IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 November 2022, accepted 15 December 2022, date of publication 22 December 2022,
date of current version 29 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3231747

SURVEY

Long-Term Video QoE Assessment Studies: A Systematic Review

NATALIA CIEPLIŃSKA[1], LUCJAN JANOWSKI[1], KATRIEN DE MOOR[1,2], (Member, IEEE), AND MICHAŁ WIERZCHOŃ[1,3,4,5]

[1]Faculty of Computer Science, Electronics and Telecommunications, AGH University of Science and Technology, 30-059 Kraków, Poland
[2]Department of Information Security and Communication Technology, NTNU, 7034 Trondheim, Norway
[3]Consciousness Laboratory, Institute of Psychology, Jagiellonian University, 31-007 Kraków, Poland
[4]Centre for Brain Research, Jagiellonian University, 31-007 Kraków, Poland
[5]Jagiellonian Human-Centered Artificial Intelligence Laboratory, Jagiellonian University, 31-007 Kraków, Poland

Corresponding author: Natalia Cieplińska (cieplins@agh.edu.pl)

The research leading to these results has received funding from the Norwegian Financial Mechanism 2014-2021 under project 2019/34/H/ST6/00599.

ABSTRACT Although longitudinal studies are common in other disciplines, such as psychology, medicine, or User Experience, they are rarely used in Quality of Experience (QoE). However, observing users over time can provide useful information on the QoE and help to better understand its influencing factors. Here, we present a systematic review of the methodologies of longitudinal studies in the QoE domain. We review papers selected through a systematic search and discuss various aspects of described studies, such as methods of gathering subjective assessment or length of the studies. Our work recognizes common practices that can be used to reproduce and extend the proposed long-term study designs. Additionally, based on our review, we propose future work directions for the longitudinal QoE studies.

INDEX TERMS Quality of experience, longitudinal study.

I. INTRODUCTION

Recently, the research community is facing the so-called replication crisis in science [11], [14]. There are multiple examples of studies (e.g., pertaining to key studies within the field of psychology) that were not replicated, even though they use the same set-up [48]. An important implication of this fact is that one should not conclude based on a single study and take its result as ground truth. Instead, researchers should accumulate findings from multiple studies and draw conclusions by reviewing, aggregating, and comparing them. One of the ways to do this is to conduct a systematic review of the relevant literature.

A systematic review is different from the state-of-the-art presentation that is typically included in many publications, as described in the overview of conducting systematic review [15]. The main advantages of the systematic review are: the detailed description of the search process, and the fact that conclusions are based on a number of papers with the similar claims. The formalized search process enables the replication and validation of the systematic review. The number of publications supporting a given claim help allows to indicate solid findings as well as to detect those potentially under-powered or false.

In telecommunication technology, systematic reviews are not common. This is probably due to rapid changes and developments of technologies. We found a few examples of the systematic reviews of the Quality of Experience (QoE) or living labs studies [17], [18], [19]. As QoE measures the delight or annoyance of a customer's experiences with a service it is an important aspect, especially for service and network providers. However, to our knowledge, no systematic review scrutinizes longitudinal QoE studies. Such a systematic review is the key focus of this article. More precisely, the main contribution of this paper is a systematic review of methodologies of long-term video QoE studies. We summarize the methods used in the longitudinal studies assessing video quality judgements with human subjects and identify future work directions to better understand the variability of the QoE judgements over time.

A Systematic review approach seems especially advisable for longitudinal QoE studies, because quality assessment over a longer period of time is not common. There is no standardized methodology of this kind of study, resulting in a range of different methodological designs that are difficult to compare. Most of the QoE studies are cross-sectional,

The associate editor coordinating the review of this manuscript and approving it for publication was Alessandro Floris.

i.e., they conduct experiments in a single session. Such solution is less time-consuming for the researchers, and does not require a large budget to pay the participants. However, observation of the QoE over time is an important aspect of user experience with a service. In the real life conditions, users typically interact with services for long periods of time, e.g., when using a streaming platform daily. This prolonged interaction may change the importance of certain features and help to identify relevant variables in a more robust and ecologically valid way. Ecological validity is defined as a degree of correspondence between the research conditions and the phenomenon being studied as it occurs naturally or outside of the research setting. As an example, the tolerance to video stalling in the lab may be larger when it happens unnaturally than when someone is using a video service over a longer period of time and mostly experienced video without stalling. A prolonged interaction with a service may also reveal an increased number of factors influencing a user's experience. Depending on the study design, it can be beneficial to observe the effect of these factors when they can be captured and analyzed meaningfully. The long-term studies may also just increase the results' variance, making them more comparable to the real world situation.

Long-term QoE studies allow to understand what influences the experience in more natural settings. Prolonged users' observation enables insight into their behavior. Use of a service on the macro level can be assessed in terms of behavioral economics, and may result in business [12]. As a result, in addition to research purposes, long-term QoE studies are also important for service providers. This is because the method is more ecologically-valid and thus the results are better correlated with churn [49] and crucial customer decisions (i.e. when user is purchasing additional service or recommending it to other people).

Given the mentioned considerations, the goal of the paper is to summarize the current state of longitudinal QoE studies in the context of video services evaluations, systematically review the results of the previous studies and discuss future directions of studies. More precisely, relevant, long-term QoE studies related to the video quality evaluation are described in terms of their methodology, e.g., with regard to how the subjective judgments were gathered, what was a study duration, which factors influencing the QoE were investigated, and which directions of studies are still not covered.

The remainder of the paper is organized as follows: the background of longitudinal studies is shortly described in the following section. The methodology behind this systematic review is shared in Section III. Section IV describes the data, methods, and results gathered from the analyzed papers. Finally, Section V is dedicated to the discussion, and Section VI to the final conclusions.

II. STATE OF THE ART

Because systematic review articles are not often proposed in the QoE community, let us start with presenting how one can properly describe and systematize previous work. First, we present the types of review articles and methodologies (more or less formal) applied within those articles. Then, we describe the types of longitudinal studies we aim at reviewing.

A. TYPES OF THE REVIEW ARTICLES

The progress of science depends on a current consensus on the scientific facts [52]. The consensus can be reached through the analysis of numerous papers describing the phenomena of interest. Such an analysis is usually done with review papers. A well-prepared review paper can clarify the state of knowledge on the phenomena of interest, notice contradictions in the results or methodologies, and identify missing questions that should be answered. The nine main types of review articles identified by [55] are:

1) *Literature review* lists current literature on a phenomena of interest selected by an author, but does not require formal, systematic search rules;
2) *Scoping review* evaluates the size and scope of literature available on a particular topic. It aims at presenting the overview of the literature, but does not provide formal quality assessment of picked studies;
3) *Critical review* evaluates the quality of cited resources, but does not apply systematic search rules or formal quality assessment;
4) *Systematic review* applies systematic search rules and follows a protocol, which defines formal method for selecting the studies. It it usually exhaustive, but may have a selection bias, because the search rules (e.g. keywords) and the protocol (e.g. exclusion criteria) are defined by a researcher. However, this could always be corrected because the protocol is clearly defined and published;
5) *Meta-analysis* develops a precise, statistical summary of multiple quantitative studies. The statistical analysis requires that all the studies have the same (or very similar) measures;
6) *Mapping review* delineates and classifies existing literature on a given topic, but is usually time-constrained and does not apply a systematic search rules;
7) *Qualitative systematic review* integrates and compares outcome of qualitative studies but does not use any specific methodology;
8) *Meta-synthesis* aims to explain a particular phenomena;
9) *Realist review* aims to investigate complex interventions, e.g. performance measures, regulation and inspection, or funding reforms, and discovers and identifies the reasons for thriving or failing of the intervention;
10) *Umbrella review* extracts results from multiple reviews.

Reference [55] also provides examples of each type of the review article. Note that the list is not necessarily exhaustive - e.g. meta-analysis can apply very different approaches and types of datasets (meta-analysis of meta-analyses is also sometimes performed). Nevertheless, it is important to select

the type of review article, keeping in mind the goal we want to achieve. Our main aim was to enable validation and replication of our review, so we have decided to follow a systematic review structure.

B. LONGITUDINAL STUDIES IN QoE

Before exploring the details of the studies reviewed in this paper, let us introduce the main approaches used in QoE to understand changes in users' experience and behavior over time. According to [44], there are three dominant approaches. The first and most commonly used is the, *cross-sectional approach*. It is a type of observational study that analyzes data from a population, or a representative subset based on data collected at one, specific point in time. The change of experience in time is probed by comparing different user groups (e.g., with different levels of expertise). Cross-sectional studies are usually focused on a very specific aspect measured in isolation. However, the approach has some serious limitations [46]. For example, participants are not observed or queried over a time, thus the context and human factors [45] increase the observed variance. To be more specific, changes in a single user behavior cannot be observed.

The second approach to probe changes over the time is *within-subject repeated sampling approach* [44]. It assumes that the same users should be tested twice - at the beginning and at the end of a study. Because these types of studies are based on only two tests, it may not correctly recognize time effects due to random contextual variation. It may also not correctly capture the changes in time that are not linear (the second test could be too early or too late to show the temporal effect of the manipulation of interest).

Finally, *longitudinal approach* [44] is based on more than two measurements. Therefore, it provides more insight on how the experience changes over time. In the longitudinal studies one may also identify factors that are constant regardless of a situation. The factors that can be observed with the longitudinal approach could be [45]:

- Human related (i.e. expectations, needs, mood) - for example, Using qualitative methods such as interviews, testers could be encouraged to provide feedback on their experiences and expectations and their changes over a time. This way the researchers can get more information about the changes in testers experience over time;
- Context related (i.e. physical, temporal and social) - for example an environment, time of a task performance, etc. can change human experience and behavior. By tracking these data longitudinally, we observe how the context affect the experience.

Observing users over time gives an unique opportunity to investigate memory effects. Memory effects are very important for service providers, as the memory may affect users' motivation to change the provider, especially if they are annoyed with a service. According to [6] and [50], the memory of a negative experience at a specific time point depends on:

- The strength of memory - the more annoying an event, the stronger the user memorizes it and the longer memory lasts;
- The time that has passed since the occurrences of an event;
- Repetition - the more frequently an event occurs, the more likely a user remembers it.

Longitudinal experiments could be performed in the so-called *living lab* environment [47]. As opposed to typical laboratory experiments, living lab studies aim to put end-users and the evaluated QoE components (i.e. network, apps, etc.) as close as possible to their daily media usage scenarios and natural environment. As a result, they provide more representative evaluations of quality [31]. This approach allows to identify real-life user's behavior that is not influenced by an experimental situation (i.e. feeling of being observed or judged, sitting in a laboratory environment, not using their own devices). Such effects are unavoidable in a typical laboratory setting. For these reasons, the relevance of living lab environment in QoE studies seems well justified and needs further exploration. Studies using this approach are further investigated in this review.

III. METHODOLOGY

Conducting a systematic review requires a systematic and thorough documentation on how we selected the papers to be analysed. Thanks to the documentation, it is possible to validate the results and replicate the process, e.g., with an update. Some of the key features of the systematic review, as defined in [13], are:

- clearly defined search strategy that should detect as much of the relevant literature as possible.
- clear documentation of the search strategy so that readers can access its rigour and completeness.
- explicit inclusion and exclusion criteria that could be applied to assess primary studies.

To meet those criteria, and to select the primary papers presenting long-term studies investigating QoE, we used a 7-step search procedure.

1) **The first step** of the search protocol was to check several preliminary queries in three different databases (IEEE, ACM and Scopus) and to review a number of the resulting papers. To develop the queries, we analysed whether given keywords allow to deliver 18 target papers, selected based on prior knowledge of the field [1], [2], [3], [4], [5], [7], [8], [9], [24], [25], [26], [27], [28], [33], [34], [35], [38], [41]. We will call those 18 papers "Related". 10 of them were finally included in our systematic review. The other eight were closely related to the field and their citations helped us finding other relevant papers. However, we have not included them in the systematic review, because they were not related to video quality. This step lead to the selection of the following keywords: quality of experience, QoE, subjective, multi-episodic, perceived quality, user experience, long term, living lab, user

engagement, longitudinal, macro-temporal, over time, prolonged use. We also excluded medicine-related keywords, like: patient, disease, health, since papers with such keywords typically describe longitudinal studies investigating patients' health, but not the role of media services or videos. We ended up this step with the final list of keywords.

2) The second step of the search procedure, was to **pick 10 queries** from ACM, IEEE and Scopus databases and **analyse the abstracts** of 20 random output papers from each query to see how many of the papers are relevant (for the criteria see below). We discarded queries with over 300 results if the number of relevant papers was lower than 30%. As the result, we finally applied eight queries to select the papers.
3) The third step was to **merge the papers** from those eight queries and to remove duplicates. This resulted with an initial database of 318 papers.
4) In the forth step involved screening the abstracts of the above-mentioned papers by two researchers independently. We **divided the papers** into 2 categories: ''Not Relevant'' and ''Relevant''. The latter was further divided into ''Possibly Relevant'' and ''Definitely Relevant''. A paper was considered ''Not Relevant'' when:
   - the topic was related to medicine
   - reported study did not involve users/human subjects (e.g., a study based on simulations but not containing actual subjective evaluations was excluded)
   - reported study does not fall into the category of long-term study or was not connected to QoE (e.g., LTE, which appeared in a search, because its name consists of ''Long Term'' but is not relevant for this research)

   ''Definitely Relevant'' papers were those describing the methodology of long-term studies or describing long-term experiments related to video. ''Possibly relevant'' papers were those describing UX-focused long-term experiments, user engagement and living labs. A paper was marked as ''Relevant'' or ''Not Relevant'' based on its abstract. As the result of this step, we identify 32 ''Possibly Relevant'' papers and 19 ''Definitely Relevant''.
5) In the fifth step we **used the Connected Papers tool** [10]. The tool uses a similarity metric based on Co-citations and Bibliographic Coupling. Using *Connected Papers* one can select those with the strongest connection to an input paper. We observed QoE has no standardised keywords, so similar studies can be named differently (for example ''living labs'', ''mobile human'', ''longitudinal'' and ''multi-episodic'' keywords refers to the same methodology). This is why the step was so important, as it adds relevant publications that might have been overseen during the initial query search. We uploaded all papers marked as ''Surely Relevant'' and ''Related''. Then we downloaded all papers found by *Connected*, *Prior* and *Derivative* search functions of the *Connected Papers* tool [10]. *Connected* function selects those papers which have similar citations to the publication we uploaded, *Prior* function selects those which were cited by many connected papers, and *Derivative* function selects the papers which cite many of the connected papers. Finally, we arrive at 1263 papers from Connected Papers tool, excluding duplicates.
6) In the sixth step, to further narrow the selection, we **analysed only those papers which appeared at least three times in the fifth step of our procedure** in any of the categories (*Prior*, *Derivative* or *Connected*). We selected this threshold as we have noted in the preliminary screening that papers with less than 3 appearances had no connection to longitudinal QoE. We also assumed the number of papers selected at this step should not exceed the number of the papers selected in the main query. Using the threshold we arrived at 206 connected papers and analysed their abstracts. Using the judgement of relevance identical to the one described above we finally selected 37 new ''Definitely Relevant'' papers.
7) Finally, in the seventh step, we **analysed in details all the papers marked as "Related", "Definitely Relevant" and "Possibly Relevant"** (106 papers in total). We discarded papers which did not describe video services and those, which did not use longitudinal approach, but measure quality in one session. The keyword ''video'' was not used in each query, as in most cases it generated too few results. Finally, we obtained 22 papers describing long-term studies of video QoE assessment. The result of our review is described in the next section of this paper.

IV. RESULTS

Longitudinal studies may investigate multiple aspects of QoE. To make the review more comprehensive, we divided the reviewed studies into 3 categories defined based on a task users performed - *using an app*, *audio-visual calls*, *watching videos*. Below, we analyze most important aspects of the studies in the following subsections:

- ''Focus of the studies'' provides general information, such as task description, number of participants and the length of the study.
- ''Objective data'' lists technical aspects measured (e.g. those related to the network)
- ''Subjective data'' specifies methods used for gathering the subjective judgments
- ''Results'' sums up results obtained in a given study

Each of these sections was divided into 3 parts, depending on a task users performed.

A. FOCUS OF THE STUDIES

The reviewed papers described 3 types of studies. We identified the study types according to the type of the activity the

testers were performing during the study. *Using an app* studies are based on the observation of the users' interactions with applications. Typically, participants were asked to use their phones as usual, and questioned about their QoE after using a specific application. *Watching videos* studies are focused on watching a pre-selected content provided by the researchers. Typically, movie trailers or fragments of TV series were used, and the video quality was controlled. Finally, *audio-visual calls* studies are based on video calls under pre-specified conditions. The participants were given a scenario of their video call. We describe in more detail each type of study in the subsections below.

1) USING AN APP STUDIES

In most of the *using an app* studies, except two ([32], [37]), the quality of the videos and network limitations were not controlled. If they were, either down-link traffic was passively modified through traffic shaping, participants were provided with specific Internet access connections for the study [32]. In the second paper with controlled conditions, a customized video player was implemented to simulate artifacts (e.g. rebuffering and change in video quality) and to record video streaming performance in the background [37]. When the network was not controlled the users were using the app in any environment (e.g. at home, outside, in the bus), and using the network available (either Wi-Fi or data connection).

Multiple papers ([21], [22], [23], [29], [31]) monitored the user's interaction with YouTube. The studies used the YoMoApp (YouTube Performance Monitoring Application) - which make use of the YouTube mobile website and the YouTube HTML5 API to exactly replicate the YouTube service. YoMoApp employs HTTP adaptive streaming technology based on bit rate (BE) adaptation. It additionally monitors and stores multiple Key Performance Indicators (KPIs) of the video streaming (such as e.g., player state, events and video quality level) via the YouTube API [42]. Other papers ([24], [25], [26], [27], [28], [30]) describes participants' interactions with different mobile phone applications. For example [24], [27], and [28] make use of the mQoL-Log application, which allows to monitor users' activity and trigger surveys after specific events. References [25] and [26] reports the studies with a similar methodology. However, two separate applications were used: (1) the Android Context Sensing Software (CSS) monitoring users' activity at their phones and (2) a customized survey application. Similarly, [30] reports a study using two customized applications: a traffic monitoring tool and a web-based application providing QoE feedback. Finally, [20] describes a study using a public display situated on university's corridor. Students were asked to look at the display and participate in some interactions with its content (e.g., choose a lecture-related video to watch, take part in a quiz).

Table 1 depicts the reviewed studies duration, type and the number of testers involved. As can be seen, the duration of most of the *using an app* studies was two to four weeks. A study involving YoMoApp was longer (four years) and had higher number of participants (360). It is important to mention that the YouMoApp study did not require participation of a given tester throughout the entire duration of the study. Thus it is difficult to compare it to other studies described in this paper. Majority of the studies involved 30-38 participants, with the exception of [24] and [20] reporting studies on a few participants. Only [29] and [31] mentioned the participation fee (participants received vouchers).

TABLE 1. Duration, type of study and number of testers.

Type	Duration	No. of testers	References
using an app	1431 days	360	[21]–[23]
	28 days	38	[28]
	28 days	33	[27]
	28 days	30	[25], [26]
	28 days	5	[24]
	14 days	33	[32]
	14 days	30	[29], [30]
	14 days	29	[31]
	7 days	70	[37]
	6 days	8	[20]
watching videos	14 days	21	[33], [35]
	14 days	20	[34], [35]
	7 days	29	[38]–[41]
audio-visual calls	14 days	56	[33]
	12 days	56	[36]

2) WATCHING VIDEOS STUDIES

In this type of studies, the testers were requested to watch a pre-selected content (e.g., cut scenes from TV series, soap operas or movie trailers). The quality was controlled. Typically, two conditions have been introduced: high- and low-performance quality.

Paper [35] presents two studies. In both performance levels were manipulated daywise (High Performance (HP) or Low Performance (LP)). Scenes from TV series were used. In the first study HP was defined as 2 Mbit/s and LP's 125 kbit/s bandwidth respectivelly (encoded with h.264 at 720×576px resolution). Presented content were 12-17 minute scenes from ''Friends''. Second study make use of 3 Mbit/s HP and 0.25 Mbit/s and LP bandwidth respectively (encoded with h.264 at 800×480 px resolution). The presented content consisted of 8-12 minute scenes from ''The Big Bang Theory''.

Paper [33] report a study with 13-17 minute fragments of soap operas. Again, HP and LP conditions were applied. In the HP, audio was encoded with MP4A at 320 kbps and video with h264 at 5 Mbps at a device resolution. In the LP, the audio was encoded using GSM full-rate and video bandwidth of 250 kbps. In the paper [34] testers had to watch a 15-minute movie. The HP video bandwidth was 2 Mbit/s, and LP video bandwidth was 125 Kbit/s.

In studies [38], [39], [41] the participants were asked to watch pre-defined videos (2-3 minutes movie trailers) using a mobile device (Nexus One phone running Android 2.1) in their natural environment. The first two papers describe video quality using:

- Seven low-quality videos using the Real-time Transport Protocol (RTP);

- Seven high-quality videos using RTP;
- Seven low-quality videos using progressive download (based on HTTP);
- And seven high-quality videos using progressive download.

Studies presented in this category lasted 7-14 days and number of participants was between 20 and 29. Studies [38], [39], [41] offered compensation in a form of a gift voucher (worth 10 Euro). Study [35] offered 40 Euro participation fee and study [34] 70 Euro participation fee.

3) AUDIO-VISUAL CALLS STUDIES

Following our exclusion criteria, only two reviewed papers belong to the audio-visual calls category. In both studies, participants had to carry out at least two conversations daily (first between 6 am and 3 pm; second between 3 pm and 12 am) using a purposely modified Skype client. The calls were run in pairs. A test participant from each pair received a scenario to ensure the conversations have approximately the same length and structure. The scenario involved participants in a role-play tasks based on everyday situations (e.g., train ticket booking, appointment with a doctor, ordering a pizza). Only about a half of the information about the scene was available to a single participant, so bidirectional conversations were necessary to resolve the tasks proposed in the scenario [43]. The calls were run with modified Skype software, where the joint audio-and-video bandwidth was artificially restricted up to a certain maximal value (62500, 18750, or 4000 bytes per second) for a given day. Both studies reports the results of 56 participants. The duration was around two weeks [33], [36]. As visible in Table 1, the studies of this type lasted 1 or 2 weeks, and all except one study had between 20 and 30 participants.

B. OBJECTIVE DATA

The studies presented in all analyzed papers, except five ([32], [33], [34], [35], [36]) collected technical data from participants. Data can be divided into 4 categories: network-related (e.g., data throughput or video packet loss rate), activity-related (e.g., name of the opened app or screen orientation), cell-related (e.g., signal strength or operator) and context-related (e.g., GPS location or brightness). Tables 2 and 3 present an overview of data collected in the studies. For the *'audio-visual calls'* experiments, no technical data was collected, as the quality of the calls was controlled.

We observed that most of the studies gathered context data - either only GPS position, or the related activity (walking, sitting or in a vehicle while performing a task).

Studies from *using an app* category often tracked screen size, orientation and apps used simultaneously. They also gathered network and cell related data (e.g., Wi-Fi level, connection provider, and network usage).

Watching videos studies gathered mainly information about which video the tester watched at the time, the quality, and session start and end times. They also recorded connectivity data (e.g. network type, loading time).

C. SUBJECTIVE DATA

This section describes what questions were the testers asked and how the feedback about their experience was collected (via questionnaire, interviews etc.). Table 4 presents summarized subjective data from all the papers. During almost all of the studies, the researchers ask the testers about overall quality. Eight studies ask about acceptability of the session or service, six ask about context (either social or location). Questions about multi-episodic quality appear in only 4 of the papers.

1) SUBJECTIVE ASSESSMENT FOR USING AN APP STUDIES

As mentioned before, some of the studies were focused on users' interaction with YouTube specifically, and some were focused on users' interaction with a multiple applications (for example Facebook, Spotify, WhatsApp etc.). YouTube related questions were mostly focused on the quality itself (sometimes some specific aspects, as quality of streaming or service acceptability). On the other hand, studies that focused on interacting with multiple applications asked about users' expectations and what action did they user try to accomplish in this app. During some of those studies interviews were conducted. Different methods are described below.

In the studies that focused on observing user interactions with YouTube, subjective user feedback was expected after each session. In papers [21] and [37] the participants were only asked about overall quality on 5-point ACR scale. In papers [22] and [23], the questions included the user's feedback on the quality of the video, the quality of the streaming, the user's opinion on the video content (using ACR scale), as well as the service acceptability (here the possible answers were yes/no). In the study described in the paper [29] participants were asked to rate the overall quality on a 5-point ACR scale and to indicate whether the session quality was acceptable. On the same scale, they had to indicate to what extend they were annoyed by the initial delay. They were also asked if they noticed any interruptions or stops during the streaming. If yes, they had to indicate whether they experienced these interruptions as annoying on the same scale as for initial delay. In the experiments from [30], [31], and [32], the testers were asked to rate overall experience and acceptability after each session. Additionally, in paper [32], participants indicated their location at the moment of watching the YouTube video (e.g. at home, in the metro, walking). In the studies described in papers [25] and [26] the questions were about: QoE (using 5-point ACR scale), location, social context, mobility (e.g. moving, sitting). They also conducted a weekly interview based on Day Reconstruction Method (DRM). The interview was conducted after completion of a detailed diary of the previous 24-hour period. In the study described in paper [24] only one question was asked: What action were you trying to accomplish?

In the papers [27] and [28], the questions the testers answered after using an app were as following: 1. Did your usage of this app went as expected? 2. How was your last usage session of this app (on a scale 1-5)? 3. What action were

TABLE 2. Objective data collected in *using an app* studies.

paper	network	activity	cell	context
[20]	-	interactions with quizzes and videos that led to results/ratings in the database, page views of the mobile quiz/ratings web apps, and registrations to the system	-	-
[21]	network usage, the amount of downloaded and uploaded bytes on the device	playback, screen size,screen orientation, volume, player size, and player mode(normal/full screen)	changes of operator, RAT, cell ID, signal strength	GPS position
[22]	network usage, the amount of downloaded and uploaded bytes on the device	playback, screen size, screen orientation, volume, player size, and player mode(normal/full screen)	changes of operator, RAT, cell ID, signal strength	GPS position
[23]	network usage, the amount of download and upload bytes for the device	playback, screen size, screen orientation, volume, player size, and player mode(normal/full screen)	changes of operator, RAT, cell ID, signal strength	GPS position
[24]	basic service set identifier and service set ID (network name) of WiFi Access Point, routing tables, IP address, domain name server, ping to app server, packet and kilobytes send and received per an interface	package name of app on the user's screen, number and distribution of user touch while interacting with the screen, per a user session	signal strength, cell ID, operator name, network code and network time of the connected and neighbour cells	GPS position, user physical activity from the Google Play Services Activity(still, tilting: between two states,in vehicle, on bicycle,on foot, running)
[25]	sent/received data throughput (in kilobytes per second)	number of calls and short message service messages (SMSs), running applications, screen orientation	cellular network, Bluetooth, WiFi connectivity and its received signal strength indication (RSSI)	GPS position, location (walking, home etc), acceleration, brightness
[26]	current running applications with total amounts of application data throughput, (i.e., Bytes/second) sent and received	phone activity for interpersonal interaction in terms of number of calls, SMS and MMS, screen orientation	wireless access network technology (cellular or WLAN), cell-ID (for 2.5G, 3G or 4G) an Access Point name (for WLAN), wireless access signal strength (RSSI), Bluetooth network status	GPS position, location (walking, home etc), device's proximity status, device acceleration, magnetometer and orientation
[27]	internet connection status, netstat (TCP network statistics), IP address, proxy information, domain name, number of packet and bytes send and received on each wireless interfaces DNS's IP address, routing table information	application name on the user screen, number of user touches on the screen and duration during an usage session	WiFi level, WiFi BSSID, WiFi SSID, WiFi interface speed, cell ID, cell operator, cell strength, cell Radio Access Technology, cell network code, cell bandwidth up and down stream, battery state(e.g. charging, full, discharging), battery level, battery temperature	GPS position, user physical activity from the Google Play Services Activity(still, tilting: between two states,in vehicle, on bicycle,on foot, running)
[28]	internet connection status, netstat (TCP network statistics), IP address, proxy information, domain name, number of packet and bytes send and received on each wireless interfaces, DNS's IP address, routing table information	application name on the user screen, number of user touches on the screen and duration during an usage session	WiFi level, WiFi BSSID, WiFi SSID, WiFi interface speed, cell ID, cell operator, cell strength, cell Radio Access Technology, cell network code, cell bandwidth up and down stream, battery state(e.g. charging, full, discharging), battery level, battery temperature	GPS position, user physical activity from the Google Play Services Activity(still, tilting: between two states,in vehicle, on bicycle,on foot, running)
[29]	flow start times, flow direction (up/down), flow duration, flow size, avg. flow throughput	app	signal strength, operator, cell ID, cell location, RAT, Device ID (IMEI)	GPS position, location (walking, home etc)
[30]	flow start times, flow direction (up/down), flow duration, flow size, avg. flow throughput	app	signal strength, operator, cell ID, cell location, RAT, Device ID (IMEI)	GPS position, location (walking, home etc)
[31]	flow start times, flow direction (up/down), flow duration, flow size, avg. flow throughput, total number of frames that would have been displayed if no frames are dropped, total number of frames dropped predecode or dropped because the frame missed its display deadline, total number of corrupted frames that have been detected	app, height and width of the video's display area, all the ranges of the video that have been played, current video playtime, YouTube identifier of the video content, session timestamp	signal strength, operator, cell ID, cell location,RAT, Device ID (IMEI)	-
[37]	network throughput	-	-	-

you trying to accomplish? (Consume content, share or create content, read text message, write text message, control an app (start/stop music), video call or audio call) 4. Did your last usage of this app meet your expectations (on a scale 1-5)? 5. If something went wrong, please tell us more about it. In all five studies ([24], [25], [26], [27], [28]) the survey was triggered after using an app, but maximum 12 times during a day.

TABLE 3. Objective data collected in *watching videos* studies.

paper	network	activity	cell	context
[38]	transport protocol (RTS or progressive download), packet-loss rate for the audio and video track, the mean and maximum jitter (i.e., the variability over time of the packet latency across the network) for audio and video video resolution, early interruption of the video (e.g., due to network disconnection or a disinterest of the user)	metadata about the video (id, title, length) and the start and end of the session (timestamp)	network type (e.g., UMTS, HSDPA, GPRS), number of handovers (i.e., all kinds of radio cell reselections) and inter-system handovers (i.e., different data connection-type cell reselections e.g., between UMTS and HSDPA), RSSI (received signal strength indicator)	GPS position, location (walking, home etc)
[39]	video jitter, audio jitter, video packet loss rate, audio packet loss rate, transport protocol, eebuffering time, loading time	percentage watched, video quality	GPRS percentage, GPRS average RSSI [dBm], EDGE percentage [%] EDGE average RSSI [dBm], 3G & B3G percentage [%] , 3G & B3G average RSSI [dBm], WiFi percentage, WiFi average RSSI [dBm] num. of inter-system handovers, num. of handovers	GPS position, location (walking, home etc)
[40]	the video packet-loss rate (i.e. the percentage of video packets lost during transmission to the device) and the loading time of the video (which is measured as the time period between selecting a video and the moment when the video starts playing)	-	-	-
[41]	transmission protocol (RTP or progressive download), packet-loss rate for the audio and video track, the mean and maximum jitter (i.e., the variability over time of the packet latency across the network) for audio and video early interruption of the video (e.g., due to network disconnection)	metadata about the video (id, title, length),the start and end of the session (time-stamp), video quality (resolution and bitrate)	network type(e.g., UMTS, HSDPA, GPRS), number of handovers (i.e., all kinds of radio cell reselections), and inter-system handovers(i.e., different data connection-type cell reselections e.g., be-tween UMTS and HSDPA), and RSSI (received signal strength indicator)	GPS position, location (walking, home etc)

TABLE 4. Types of subjective data collected in the studies.

Category	Papers
overall quality	[21]–[23], [25], [26], [29]–[37], [39], [41]
acceptability	[22], [23], [29]–[32], [39], [41]
context	[20], [25], [26], [32], [38], [39], [41]
multi-episodic quality	[33]–[36]
other	[20], [22]–[29], [38]–[41]

In the paper describing the public display study ([20]), the participants were asked to daily submit their reports using an online form. The diary report consisted of several closed and open questions, asking about the context (how many people were at the display and what did they do, how long and how many times were the testers there, what did they see etc.). Additionally, since the diary report form did not specifically ask for or mention the notifications and the result visualizations, a focus group session was conducted to directly highlight these elements. Then, they were asked about elements that grabbed users' attention and/or engaged to interact; progression of these over time; perception/experience of the pop-up notifications; users' observations of others' behavior around the display; further envisioned dynamic features of the display.

2) SUBJECTIVE DATA FOR AUDIO-VISUAL CALLS STUDIES

For the *audio-video calls* experiments, after each call, participants rated the overall quality, the audio quality and the video quality of that particular call. For this purpose, 7-point continuous rating scale was used, to avoid saturation effects at the scale extremities and frequently shows a more Gaussian distribution of the judgments [36]. In paper [33], after the 4th, 14th and 24th call (i.e. after 2, 7 and 12 days), the multi-episodic judgment for the same dimensions needed to be rated. In paper [36] after the 4th, 14th and 24th call (i.e. after 2, 7 and 12 days), an extended questionnaire was provided. In addition to the mentioned ratings related to the individual call, this questionnaire also contained questions regarding the (overall, audio and video) quality of the service experienced so far, as well as any expected future service usage and recommendation to friends.

3) SUBJECTIVE DATA FOR WATCHING VIDEOS STUDIES

In this category of studies, different approaches were used. Either the testers were only asked about overall quality of their experience (in most cases both right after watching a video and after several days to measure the multi-episodic quality). In the other studies the questions to testers were more specific, for example accessing loading speed or social context. Mostly, the study using more specific questions implemented traditional diaries, where the testers could write down more information about their experience. More details are presented below.

In the study described in paper [33], apart from rating the quality of each episodic use, participants also rated

TABLE 5. Summary of the most commonly mentioned factors influencing QoE.

Factor	Yes	Uncertain	No
throughput	[25], [29]–[32], [37]	-	-
packet loss	[38]–[41]	-	-
other QoS factors	[21], [22], [25], [27], [28]	[28], [29]	[21], [22], [28], [30]
loading	[40], [41]	-	-
stalling	[29], [31]	-	-
pixel quality	[29]	-	[31]
user behavior/state and phone	[28]	[25]–[27]	[41]

multi-episodic quality after every 3rd episodic use. Similar methods were used in the paper [34], where two questionnaires were used: one per usage presented directly after finishing watching 15-minutes long movie and one to measure the integrated QoE over several interactions. The second questionnaire was presented after day 4, 7, 10 and 14 to measure the integrated QoE and determine the users satisfaction with each system. In study [38] the testers were asked to answer a few questions (on a 5-point star rating scale) to evaluate the audiovisual aspects and the QoE: the loading speed, fluentness, general experience, and noticeable distortion in the watched video. They also asked a question about the physical context of the test user (''Are you at home, on the move, at work, or somewhere else?''). Additionally, extra feedback was collected in the user diaries. Very similarly, in the study described in [39] and [41] the questionnaire included questions about the appreciation of the content, general technical quality, fluentness of the image, loading speed and the physical context of the user. The diaries contained 28 sheets (one for every video) with a number of questions: about the interest in the content, general experience evaluation, positive and negative aspects with an influence upon this overall experience, the social context in which the user watched. Finally, users were also asked to evaluate the acceptability of the audiovisual quality of the videos using the following categories: ''Acceptable in every context'' (accept), ''Acceptable but only in the context that I watched the video'' (context) and ''Not acceptable'' (not accept). The study described in [40] was assessing only perceived distorting and the loading speed, also using a 5-level subjective quality scale, and used a diary as well, to give a possibility for additional feedback. Paper [35] described 2 different experiments, in both of them the participants assessed the episodic perceived quality on the 7-point scale after finishing an episode. In the first one, the multi-episodic perceived quality for each service was assessed on the 4th, 7th, 10th, and 14th day directly after finishing the daily episode with the VoD service (same scale). In the second one, multi-episodic judgments were taken at the 2nd, 5th, 8th, 11th, and 14th day.

D. RESULTS PRESENTED IN THE PAPERS

As mentioned before, the factors that especially can be observed with longitudinal approach are Context and System factors. However, according to our analysis, mainly System influenced factors (network, device and media related) and Context influenced factors (physical and social) were analyzed in the studies described in this paper. When it comes to Human factors, most of the researchers were only mentioning the testers' age and gender. In many studies, the participants were students or working at the university. Sometimes also socio-economic status was checked (occupation, education level, family status [26]), and their experience in a field of interest (how long were they using mobile phones [25], [26] or are they interested in presented content [38], [39], [41]). For each category of papers we present tables (6, 7, 8) with goals of the studies and type of technical and subjective data described in results.

The analysis of the results is quite challenging. Especially for exploratory studies, the conclusions are often more observations than findings supported by statistically significant results. Table 5 sums up the most commonly mentioned factors influencing QoE. The column ''Yes'' is for papers, which results indicate that this factor is significantly influencing QoE. The column called ''Uncertain'' is for papers, which results do not show the correlation precisely, and the column called ''No'' is for papers claiming that this factor has no influence.

Besides research on influencing factors, some publications compared a lab study and long term study. The obtained results are ambiguous, since the papers [30] and [31] show that the results from laboratory are similar to the ones obtained with living lab. However, papers [33], [34], [35], [36] claim the opposite.

Another research direction is time influence. Surprisingly, taking into account that all papers describe a long term study, not many of the papers are mentioning how the users' behaviour was changing in time. The two papers investigating time influence are [33], [35] described below in more details.

In the study [35], the largest effect on multi-episodic judgments was observed for an increasing number of degraded episodes. The final multi-episodic judgment decreased until two consecutive degraded episodes/days were presented. Then, no further decrease was observed although the multi-judgment remained well above the episodic judgments of degraded episodes, i. e., a saturation effect was observed.

The study from paper [33] stresses that approaches to model multi-episodic quality in a period of approximately 2 weeks cannot be directly deduced from episodic-models. Short-term effects like the recency effect and the peak-rule do not improve the prediction accuracy. Further research is

TABLE 6. Overview of technical results and subjective measures in *using an app* studies.

Paper	Goal of the study	Technical results	Subjective measures
[20]	answering the question: How does the awareness of peer interactions on the display system affect user engagement?	interactions with quizzes and videos that led to results/ratings in the database; page views of the mobile quiz/ratings web apps	user engagement, context
[21]	predicting QoE relevant matrics, users QoE and user engagement	initial delay, number of stalling events, total stalling time, stalling ratio, quality switches, re-buffering ratio	MOS, user engagement (fraction of the total video length a user watched)
[22]	analyzing a dataset of crowdsourced measurements on YouTube QoE, passively collected in real users' smartphones	Network parameters: average throghput, max. throughput, duration, volume, signal strength; Streaming parameters: initial delay, total stalling time, avr. stalling event length, max. stalling event length, number of stalling events, number of quality changes, start quality, end quality, recency time, weighted time on layer, playback time, resolution	MOS, user engagement (fraction of the total video length a user watched)
[23]	observing a systematic performance and QoE improvement of YouTube in mobile devices over time, accompanied by optimization of the YouTube streaming behavior for smartphones	initial delay, number of stalling events, total stalling time, stalling ratio, quality switches, re-buffering ratio	MOS, user engagement (fraction of the total video length a user watched)
[24]	testing mQoL Lab on a small scale study	time spent on each app	use of selected apps, context
[25]	improving understanding of factors influencing QoE, and deriving implications for mobile application design and QoE management	Mean received throughput, choice of wireless access technology, WiFi strength, SRT, RTT	context (location), QoE rating, most commonly mentioned keywords in surveys and interviews, phone charging behaviours
[26]	improving understanding of factors influencing QoE, and deriving implications for mobile application design and QoE management	choice of wireless access technology	QoE rating, context (location), social context, habbits, previous experience, phone charging behaviours
[27]	predicting the QoE of smartphone application usage based on the data collected from the smartphone and the QoE ratings from the participant	network, activity, cell and context parameters	QoE rating
[28]	building an accurate QoE model from a dataset obtained in-the-wild, predicting the "High" or "Low" QoE of smartphone application usage based on the on-board data collected from the smartphones labeled by the participants' QoE ratings	network, activity, cell and context parameters	QoE rating, context (location), expectations
[29]	presenting the characteristics of current mobile YouTube streaming	Network parameters: average throghput, max. throughput, duration, volume, signal strength; Streaming parameters: initial delay, total stalling time, avr. stalling event length, max. stalling event length, number of stalling events, number of quality changes, start quality, end quality, recency time, weighted time on layer, playback time, resolution	MOS, acceptability
[30]	dissussing on end-device passive measurement and analysis, providing sound basis to better understand QoE requirements of popular mobile apps, as well as for monitoring the underlying provisioning network	used app, maximum flow throughput	QoE ratings, context (location)
[31]	studying the QoE of popular apps in smartphones (YouTube, Facebook, Gmaps), considering the impact of the most relevant QoS-based characteristics of the access network: the downlink bandwidth	Network parameters: average throghput, max. throughput, duration, volume, signal strength; Streaming parameters: initial delay, total stalling time, avr. stalling event length, max. stalling event length, number of stalling events, number of quality changes, start quality, end quality, recency time, weighted time on layer, playback time, resolution	MOS, acceptability
[32]	estimating the QoE of popular end-customer services in both cellular and fixed-line networks	YouTube flow ADT values and the distribution of the β= ADT/VBR, number of stallings, downlink bandwidth	acceptance rate, MOS
[37]	presenting a novel measurement framework to unify network measurement and QoE assessments	NDT download throughput	QoE rating

required to understand how quality evolves over meaningful time periods from the user's perspective that cannot be deduced from short stimuli in the range of seconds. That is why more research on the long term video quality is needed.

V. DISCUSSION

The results analysis shows that the longitudinal studies are not very popular in the QoE field, more precisely in the studies investigating video quality. Furthermore, even in case of relatively long-term studies (2 or 4 weeks), changes in the subjective quality judgements over time are not investigated. This constitutes a challenge for future work, especially when designing experiments that aim to take longer than a month.

We found that in most of the studies, especially those from *Using an app studies* category, objective data have been collected. However, the results are rarely presented, but rather applied to propose machine learning models. Unfortunately, the models are not described in detail, so it is difficult to grasp

TABLE 7. Overview of technical results and subjective measures in *watching videos* studies.

Paper	Goal of the study	Technical results	Subjective measures
[33]	modeling multi-episodic quality	bitrate	episodic and multi-episodic quality ratings
[34]	investigating temporal effects on QoE over several interactions (macro-temporal variations of performance, i.e. variation only between interactions, and not micro-temporal variation, i.e. within-interaction)	Video encoding bandwidth	QoE per interaction (video, audio, overall), integrated QoE
[35]	understanding the impact of potential factors (i. e., varying performance, usage situation, and usage period) affecting the quality formation process; implementing a model using episodic judgments for the prediction of multi-episodic judgments in terms of MOS	Compression: (for HP, content downscaled to a resolution of 1280x720px and encoded with h.264 (25FPS,5Mbit/s, two-pass). For LP, the video signal is degraded by setting the QP factor to 50)	MOS
[38]	quantifying the QoE of mobile video consumption in a real-life setting based user behavior and technical parameters such as network and video quality	transport protocol, quality of the video source, transport protocol, the quality of the video source, the types of data network that were used to transmit the video, the number of handovers during transmission, and the percentage of the video that was actually watched by the user	QoE rating
[39]	observing the users' behaviors in a natural context, to gain insight in the subjective evaluation of the offered video qualities and to determine the acceptability of the audiovisual quality based on objectively measured parameters	transport protocol and the quality of the video source, for modelling: the transport protocol, the quality of the video source, the type of data network that was used, and the percentage of the video that was actually watched by the user	acceptability
[40]	discussing results froman exploratory study in which QoE aspects related to mobile video-on-demand were investigated in a living lab setting	mean loading time, video packet loss rate	context (location), context (social), evaluation of loading time, evaluation of distortion
[41]	exploring contextual aspects and subjective quality evaluations related to mobile video watching in a natural environment	video quality (resolution and bitrate), transmission protocol (RTP or progressive download) mean loading time, video packet loss rate	context (location), context (social), evaluation of loading time, evaluation of distortion, evaluation of fluentness, technical quality rating, user feedback (positive/negative comments, change reports)

TABLE 8. Overview of technical results and subjective measures in *audio-visual calls* studies.

Paper	Goal of the study	Technical results	Subjective measures
[33]	modeling multi-episodic quality	bitrate	episodic and multi-episodic quality ratings
[36]	showing how the long-term rating of a service quality is impacted by variations on transmission quality	bitrate	overall MOS, audio MOS and videoMOS, service quality after N calls

which factors are important. Consequently, a formal meta-analysis combining results from different papers cannot be performed. The most commonly investigated aspects of QoE are throughput and packet loss [25], [29], [30], [31], [32], [37], [38], [39], [40], [41]. Loading and stalling times are also often measured [29], [31], [40], [41].

Using an app studies category includes most ecologically valid studies described in this review. Unlike in a typical QoE study, testers are not forced to watch content they are not interested in. Similarly, they are also not obliged to use an application if they would not use it in their daily routine [24], [25], [26], [27], [28]. Taking ecological validity as a priority, future studies should adjust their methodology. Given that all the reviewed studies probed quality, one can propose a study where participants report whenever they are annoyed by any service-related issues and only then researchers would check, if the issues are related to the quality. We can also test whether participants would perform any actions to increase the quality (e.g., by reconnecting WiFi, switching from WiFi to the cellular connection, refreshing the website or the application). We could also investigate long-term effects of being annoyed with the service, e.g., ask participants to report when the quality over time drop to the level at which they would stop using a service and look for some alternative. This further justifies the long-term component of the study, because such customer decisions take time.

While ecologically valid, *using an app* studies have some limitations. Measuring QoE in an uncontrolled environment is challenging, as there are multiple factors to be taken into account. Users who rarely use certain applications on their own devices (e.g., only at home) can give biased responses. Another issue is that replication of those studies, or designing new ones becomes challenging. This is because the safety and privacy settings on most of modern cellphones do not allow to track some of technical data. Users also become more aware of their data privacy, and are not willing to participate in an

experiment if sensitive personal data, such as GPS location, is tracked. Consequently, future studies of high ecological validity might not be possible. For that reason future work will likely focus on *audio-visual calls* and *watching video* studies.

To increase the ecological validity of *audio-visual calls*, researchers may propose more scenarios for testers to choose from. They can also make the scenarios more realistic (i.e., conversation may be more flexible in terms of the topic). Future work within *watching videos* studies category may include a short preliminary questionnaire asking testers about their interest and habits so to make experimental conditions of the study more ecologically valid. This could be done e.g., by using users favorite movie genre as content in the study or by adjusting the length of the videos to typical time of materials they watch on their phones.

The subjective judgement in most of the studies was collected in two ways. Either questions pop up after watching a video or a diary was administered. It is difficult to indicate the best practice here. The diary may give more detailed information and allow testers to respond with more details with open questions. However, especially in case of the longitudinal studies, it can become exhausting for participants so their motivation to properly fill the diaries might drop. Perhaps the solution for studies taking more than 4 weeks is to apply pop-up questions on daily basis and on top of that ask users to fill out a diary from time to time. Finally, the solution may be to offer some additional remuneration or other type of incentive.

We recognise two main areas for improvement in longitudinal QoE studies. First, the studies are typically focused on System factors influencing QoE. Second, if Context factors are described, the studies lack systematization and statistical analysis (e.g., the studies mention that users' routines in the quality rating changes over time, but how exactly is not reported [25], [26].

When comparing results from lab studies and living lab studies, the conclusions are ambiguous - some studies claim that the results are similar [30], [31], and some claim otherwise [33], [34], [35], [36]. This discrepancy suggests an important direction of future work - the differences and similarities between results obtained with the living lab and laboratory approaches should be further investigated. Such a comparative approach allows us to understand better which type of a study design is more suitable to investigate a given type of factors influencing QoE.

VI. CONCLUSION

The goal of this paper was to review the literature discussing the longitudinal studies investigating video QoE assessment. Our review shows that long term studies are not common in the QoE field. If the longitudinal study is performed, it seems to aim at gathering more data rather than observing changes in users' behavior over time. In our opinion, this constitutes a significant gap in QoE research. This is because in real-life scenarios, the majority of user interactions with services are not one-time or temporary, but rather long-lasting and dynamic. Thus, it seems beneficial to broaden the scope of the research in this direction. To better understand the QoE and to propose more reliable models, one should investigate how the QoE assessment changes over the time and how it differs from the results obtained from a single experimental session.

Importantly, most of the QoE studies are focused on the influence of network parameters on the quality ratings. Human factors are still not sufficiently investigated. Thus, one should apply qualitative methods (such as diaries and extended interviews), rather than simple QoE ratings to understand better how users perceive the quality and which factors cause changes in their behavior. Even though qualitative data analysis is more time-consuming, it proves its importance in predicting users' experience with a service.

Another conclusion we derived from this review is that the terminology and standards of reporting studies procedures and results in the QoE field should be more standardized. As we show above, the terminology referring to the group of experiments using the very same methodology is very broad ("over time", "long term", "longitudinal" or "multi-episodic"), which makes it difficult to identify all relevant papers. The same applies to the presentation of the results. The lack of standardization leads to the situation, where in some papers no statistical effects have been presented. Without this kind of information, it seems impossible to properly identify the factors influencing the QoE.

Finally, in the discussion section, we present our conclusions and suggestions on how to improve the ecological validity of the longitudinal studies in the QoE field.

APPENDIX A
LIST OF THE QUERIES USED TO GENERATE PAPERS FOR THE REVIEW

- Used in ACM: [[Keywords: "quality of experience"] OR [Keywords: "perceived quality"] OR [Keywords: "user experience"] OR [Keywords: "user engagement"] OR [Keywords: "qoe"] OR [Keywords: "quality of life"] OR [Keywords: "mobile human"]] AND [[Keywords: "longitudinal"] OR [Keywords: "macro-temporal"] OR [Keywords: "multi-episodic"] OR [Keywords: "over time"] OR [Keywords: "prolonged use"] OR [Keywords: "long term"] OR [Keywords: "living lab"] OR [Keywords: "live lab"]] AND NOT [[All: "patient*"] OR [All: "treatment*"] OR [All: "medic*"] OR [All: "hospital*"] OR [All: "health"] OR [All: "cancer"] OR [All: "care"] OR [All: "diagnos*"] OR [All: "surgery"] OR [All: "pain"] OR [All: "disease"] OR [All: "disorder"]]
- Used in IEEE: ("Author Keywords":"Quality of Experience" OR "Author Keywords": "perceived quality" OR "Author Keywords": "User Experience" OR "Author Keywords": "User engagement" OR "Author Keywords": "QoE" OR "Author Keywords": "Quality of life" OR "Author Keywords": "Mobile human") AND ("Author Keywords":"longitudinal" OR "Author

Keywords'': ''macro-temporal'' OR ''Author Keywords'': ''multi-episodic'' OR ''Author Keywords'': ''over time'' OR ''Author Keywords'': ''prolonged use'' OR ''Author Keywords'': ''long term'' OR ''Author Keywords'': ''living lab'' OR ''Author Keywords'': ''live lab'') NOT (''All Metadata'':''patient*'' OR ''All Metadata'': ''treatment*'' OR ''All Metadata'': ''medic*'' OR ''All Metadata'': ''hospital'' OR ''All Metadata'': ''health'' OR ''All Metadata'': ''cancer'' OR ''All Metadata'': ''care'' OR ''All Metadata'': ''diagnos*'' OR ''All Metadata'': ''surgery'' OR ''All Metadata'': ''pain'' OR ''All Metadata'': ''disease'' OR ''All Metadata'': ''disorder'')

- Used in IEEE: (TITLE (''Quality of Experience'' OR ''perceived quality'' OR ''User Experience'' OR ''User engagement'' OR ''QoE'' OR ''Quality of life'' OR ''Mobile human'') AND TITLE (''longitudinal'' OR ''macro-temporal'' OR ''multi-episodic'' OR ''over time'' OR ''prolonged use'' OR ''long term'' OR ''living lab'' OR ''live lab'') AND NOT ALL (''patient*'' OR ''treatment*'' OR ''medic*'' OR ''hospital'' OR ''health'' OR ''cancer'' OR ''care'' OR ''diagnos*'' OR ''surgery'' OR ''pain'' OR ''disease'' OR ''disorder''))
- Used in IEEE: (''Author Keywords'':''Quality of Experience'' OR ''Author Keywords'': ''perceived quality'' OR ''Author Keywords'': ''User Experience'' OR ''Author Keywords'': ''User engagement'' OR ''Author Keywords'': ''QoE'' OR ''Author Keywords'': ''Mobile human'') AND (''Author Keywords'':''longitudinal'' OR ''Author Keywords'': ''macro-temporal'' OR ''Author Keywords'': ''multi-episodic'' OR ''Author Keywords'': ''over time'' OR ''Author Keywords'': ''prolonged use'' OR ''Author Keywords'': ''long term'' OR ''Author Keywords'': ''living lab'' OR ''Author Keywords'': ''live lab'') AND (''Author Keywords'':''video'' OR ''Author Keywords'': ''multimedia'' OR ''Author Keywords'': ''subjective'')
- Used in Scopus: (KEY (''Quality of Experience'' OR ''perceived quality'' OR ''User Experience'' OR ''User engagement'' OR ''QoE'' OR ''Mobile human'') AND KEY (''longitudinal'' OR ''macro-temporal'' OR ''multi-episodic'' OR ''over time'' OR ''prolonged use'' OR ''long term'' OR ''living lab'' OR ''live lab'') AND KEY (''video'' OR ''multimedia'' OR ''subjective''))
- Used in Scopus: (''Abstract'':''Quality of Experience'' OR ''Abstract'': ''perceived quality'' OR ''Abstract'': ''User Experience'' OR ''Abstract'': ''User engagement'' OR ''Abstract'': ''QoE'' OR ''Abstract'': ''Mobile human'') AND (''Abstract'':''longitudinal'' OR ''Abstract'': ''macro-temporal'' OR ''Abstract'': ''multi-episodic'' OR ''Abstract'': ''over time'' OR ''Abstract'': ''prolonged use'' OR ''Abstract'': ''long term'' OR ''Abstract'': ''living lab'' OR ''Abstract'': ''live lab'') AND (''Abstract'':''subjective'')
- Used in Scopus: (KEY (''Quality of Experience'' OR ''perceived quality'' OR ''User Experience'' OR ''User engagement'' OR ''QoE'' OR ''Mobile human'') AND ABS (''subjective'') AND TITLE (''longitudinal'' OR ''macro-temporal'' OR ''multi-episodic'' OR ''over time'' OR ''prolonged use'' OR ''long term'' OR ''living lab'' OR ''live lab''))
- Used in Scopus: (ABS (''Quality of Experience'' OR ''perceived quality'' OR ''User Experience'' OR ''User engagement'' OR ''QoE'' OR ''Mobile human'') AND ABS (''subjective'') AND ABS (''longitudinal'' OR ''macro-temporal'' OR ''multi-episodic'' OR ''over time'' OR ''prolonged use'' OR ''long term'' OR ''living lab'' OR ''live lab''))

REFERENCES

[1] D. Guse, A. Wunderlich, B. S. Weiss, and Möller, ''Duration neglect in multi-episodic perceived quality,'' in *Proc. 12th Int. Conf. Quality Multimedia Exper. (QoMEX)*, Jun. 2016, pp. 1–3.

[2] D. Guse, O. Hohlfeld, A. Wunderlich, B. S. Weiss, and Möller, ''Multi-episodic perceived quality of an audio-on-demand service,'' in *Proc. 12th Int. Conf. Quality Multimedia Exper. (QoMEX)*, May 2020, pp. 1–6.

[3] V. Manea, A. Berrocal, A. De Masi, N. H. Møller, K. Wac, H. Bayer, S. Lehmann, and E. Ashley, ''LDC 19: International workshop on longitudinal data collection in human subject studies,'' in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput., ACM Int. Symp. Wearable Comput.*, Sep. 2019, pp. 878–881, doi: 10.1145/3341162.3347758.

[4] S. Möller, C. Bang, T. Tamme, M. Vaalgamaa, and B. Weiss, ''From single-call to multi-call quality: A study on long-term quality integration in audio-visual speech communication,'' in *Proc. Interspeech*, Aug. 2011, pp. 1–4.

[5] J. Lee, J. Lee, and L. Feick, ''The impact of switching costs on the customer satisfaction-loyalty link: Mobile phone service in France,'' *J. Services Marketing*, vol. 15, no. 1, pp. 35–48, Feb. 2001.

[6] T. N. Duc, C. M. Tran, P. X. Tan, and E. Kamioka, ''Modeling of cumulative QoE in on-demand video services: Role of memory effect and degree of interest,'' *Future Internet*, vol. 11, no. 8, p. 171, Aug. 2019.

[7] S. Kujala, V. Roto, K. V. Vainio-Mattila, and E. A. A. Karapanos, ''UX Curve: Method for evaluating long-term user experience, *Interacting Comput.*, vol. 23, pp. 473–483, Sep. 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0953543811000737

[8] S. Egger, P. Reichl, T. Hosfeld, and R. Schatz, '''Time is bandwidth'? Narrowing the gap between subjective time perception and quality of experience,'' in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2012, pp. 1325–1330.

[9] T. Hoßfeld, S. Biedermann, R. Schatz, A. Platzer, S. Egger, and M. Fiedler, ''The memory effect and its implications on web QoE modeling,'' in *Proc. 23rd Int. Teletraffic Congr. (ITC)*, 2011, pp. 103–110.

[10] *Connected Papers Tool*. Accessed: Jun. 20, 2021. [Online]. Available: https://www.connectedpapers.com/

[11] R. McElreath. *The Evolution of Statistical Methods for Studying Human Evolution*. YouTube. Accessed: Jun. 10, 2022. [Online]. Available: https://www.youtube.com/watch?v=Wu0hAjlMqUQ&abchannel=MizzouVisualProductions

[12] M. Fiedler, S. Möller, P. Reichl, and M. Xie, ''QoE Vadis? (Dagstuhl perspectives workshop 16472),'' *Dagstuhl Manifestos*, vol. 7, no. 1, pp. 30–51, 2018. [Online]. Available: http://drops.dagstuhl.de/opus/volltexte/2018/8683

[13] B. Kitchenham, ''Procedures for performing systematic reviews,'' *Keele, U.K., Keele Univ.*, vol. 33, pp. 1–26, Jul. 2004.

[14] (2022). *Wikipedia Replication Crisis*. [Online]. Available: https://en.wikipedia.org/wiki/Replication_crisis

[15] E. Aromataris and A. Pearson, ''The systematic review: An overview,'' *AJN Amer. J. Nursing*, vol. 114, no. 3, pp. 53–58, 2014.

[16] B. T. Truong and S. Venkatesh, ''Video abstraction: A systematic review and classification,'' *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 3, no. 1, p. 3, Feb. 2007, doi: 10.1145/1198302.1198305.

[17] M. Pallot and P. Krawczyk, ''Landscaping user centered related methods applied in the context of living labs,'' in *Proc. IEEE Int. Conf. Eng., Technol. Innov./Int. Technol. Manag. Conf.*, Jun. 2015, pp. 1–8.

[18] I. D. L. T. Díez, S. G. Alonso, S. Hamrioui, M. López-Coronado, and E. M. Cruz, ''Systematic review about QoS and QoE in telemedicine and eHealth services and applications,'' *J. Med. Syst.*, vol. 42, no. 10, pp. 1–10, Oct. 2018.

[19] E. Bañuelos-Lozoya, G. González-Serna, N. González-Franco, O. Fragoso-Diaz, and N. Castro-Sánchez, "A systematic review for cognitive state-based QoE/UX evaluation," *Sensors*, vol. 21, no. 10, p. 3439, May 2021.
[20] M. Müller, N. Otero, A. Alissandrakis, and M. Milrad, "Increasing user engagement with distributed public displays through the awareness of peer interactions," in *Proc. 4th Int. Symp. Pervasive Displays*, Jun. 2015, pp. 23–29.
[21] S. Wassermann, N. Wehner, and P. Casas, "Machine learning models for YouTube QoE and user engagement prediction in smartphones," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 46, no. 3, pp. 155–158, Jan. 2019, doi: 10.1145/3308897.3308962.
[22] S. Wassermann, P. Casas, M. Seufert, and F. Wamser, "On the analysis of YouTube QoE in cellular networks through in-smartphone measurements," in *Proc. 12th IFIP Wireless Mobile Netw. Conf. (WMNC)*, Sep. 2019, pp. 71–78.
[23] N. Wehner, S. Wassermann, P. Casas, M. Seufert, and F. Wamser, "Beauty is in the eye of the smartphone holder a data driven analysis of YouTube mobile QoE," in *Proc. 14th Int. Conf. Netw. Service Manag. (CNSM)*, 2018, pp. 343–347.
[24] A. De Masi and K. Wac, "You're using this app for what: A mQoL living lab study," in *Proc. ACM Int. Joint Conf. Int. Symp. Pervasive Ubiquitous Comput. Wearable Comput.*, Oct. 2018, pp. 612–617.
[25] S. Ickin, K. Wac, M. Fiedler, L. Janowski, J.-H. Hong, and A. K. Dey, "Factors influencing quality of experience of commonly used mobile applications," *IEEE Commun. Mag.*, vol. 50, no. 4, pp. 48–56, Apr. 2012.
[26] K. Wac, S. Ickin, J. H. Hong, L. Janowski, M. Fiedler, and A. K. Dey, "Studying the experience of mobile applications used in different contexts of daily life," in *Proc. 1st ACM SIGCOMM Workshop Meas. Stack*, 2011, pp. 7–12.
[27] A. D. Masi and K. Wac, "Predicting quality of experience of popular mobile applications from a living lab study," in *Proc. 11th Int. Conf. Quality Multimedia Exper. (QoMEX)*, Jun. 2019, pp. 1–6.
[28] A. De Masi and K. Wac, "Towards accurate models for predicting smartphone applications QoE with data from a living lab study," *Qual. User Exper.*, vol. 5, no. 1, pp. 1–18, Dec. 2020.
[29] M. Seufert, P. Casas, F. Wamser, N. Wehner, R. Schatz, and P. Tran-Gia, "Application-layer monitoring of QoE parameters for mobile YouTube video streaming in the field," in *Proc. IEEE 6th Int. Conf. Commun. Electron. (ICCE)*, Jul. 2016, pp. 411–416.
[30] P. Casas, B. Gardlo, M. Seufert, F. Wamser, and R. Schatz, "Taming QoE in cellular networks: From subjective lab studies to measurements in the field," in *Proc. 11th Int. Conf. Netw. Service Manag. (CNSM)*, Nov. 2015, pp. 237–245.
[31] P. Casas, M. Seufert, F. Wamser, B. Gardlo, A. Sackl, and R. Schatz, "Next to you: Monitoring quality of experience in cellular networks from the end-devices," *IEEE Trans. Netw. Service Manag.*, vol. 13, no. 2, pp. 181–196, Jun. 2016.
[32] P. Casas, B. Gardlo, R. Schatz, and M. Mellia, "An educated guess on QoE in operational networks through large-scale measurements," in *Proc. ACM SIGCOMM Workshop QoE Based Anal. Manag. Data Commun. Netw.*, 2016, pp. 1–6.
[33] D. Guse, B. Weiss, and S. Moller, "Modelling multi-episodic quality perception for different telecommunication services: First insights," in *Proc. 6th Int. Workshop Quality Multimedia Exper. (QoMEX)*, Sep. 2014, pp. 105–110.
[34] D. Guse and S. Möller, "Macro-temporal development of QoE: Impact of varying performance on QoE over multiple interactions," in *Proc. DAGA* Mar. 2013, pp. 452–455.
[35] D. Guse. (2016). *Multi-Episodic Perceived Quality of Telecommunication Services Spatial Telephone Conferencing for Asterisk (STEAK) View Project*. [Online]. Available: https://www.researchgate.net/publication/308645878
[36] S. Möller, C. Bang, T. Tamme, M. Vaalgamaa, and B. Weiss, "From single-call to multi-call quality: A study on long-term quality integration in audio-visual speech communication," in *Proc. Interspeech*, Aug. 2011, pp. 1485–1488.
[37] R. K. P. Mok, G. Kawaguti, and K. Claffy, "QUINCE: A unified crowdsourcing-based QoE measurement platform," in *Proc. ACM SIGCOMM Conf. Posters Demos*, 2019, pp. 60–62, doi: 10.1145/3342280.3342307.
[38] T. De Pessemier, K. De Moor, A. Juan, W. Joseph, L. De Marez, and L. Martens, "Quantifying QoE of mobile video consumption in a real-life setting drawing on objective and subjective parameters," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, Jun. 2011, pp. 1–6.
[39] T. De Pessemier, K. De Moor, A. J. Verdejo, D. Van Deursen, W. Joseph, L. De Marez, L. Martens, and R. Van De Walle, "Exploring the acceptability of the audiovisual quality for a mobile video session based on objectively measured parameters," in *Proc. 3rd Int. Workshop Quality Multimedia Exper.*, Sep. 2011, pp. 125–130.
[40] T. De Pessemier, L. Martens, and W. Joseph, "Modeling subjective quality evaluations for mobile video watching in a living lab context," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, Jun. 2013, pp. 1–5.
[41] T. De Pessemier, K. De Moor, W. Joseph, L. De Marez, and L. Martens, "Quantifying subjective quality evaluations for mobile video watching in a semi-living lab context," *IEEE Trans. Broadcast.*, vol. 58, no. 4, pp. 580–589, Dec. 2012.
[42] F. Wamser, M. Seufert, P. Casas, R. Irmer, P. Tran-Gia, and R. Schatz, "YoMoApp: A tool for analyzing QoE of YouTube HTTP adaptive streaming in mobile networks," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Jun. 2015, pp. 239–243.
[43] U. Heute, "Telephone-speech quality," in *Speech and Audio Processing in Adverse Environments*. Berlin, Germany: Springer, 2008, pp. 287–337, doi: 10.1007/978-3-540-70602-1_9.
[44] E. Karapanos, J. Martens, and M. Hassenzahl, "On the retrospective assessment of users' experiences over time: Memory or actuality?" in *Proc. 28th Int. Conf. Hum. Factors Comput. Syst. (CHI)*, 2010, pp. 4075–4080, doi: 10.1145/1753846.1754105.
[45] R. Al, "Factors influencing quality of experience," in *Quality of Experience* (T-Labs Series in Telecommunication Services). Cham, Switzerland: Springer, 2014, pp. 55–72, doi: 10.1007/978-3-319-02681-7_4.
[46] J. Prumper, D. Zapf, F. C. Brodbeck, and M. Frese, "Some surprising differences between novice and expert errors in computerized office work," *Behaviour Inf. Technol.*, vol. 11, no. 6, pp. 319–328, Nov. 1992.
[47] A. Følstad, "Living labs for innovation and development of information and communication technology: A literature review," *Electron. J. Organizational Virtualness*, vol. 10, pp. 99–131, Jan. 2008.
[48] O. Collaboration, "Estimating the reproducibility of psychological science," *Science*, vol. 349, no. 6251, 2015, Art. no. aac4716. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.aac4716
[49] Z.-Y. Chen, Z.-P. Fan, and M. Sun, "A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data," *Eur. J. Oper. Res.*, vol. 223, no. 2, pp. 461–472, Dec. 2012.
[50] D. A. Redelmeier and D. Kahneman, "Patients memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures," *Pain*, vol. 66, no. 1, pp. 3–8, Jul. 1996.
[51] M. J. Grant and A. Booth, "A typology of reviews: An analysis of 14 review types and associated methodologies," *Health Inf. Libraries J.*, vol. 26, no. 2, pp. 91–108, Jun. 2009. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1471-1842.2009.00848.x
[52] G. W. Suter, "Review papers are important and worth writing," *Environ. Toxicology Chem.*, vol. 32, no. 9, pp. 1929–1930, Sep. 2013. [Online]. Available: https://setac.onlinelibrary.wiley.com/doi/abs/10.1002/etc.2316
[53] K. O'Gorman and R. MacIntosh, *Research Methods for Business Management*. Oxford, U.K.: Goodfellow Publishers Limited, Sep. 2014.
[54] W. Bandara, E. Furtmueller, E. Gorbacheva, S. Miskon, and J. Beekhuyzen, "Achieving rigor in literature reviews: Insights from qualitative data analysis and tool-support," *Commun. Assoc. Inf. Syst.*, vol. 37, no. 1, p. 8, 2015.
[55] S. Samnani, M. Vaska, S. Ahmed, and T. Turin, "Review typology: The basic types of reviews for synthesizing evidence for the purpose of knowledge translation," *J. College Physicians Surgeons Pakistan*, vol. 27, no. 10, pp. 635–641, 2017.

NATALIA CIEPLIŃSKA received the bachelor's degree in electronics and telecommunications and the master's degree in mechatronics from the AGH University of Science and Technology, Kraków, Poland, where she is currently pursuing the Ph.D. degree. Her research interests include QoE and psychology.

LUCJAN JANOWSKI received the Ph.D. degree in telecommunications from the AGH University of Science and Technology, Kraków, Poland, in 2006. In 2007, he was a Postdoctoral Researcher with the Laboratory for Analysis and Architecture of Systems, Centre National de la Recherche Scientifique, Paris, France. From 2010 to 2011, he was a Postdoctoral Researcher with the University of Geneva, Geneva, Switzerland. From 2014 to 2015, he was a Postdoctoral Researcher with the Telecommunications Research Center Vienna, Vienna, Austria. He is currently an Assistant Professor with the Department of Telecommunications, AGH University of Science and Technology. His research interests include statistics and probabilistic modeling of subjects and subjective rates used in QoE evaluation.

KATRIEN DE MOOR (Member, IEEE) received the Ph.D. degree in social sciences from Ghent University, 2012, with a thesis on bridging gaps in quality of experience research and its challenges. She is currently an Associate Professor at the Department of Information Security and Communication Technology, NTNU, mainly focusing on socio-technical approaches in ICT research. She is the Co-Editor-in-Chief of the multidisciplinary journal *Quality and User Experience* (Springer). Her research interests include human/user experiences with technology (quality of experience, user experience, user engagement, and immersive experiences), related methodological challenges (ecological validity and user diversity), and ethical implications (e.g., data privacy, human agency, power dynamics in design processes, and ecological footprint of ICT). She has published her work in a range of international, peer-reviewed journals, conferences, and books, and has served in various program committees of international conferences and workshops. She is one of the 20 founding members of the Young Academy of Norway.

MICHAŁ WIERZCHOŃ is currently a Professor with Jagiellonian University, the Director of the Centre for Brain Research, Jagiellonian University, the Vice-Dean for the Scientific Affairs, Faculty of Philosophy, Jagiellonian University, and the Director of the International Ph.D. Program in Cognitive Neuroscience (CogNeS). He is also the Founder and the Head of the Consciousness Laboratory (C-Laboratory). He is also a fellow of the Psychonomic Society. He was an Elected Member of the Committee of Psychology of the Polish Academy of Science (2020–2023). He was the Past Director of the Institute of Psychology, Jagiellonian University, from 2016 to 2020, a Past Member, from 2012 to 2017, and the Past Deputy Chairperson of the Polish Young Academy, Polish Academy of Sciences, from 2014 to 2017. He was also the Past Secretary of the European Society for Cognitive Psychology Executive Committee (2010–2017) and a Past Member of ESCoP Executive Committee (2015–2018). He worked as a Visiting Researcher at the CO3 Laboratory, Universite Libre de Bruxelles, from 2010 to 2011.