective number of stalling events the participants noticed, with dark green indicating zero or one stalling event and light green indicating two recognized stalling events. The small sample size of participants recognizing less than two stalling events leads to large confidence intervals. When comparing the obtained MOS values for both interface designs, we again do not observe any effect. This observation is supported by non-significant Mann-Whitney U tests applied for each video and number of noticed stalling events with $p > 0.05$. As expected, the perceived video quality was higher when no stalling events were noticed, except for the *in vitro* design of the animals video. Here, the MOS was higher in the case that the participants noticed both stalling events. Mann-Whitney U tests, with Bonferroni-Holm correction due to multiple comparisons applied for each video and design, revealed that these differences are only significant for the animal ($p < 0.01$) and the concert video ($p < 0.01$) watched in the *in vivo* design. Regarding the interface design, we found that there is no significant difference between the MOS values for the *in vivo* and *in vitro* designs. Thus, no significant influence of the *in vivo* and *in vitro* design could be determined regarding the perceived streaming quality.

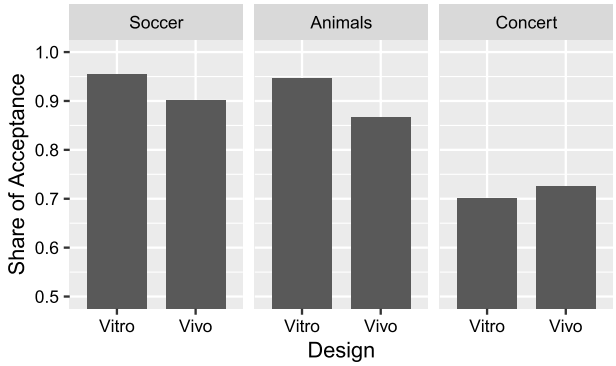
Table 3 Kendall rank correlation of QoE ratings and level of liking the video content

Video	Design	τ	$p - value$
Sub-study A: No stalling, 30 s video			
Soccer	In Vitro	0.35	< 0.001
	In Vivo Light	0.37	< 0.01
Talk	In Vitro	0.36	< 0.001
	In Vivo Light	0.56	< 0.001
Animals	In Vitro	0.38	< 0.001
	In Vivo Light	0.50	< 0.001
Concert	In Vitro	0.51	< 0.001
	In Vivo Light	0.62	< 0.001
Sub-study D: Two stalling events, 60 s video			
Soccer	In Vitro	0.22	0.05
	In Vivo	0.32	< 0.05
Animals	In Vitro	0.32	< 0.05
	In Vivo	0.06	0.64
Concert	In Vitro	0.34	< 0.01
	In Vivo	0.44	< 0.001

To go more into detail, we analyze the Kendall rank correlation of the QoE ratings and the ratings of how much the participants liked the video content. Table 3 shows the result per video of sub-study *A* without stalling and when two stalling events were recognized by the participants of sub-study *D*. In sub-study *A*, we observe a positive correlation between the ratings for both design approaches for all videos. The more the participants liked the video content, the higher they rated the streaming quality. Nevertheless, the correlations for the *in vivo light* design are higher than for the *in vitro* approach, and all correlations are significant. In sub-study *D*, again, a positive correlation can be observed for both interface designs for all videos. However, the results are not significant for the *in vitro* design of the soccer video and the *in vivo* design of the animals

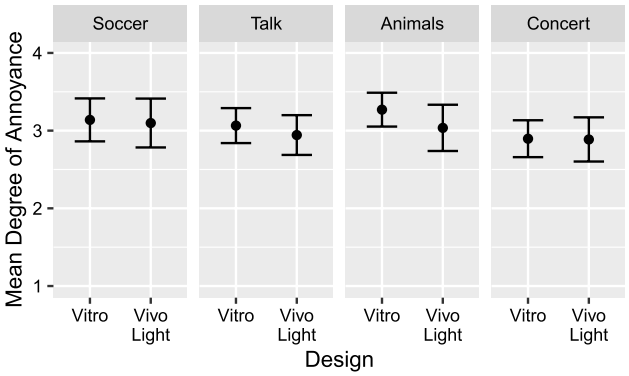


(a) Participants of sub-study A.

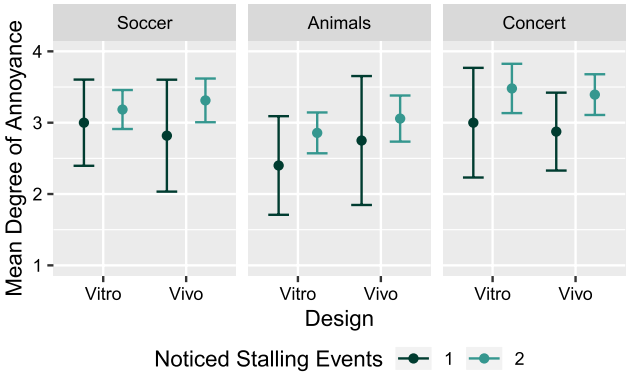


(b) Participants of sub-study D who noticed two stalling events.

Fig. 5 Share of participants who would accept a streaming service with the shown quality



(a) Participants of sub-study A who watched video sequences of 30s with 6s stalling.



(b) Participants of sub-study D who watched video sequences of 60s with two stalling events.

Fig. 6 Mean degree of annoyance with 95% confidence intervals

video. Moreover, in contrast to all other cases, the animals video has a weaker correlation for the *in vivo* design than for the *in vitro* setting. Thus, the impact on perceiving a better quality for liked content might be higher while watching videos in an *in vivo* scenario.

Further, we evaluate the impact of the design on the participants' acceptance of the streaming quality (hypothesis **H2**). The acceptance of the streaming service with and without stalling for a video duration of 30s is shown in Fig. 5a. The results exhibit significant differences between the videos indicated by chi-squared tests for both designs ($p < 0.001$), but only a slight difference between the *in vivo light* and *in vitro* designs. Fisher's exact tests show that this difference is not significant with $p > 0.05$. Regarding sub-study D having a video duration of 60s, the same behavior can be observed for the *in vivo* and *in vitro* designs as shown in Fig. 5b. Again, these differences are not significant, as indicated by an exact Fisher test with $p > 0.05$. Concluding, concerning hypothesis **H1** and **H2**, no impact of the design on the perceived quality as well as on the acceptance was visible.

Influence on the degree of annoyance

Before having a closer look at the influence of the interface design on the degree of annoyance, the influence of the study design (variation or no variation of the interface design per participant) on the annoyance ratings is again evaluated by using annoyance ratings of sub-study A and B. Again, the talk video is not considered. Mann-Whitney U tests computed for each of the remaining videos result in no differences between the perceived annoyance of participants who watched videos with a fixed or varying interface design, i.e., *in vivo light* or *in vitro*, with all $p > 0.05$.

To evaluate the influence of the interface design (*in vivo* or *in vitro*) on the degree of annoyance, we analyze the ratings of sub-study A and D. Figure 6a depicts the degree of annoyance with 95% confidence intervals for the two design approaches for a video duration of 30s. The results are grouped by video content and the values are based on the ratings of the stalling annoyance from participants who recognized stalling correctly. The higher the degree of annoyance, the more the participants are annoyed by the stalling event. Again, Mann-Whitney U tests applied for each video

results in no rejection of the null hypothesis that the samples originate from populations with the same distribution with $p > 0.05$. Thus, we found no effect of the design on the degree of annoyance.

The ratings for a longer video duration with 60 s (sub-study *D*) are shown in Fig. 6b. Here, the results are split by the number of noticed stalling events. Once again, the Mann-Whitney U test results in no rejection of the null hypothesis that the samples originate from populations with the same distribution with $p > 0.05$ and thus, no significant difference between the *in vivo* and the *in vitro* design is visible.

We investigate whether there is a connection between the enjoyment of a video and the degree of annoyance of an occurred stalling. To do so, we employ the Kendall rank to analyze the correlation between the ratings of sub-study *D* regarding how much the participants like a video and the perceived annoyance. Having a look at the correlation for the soccer match in the *in vitro* design, a low positive correlation ($\tau = 0.24$, $p < 0.05$) is found. The more the participants like the video of the soccer match, the more annoyance is perceived due to the stalling events when watching it in the *in vitro* design. In contrast to that, a low negative correlation ($\tau = -0.21$, $p < 0.05$) for the concert in the *in vivo* design is observed. Thus, the more the participants like the video of the concert, the less s/he is annoyed from the stalling in the *in vivo* design. This observation indicates that liking or disliking the content influences the way participants perceive playback interruptions. On the one hand, people liking soccer are interested in the course of the match. Hence, stalling events are more annoying for them. On the other hand, for participants who did not like the concert video, it is less annoying if the video stalls as they are not interested in the content anyway. The differences in the observation may be also slightly influenced by the design and the possibility to interact with the web page. Participants watching the soccer game in the *in vitro* design might be more focused on the content than people watching the concert in the *in vivo* design. However, these findings are highly dependent on the video content as indicated by the no observed correlation between the annoyance ratings and the degree of enjoyment for the other videos.

To sum up, no general influence of the interface design on the degree of annoyance of stalling events is visible (hypotheses **H3**). In connection with the content and possible side effects, such as enjoyment, influences can be recognized, which are however very likely negligible for general investigations of influencing factors.

Influence on focus and interactiveness

In both *in vivo* designs, we provide the possibility to scroll, while in the *in vivo* design users can additionally interact with the web page. By scrolling out of focus and by interacting with the web page, the focus is shifted away from the actual stimulus, i.e., the video with stalling events (hypothesis **H4**). Thus, it is more likely that participants miss a video impairment because of the distraction of the interactions. To analyze this, we have a closer look at the scrolling behavior during the main task in the *in vivo light* / *in vivo* designs since the share of participants who used one of the new interactions was rather low in sub-study *D*.

To investigate if the participants' focus on the stimulus is influenced by the interface design, we first have a look at sub-study *A*. Here, we analyze the user's behavior concerning scrolling based on the tracked mouse positions when watching the videos on the YouTube-like web page. One can expect that participants who scroll are more likely to miss stalling events which may influence the QoE positively. However, only for a small share of participants (14%) scrolling is observed at all. The share of scrolling users also does not differ significantly for all the videos (soccer 12.3%, talk 16.1%, animals 14.9%, concert 13.7%). To analyze the relation between scrolling and missed stalling events, the phi coefficient of correlation is computed between the dichotomous predictor (scrolling/no scrolling) and the dichotomous criterion (stalling noticed/missed) per video. Here, only participants watching the videos with one stalling event are considered. All correlation coefficients are negligibly small, thus, we found no evidence that participants are more likely to miss stalling events due to scrolling.

As the missing of stalling events may depend on the point of time at which the scrolling occurs, we further investigate the scrolling behavior of participants regarding this aspect. We categorize the first scrolling event which occurs while watching the video in three categories, early, mid, and late scrolling. The early scrolling category contains scrolling events which occurs within the first third of the video playback, i.e., the first 10 seconds of a video with a length of 30s. Accordingly, the mid category comprises events in the second third (second 11 to 20) and the late scrolling group in the last third (second 21 to 30). We found that 51.4% of the first scroll events fall in the early scrolling category. Thus, we assume it is more likely that workers miss stalling events at the beginning of a video.

To test this assumption, we use the data obtained in sub-study *A* and *C*. In sub-study *C*, the stalling event occurs earlier, after only 5s of video playback time compared to after 15s for sub-study *A*. The share of scrolling while watching the video in sub-study *C* (11.7%) is slightly lower than in sub-study *A*. However, other than expected, we found no significant correlation between scrolling, the position of stalling events and the cases the stalling event has been missed.

To investigate the effect of the stalling position more in detail and additionally analyze whether the video duration has an influence on the focus time, we consider the data collected in sub-study *D*. We assess focus and interactiveness

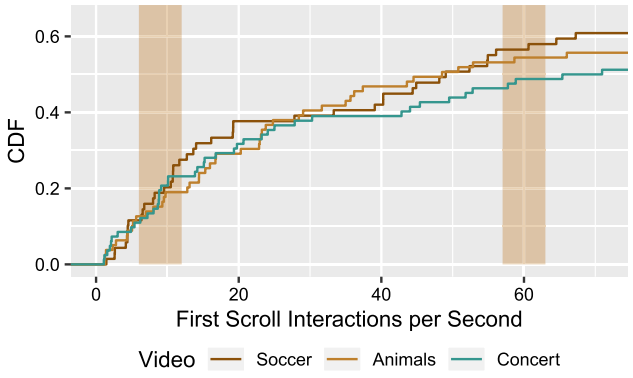


Fig. 7 Cumulative distribution function of first scroll interaction after video playback started with highlighted stalling events

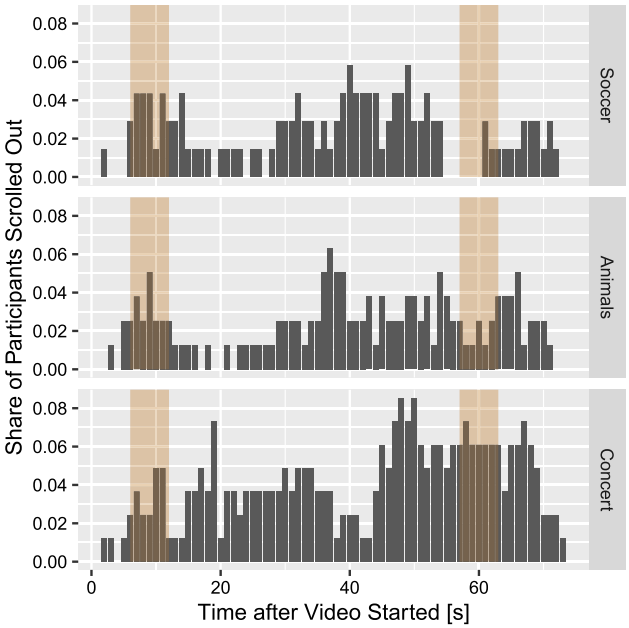


Fig. 8 Share of participants who scrolled the video out of the visible part of the browser window

of participants by analyzing their scrolling behavior and whether they used the different interaction possibilities offered by the *in vivo* design, e.g., like, dislike, comment.

In the following evaluations, we only consider the 230 participants of sub-study *D* who watched a video in the *in vivo* design. From them, 55.6% scrolled the web page during the main task. The much higher number of test takers who scrolled during the playback can be explained by the longer duration of the video on sub-study *D* and the higher sampling rate for tracking the mouse and video position compared to sub-studies *A* and *C*. We analyzed the results statistically and did not identify significant differences among the different videos. A chi-squared test allowed us to conclude that samples originate from the same population ($p > 0.05$). Additionally, 60.1% of the participants scrolled in a way that the video was sometimes no more visible on the screen (further referred as *scrolled out of focus*). Going into detail for each type of content, in the soccer video, most of the participants scrolled the web site (60.9%) and 47.6% of those scrolled out of focus for some time. Having a look at the animals video, 55.7% of the participants scrolled the web site and 59.1% of them made the video invisible occasionally. Moreover, 51.2% of the participants who watched the concert video scrolled the web page and 73.8% of them did it until making the video out focus. A chi-squared test indicated that differences between the scrolling out of focus behaviors for the different videos are significant ($\chi^2(2) = 6.042$, $p < 0.05$).

To evaluate when participants start scrolling the web page, Fig. 7 shows the CDF (cumulative distribution function) of the percentage of people who scrolled for the first time at a specific time after the video playback started. The stalling events after 5s and after 56s are highlighted in light brown. Analyzing the time interval until the first stalling event occurs, already 11.0% to 12.6% of the participants started to explore the website by scrolling. Here, their scrolling behavior does not differ between the videos. During the first stalling event the percentage of people starting to scroll increase to 27.5% for soccer game and to 23.2% for concert. The lowest percentage of first scrolling interactions is observed for the animal video with 19.0%. While first scrolling interactions occur quite frequently during the first stalling event, the percentage of participants who scrolled for the first time during the second stalling event is considerably lower with 1.5% for soccer game, 1.3% for animal video, and 2.4% for concert. As expected, the probability that users start scrolling the web page decreases towards the end of the video playback. Nevertheless, it is unclear if the behavior during the first stalling event is caused by the interruption of the video playback or by participants' web page exploration. By only focusing on the first half of the video (like in the previous sub-studies), fewer participants would have started to scroll the web page: for the soccer match 20.3% fewer, for the animal documentary 10.1% fewer, and for the concert 12.2% workers would start to scroll compared to the full playback time of 60s.

An indicator of the participants' focus is the share of participants who scrolled out of the initial visible part of the browser window. Figure 8 illustrates this behavior over time after the start of each video playback. Again, the stalling event occurrences are highlighted in light brown. As observed, being able to scroll results in some participants losing focus on the stimulus at different times of the video playback in the *in vivo* design. In addition, those test takers who scrolled out at specific time sequences missed stalling events when having the video out of focus. Contrary to our expectations, no clear increase during the stalling times is

visible. On average, 2.7% of the participants who watched the soccer match, 2.6% who watched the animals documentary, and 4.0% of the participants who watched the concert scrolled out of focus. Comparing these shares to the ones obtained during the first stalling events, an increase for all three videos can be seen (soccer 3.6%, animals 3.2%, concert 3.4%). In contrast to that, the share for the second stalling events is lower for the soccer match and the animal documentary (soccer 2.2%, animals 1.9%) and only considerably higher for the concert (concert 6.3%) Thus, like previously mentioned, no clear trend of increased scrolling out of focus caused by stalling events is visible.

To investigate the influence of the interface design on the focus, we compare the share of test takers who missed a stalling event between the *in vivo* and the *in vitro* design. From the participants who passed all the filters, 22.6% of them missed a stalling event in the *in vitro* setting and 28.3% in the *in vivo* design. Although the share is higher for the *in vivo* design, we did not find a significant effect introduced by the interactions by applying the exact Fisher test ($p > 0.05$). This may be explained by the rather low share of participants who used the additional interaction possibilities offered by the *in vivo* design. Only 12.6% of the test takers used at least one of the interaction functionalities. Specifically, 3% of them like or dislike the video, 5.2% expand or collapse the video description, 3.9% tried to add or added a comment, and 5.2% clicked one of the suggested videos. Nevertheless, no significant differences were found for the performed interactions among the three videos when applying chi-squared tests with a significance level of 0.05. Concerning the reasons to click a suggested video, five participants stated that they wanted to explore the web site, four said that the suggested video sounded interesting, and one argued to have clicked intuitively.

The effect of interactiveness can be further analyzed through the relationship between the number of noticed stalling events and interacting with the web page. Conducting a point bi-serial correlation, we found a moderate correlation (0.20) between the number of noticed stalling events and interacting with the web page ($t(228) = -3.15$, $p < 0.01$). By using Kendall's rank correlation, we found that the number of noticed stalling events is negatively correlated ($\tau = -0.33$) with the number of performed interactions, excluding scrolling ($z = -5.3374$, $p < 0.001$). Meanwhile, small correlations were found with the participants scrolling the web page with a correlation coefficient of 0.163 ($t(228) = -2.501$, $p < 0.05$), and scrolling video out of focus with a correlation coefficient of 0.209 ($t(228) = -3.239$, $p < 0.01$) conducting Pearson correlation tests.

In general, our results indicate that interactiveness, mainly scrolling, has an influence on participants' focus (**H4**). If participants start scrolling, they mostly do so at the beginning of the test and thus, start losing focus very early. Participants also tend to perform fewer scrolling and other interactive actions in short videos, compared to longer videos, but if they perform interactions, they also tend to miss stalling events. In a real-life application like YouTube, most videos are longer than 30s and thus, the results observed for the longer videos might be the more realistic behavior of user.

Relation of streaming experience and web page interaction

We still assume that the streaming experience is influenced by interacting with the web page (hypothesis **H5**). To investigate this, we analyze the ratings of videos shown in the *in vivo* design and the interaction behavior of participants of sub-study *D*. Focusing on page interactions (excluding scrolling), we found that participants who interact with the web page perceive a higher streaming quality with a MOS of 4.38. The MOS value for the group of participants who did not interact with the web page is about 4.07. The Mann-Whitney U test establishes that the samples do not originate from the same population ($W = 3540, p < 0.05$). Having a look at scrolling, no significant influence on the streaming QoE can be seen. Participants who scrolled out of focus rated the quality on average slightly higher with a MOS value of 4.16 than participants who kept the video in focus during the playback (MOS 4.08). A Mann-Whitney U test indicated that this effect is not significant with $p > 0.05$.

Besides the influence on the perceived streaming QoE, we also expect an influence of interacting with the web page on the degree of annoyance of stalling. Nevertheless, no difference can be seen between the annoyance ratings of participants who did not interact with the website at all and participants who interact in general, including scrolling and all other interaction possibilities. Furthermore, no difference is found between the degree of annoyance for workers who did not interact and workers who interacted with the website excluding scrolling. Again for both cases, Mann-Whitney U tests result in no rejection of the null-hypothesis with $p > 0.05$. Focusing on the workers who scrolled out of focus, an effect is visible. This effect is visualized in Fig. 9, which shows the mean degree of annoyance with 95% confidence intervals rated by participants who scrolled the video out of focus. Here, the difference between ratings for the concert video is significant, established by the Mann-Whitney U test ($W = 459.5$, $p < 0.05$). The participants who scrolled the video out of focus perceived stalling events as less annoying with an average rating of 2.87, while keeping the video visible results in larger annoyance with a mean degree of 3.39. A same, but not significant, trend is observed for soccer game and the animals video (both $p > 0.05$).

Our results point towards a relationship of the interactions of the user with the web page and the user's perceived video

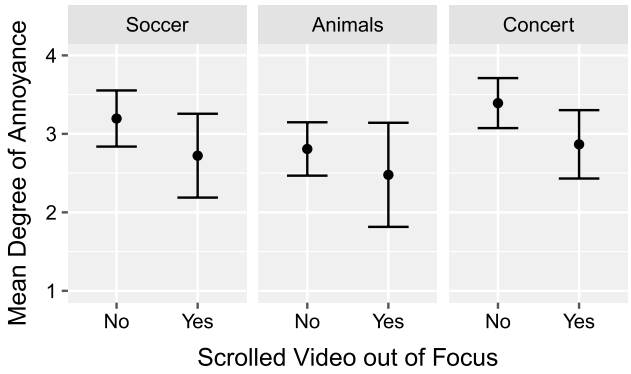


Fig. 9 Mean degree of annoyance with 95% confidence intervals for participants having the video always visible and those who scrolled it out of focus

quality and the perceived annoyance of the stalling (**H5**). However, in our study relatively few participants interacted with the test interface and thus a clear effect of interactiveness of the test interface on the perceived video quality is not observable.

Conclusion

Users' satisfaction with Internet applications and services becomes increasingly important and service providers, consequently, pay more and more attention to the concept of QoE. Subjective user studies are indispensable in this context, to gain insights into the relationship of measurable technical service parameters and user perception. In the area of video streaming QoE research, two main design alternatives are used to conduct such studies. Following an *in vivo* setting, the research aims at providing a realistic and holistic surrounding in addition to the actual stimulus, following an *in vitro* approach only the stimulus is displayed without any further context. Even if numerous works exist using either an *in vivo* or an *in vitro* approach, no direct comparison of both interface designs using the same set of stimuli was performed so far. Our work helps closing this gap and analyzes the influence of the study's interface design on the perceived video quality of participants.

Contrary to our expectations, the results show no significant influence of the interface design on the perceived video quality as well as on the acceptance (**H1** and **H2**). Nevertheless, they are in accordance to findings about the influence of advertisement banners described in [28]. Additionally, no general influence of the interface design on the degree of annoyance of stalling events is observed (**H3**). In connection with the content and possible side effects such as enjoyment, influences can be recognized, which are, however, very likely negligible for general investigations of influencing factors. On the contrary, our results indicate that the interactiveness of the study interface, mainly scrolling, has an influence on participants' focus (**H4**). Participants start scrolling and thus, can lose focus on the stimulus very early in the course of the study. Even if less interactive interfaces and shorter videos can mitigate this behaviour, we argue that these interactions are actually a more realistic behaviour of users and, thus, should be considered in the study's' interface design. Finally, we observe an influence on QoE caused by the way the user interacts with the study interface (**H5**). Further, interactions also tend to influence the degree of annoyance, if the participant scrolls the video out of focus. However, it is difficult to generalize these observations, as only few participants used the interaction possibilities in the current study. We believe that this might result from the fact that the test takers were told to watch a video as part of a paid crowdsourcing task. This setting may lead to an unnatural behavior, in particular, a stronger focus on potential video impairments and less natural interactions with the web page. In a real life streaming environment, we would expect the users to interact more often, and in that case higher influences are to be expected since viewers might be less focused on the stimulus.

In general, the comparison of the subjective ratings for video QoE from the *in vivo* and *in vitro* design indicate that the results of both interface designs are very similar in most cases in terms of MOS values, annoyance ratings, and acceptance. This enables researchers to compare results from existing and further studies on video quality, even if the interface design differed between an *in vivo* and *in vitro* approach. Moreover, the current results suggest that future studies focusing on the impact of stalling can use the simple *in vitro* design instead of the more realistic but also more complicated *in vivo* design, without introducing biases. On the other side, the results also show that researchers can include their stimuli in more complex *in vivo* settings, but then need to monitor if participants still focus on the actual video stimulus.

Future work will address some of the current limitations of this study. The visual stimulus did not contain audio tracks. Thus, stalling events were only perceivable if the video was in the visible area. However, with audiovisual content, users will likely also be aware of stalling events even when reading comments or interacting with the page. Further, even the current *in vivo* setting is far from being a fully functional interface. Therefore, future versions might include even more ways of interactivity. Here, it might be especially interesting to give the test-taker the possibility to select the videos according to their preferences freely or use a more open test phrasing that encourages more interactivity.

Acknowledgements This work is supported by Deutsche Forschungsgemeinschaft (DFG) under Project number: 239765193 (DFG Crowdsourcing). The authors alone are responsible for the content.

Funding Open Access funding enabled and organized by Projekt DEAL.

Compliance with ethical standards

Conflict of interest The authors state that there is no conflict of interest.

References

1. Borchert K, Schwind A, Hirth M, Hoßfeld T (2019) In vivo or in vitro? influence of the study design on crowdsourced video QoE. In: 2019 Eleventh international conference on quality of multimedia experience (QoMEX), IEEE, pp 1–6
2. Van den Broeck W, Jacobs A, Staelens N (2012) Integrating the everyday-life context in subjective video quality experiments. In: Workshop on quality of multimedia experience
3. Cisco: Cisco visual networking index: Global mobile data traffic forecast update, 2016–2021, (2017) San Jose. CA, USA
4. Daniel F, Kucherbaev P, Cappiello C, Benatallah B, Allahbakhsh M (2018) Quality control in crowdsourcing: a survey of quality attributes, assessment techniques, and assurance actions. ACM Comput Surv 51(1):1–40
5. Difallah D, Filatova E, Ipeirotis P (2018) Demographics and dynamics of mechanical Turk workers. In: Proceedings of the eleventh ACM international conference on web search and data mining, pp 135–143
6. Egger-Lampl S, Redi J, Hoßfeld T, Hirth M, Möller S, Naderi B, Keimel C, Saupe D (2017) Crowdsourcing quality of experience experiments. In: Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments. Springer
7. Estellés-Arolas E, González-Ladrón-De-Guevara F (2012) Towards an integrated crowdsourcing definition. J Inf Sci 38(2):189–200
8. Finnerty A, Kucherbaev P, Tranquillini S, Convertino G (2013) Keep it simple: Reward and task design in crowdsourcing. In: Proceedings of the biannual conference of the Italian chapter of SIGCHI, ACM, p 14
9. Gardlo B, Egger S, Seufert M, Schatz R (2014) Crowdsourcing 2.0: Enhancing execution speed and reliability of web-based QoE testing. In: Proceedings of the international conference on communications
10. Guse D, Egger S, Raake A, Möller S (2014) Web-QoE under real-world distractions: two test cases. In: 2014 sixth international workshop on quality of multimedia experience (QoMEX), IEEE, pp 220–225
11. Hirth M, Borchert K, De Moor K, Borst V, Hoßfeld T (2020) Personal task design preferences of crowdworkers. In: 2020 Twelfth international conference on quality of multimedia experience (QoMEX). IEEE
12. Hirth M, Hoßfeld T, Tran-Gia P (2011) Anatomy of a crowdsourcing platform—using the example of microworkers.com. In: Proceedings of the conference on innovative mobile and internet services in ubiquitous computing
13. Hirth M, Hoßfeld T, Tran-Gia P (2013) Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms. Math Comput Model 57(11–12):2918–2932
14. Hirth M, Scheuring S, Hoßfeld T, Schwartz C, Tran-Gia P (2014) Predicting result quality in crowdsourcing using application layer monitoring. In: 2014 IEEE fifth international conference on communications and electronics (ICCE), IEEE, pp 510–515
15. Hossfeld T, Keimel C, Hirth M, Gardlo B, Habigt J, Diepold K, Tran-Gia P (2014) Best practices for QoE crowdtesting: QoE assessment with crowdsourcing. IEEE Trans Multimedia 16(2):541–558
16. ITU-T Recommendation P (2008) Subjective video quality assessment methods for multimedia applications. International telecommunication union
17. Jumisko-Pyykkö S, Hannuksela MM (2008) Does context matter in quality evaluation of mobile television? In: Proceedings of the 10th international conference on Human computer interaction with mobile devices and services, ACM, pp 63–72
18. Kazai G, Kamps J, Milic-Frayling N (2011) Worker types and personality traits in crowdsourcing relevance labels. In: Proceedings of the conference on information and knowledge management
19. Keimel C, Habigt J, Horch C, Diepold K (2012) Qualitycrowd—a framework for crowd-basedquality evaluation. In: Proceedings of the picture coding symposium
20. Ketykó I, De Moor K, De Pessemier T, Verdejo AJ, Vanhecke K, Joseph W, Martens L, De Marez L (2010) QoE measurement of mobile YouTube video streaming. In: Proceedings of the 3rd workshop on mobile video delivery, MoViD '10, pp 27–32
21. Kietzmann JH (2017) Crowdsourcing: a revised definition and introduction to new research. Bus Horizons 60(2):151–153
22. Le Callet P, Möller S, Perkis A (2012) (eds.): Qualinet White Paper on Definitions of Quality of Experience. European network on quality of experience in multimedia systems and services (COST Action IC 1003), Lausanne, Switzerland
23. Martin D, Carpendale S, Gupta N, Hoßfeld T, Naderi B, Redi J, Siahaan E, Wechsung I (2017) Understanding the crowd: ethical and practical matters in the academic use of crowdsourcing. In: Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments. Springer
24. Nam H, Kim KH, Calin D, Schulzrinne H (2014) YouSlow: a performance analysis tool for adaptive bitrate video streaming. SIGCOMM Comput Commun Rev 44(4):111–112
25. Nam H, Kim KH, Schulzrinne H (2016) QoE matters more than QoS: Why people stop watching cat videos. In: Proceedings of the international conference on computer communications
26. Seufert M, Egger S, Slanina M, Zinner T, Hossfeld T, Tran-Gia P (2015) A survey on quality of experience of HTTP adaptive streaming. IEEE Commun Surv Tutorials 17(1):469–492
27. Seufert M, Hoßfeld T, Sieber C (2015) Impact of intermediate layer on quality of experience of HTTP adaptive streaming. In: Proceedings of the conference on network and service management, pp 256–260
28. Seufert M, Zach O, Slanina M, Tran-Gia P (2017) Unperturbed video streaming QoE under web page related context factors. In: Proceedings of the conference on quality of multimedia experience
29. Simperl E (2015) How to use crowdsourcing effectively: Guidelines and examples. Liber Quart, 25(1):18-39
30. Staelens N, Moens S, Van den Broeck W, Marien I, Vermeulen B, Lambert P, Van de Walle R, Demeester P (2010) Assessing quality of experience of IPTV and video on demand services in real-life environments. IEEE Trans Broadcast 56(4):458–466

31. Wamser F, Seufert M, Casas P, Irmer R, Tran-Gia P, Schatz R (2015) YoMoApp: A tool for analyzing qoe of YouTube HTTP adaptive streaming in mobile networks. In: European conference on networks and communications. IEEE
32. Xue J, Chen CW (2012) A study on perception of mobile video with surrounding contextual influences. In: 2012 Fourth international workshop on quality of multimedia experience, pp 248–253
33. Zach O, Seufert M, Hirth M, Slanina M, Tran-Gia P (2017) On use of crowdsourcing for H. 264/AVC and H. 265/HEVC video quality evaluation. In: 2017 27th international conference radioelektronika (RADIOELEKTRONIKA), IEEE, pp 1–6
34. Zhu Y, Heynderickx I, Redi JA (2015) Understanding the role of social context and user factors in video quality of experience. Comput Human Behav 49:412–426
35. Zinner T, Hirth M, Fischer V, Hohlfeld O (2016) ERWIN-Enabling the reproducible investigation of waiting times for arbitrary workflows. In: Proceedings of the international conference on quality of multimedia experience