Quality and User Experience (2023) 8:6
https://doi.org/10.1007/s41233-023-00059-2

REVIEW ARTICLE

Towards enhancing ecological validity in user studies: a systematic review of guidelines and implications for QoE research

Sruti Subramanian[1] · Katrien De Moor[1] · Markus Fiedler[2] · Kamil Koniuch[3] · Lucjan Janowski[3]

Received: 31 January 2023 / Published online: 5 July 2023

Abstract

The concept of conducting ecologically valid user studies is gaining traction in the field of Quality of Experience (QoE). However, despite previous research exploring this concept, the increasing volume of studies has made it challenging to obtain a comprehensive overview of existing guidelines and the key aspects to consider when designing ecologically valid studies. Therefor this paper aims to provide a systematic review of research articles published between 2011 and 2021 that offer insight into conducting ecologically valid user studies. From an initial count of 782 retrieved studies, a final count of 12 studies met the predefined criteria and were included in the final review. The systematic review resulted in the extraction of 55 guidelines that provide guidance towards conducting ecologically valid user studies. These guidelines have been grouped within 8 categories (*Environment, Technology, Content, Participant Recruitment, User Behavior, Study Design, Task* and *data collection*) overarching the three main dimensions (Setting, Users and Research Methodology). Furthermore, the review discusses: the flip side of ecological validity, the implications for QoE research, as well as provides a basic visualisation model for assessing the ecological validity of a study. In conclusion, the current review indicates that future research should address more in detail how and when research approaches characterized by high ecological validity (and correspondingly, low internal validity) and those characterized by low ecological validity (and normally high internal validity) can best complement each other in order to better understand the key factors influencing QoE for various types of applications, user segments, settings. Further, we argue that more transparency around the (sub)dimensions of ecological validity with respect to a particular study or set of studies is necessary.

Keywords Quality of experience (QoE) · Influence factors · Ecological validity · External validity · Guidelines · User experience (UX)

✉ Sruti Subramanian
sruti.subramanian@ntnu.no

Katrien De Moor
katrien.demoor@ntnu.no

Markus Fiedler
markus.fiedler@bth.se

Kamil Koniuch
koniuch@agh.edu.pl

Lucjan Janowski
lucjan.janowski@agh.edu.pl

[1] Department of Information Security and Communication Technology, Norwegian University of Science and Technology, Trondheim, Norway

[2] Department of Technology and Aesthetics, Blekinge Institute of Technology, Karlskrona, Sweden

[3] Department of Electronics and Telecommunications, AGH University of Science and Technology, Krakow, Poland

Introduction

Over the last decade, there have been several attempts to broaden the theoretical understanding of Quality of Experience (QoE) from a narrow, Quality of Service (QoS)-oriented to a more holistic and human-centered perspective. The latter acknowledges the highly subjective, dynamic and layered nature of technology users' experiences and the quality of these experiences. In [1], QoE has been defined as "*the degree of delight or annoyance of a person whose experiencing involves an application, service, or system. It results from the person's evaluation of the fulfilment of his or her expectations and needs with respect to the utility and/or enjoyment in the light of the user's context, personality and current state*". This definition goes beyond the previously common, more utilitarian approach to QoE. It explicitly emphasizes the importance of positive emotions

such as delight and enjoyment—which have been linked to the fulfillment of hedonic needs in psychological literature (see e.g., [2])–as a desired outcome. By equaling QoE to a dynamic affective state (degree of delight or annoyance), its transient and relative character is fully recognized. It also implies a willingness to investigate and understand how experiencing of technological artefacts involves and moves people emotionally [3].

It is furthermore acknowledged that a range of complex and strongly interrelated factors may influence QoE. In [4], these have been classified into:

1. Properties at the human level, some of which may have a dynamic character, whereas others are dispositional or are relatively stable over time
2. Factors at the level of the system and which may be content-, device-, network- and media-related
3. Factors related to the context (3) also need to be considered, since human experiences do not take place in a vacuum. The latter may relate to the physical, social, economic, task-related and temporal context [4, 5]

This more holistic perspective on QoE represented a fundamental step forward for the field. At the same time however, it entails important implications and challenges for QoE research, for instance in terms of measurement approaches, operationalisation of influence factors and used research settings. In this regard, *Ecological validity*—the extent to which study/experimental findings can be generalized to real world situations—is largely gaining interest within the QoE community [6–8]. Traditionally, QoE experiments take place in a controlled and artificial laboratory setting in which one or a set of factors are isolated, so that the influence and weight of specific factors on QoE can be quantified while guaranteeing the *internal validity* of the findings. The latter refers to the plausibility that "there is a causal relationship between treatment and outcome" [9], e.g., between the delay level and Mean Opinion Scores (MOS) in a QoE experiment. However, this also means that the human subject and the experience is taken out of the natural, real world environment and that ground truth data are collected in research settings characterized by a low *ecological validity* [9]. The concept of ecological validity is especially crucial when conducting studies with human participants and focusing on their subjective experiences. In this regard, findings obtained from studies conducted with low ecological validity are more difficult to generalize as they may not be representative of the real world and may fail to capture natural user behavior [10].

Over the last years, several researchers have therefore made a plea to explore approaches that aim to study and understand QoE also in more natural, real-world contexts (see e.g., [6–8]). Approaches and paradigms outside of the lab have therefore gained more prominence, which is also reflected in the increasing number of studies that investigate QoE outside of the artificial lab setting. For instance, this has been done through crowd-sourcing studies [11], living lab studies [12] and studies combining traditional lab and field-based approaches [13].

Hence, the notion of conducting ecologically valid user studies has gained popularity in fields such as QoE [11, 14], and Human-Computer Interaction (HCI) [15–17]. In the context of QoE, ecological validity becomes particularly significant as it focuses on assessing the quality and usability of technologies, services, or systems as they are used in everyday life, also aiming to capture the subjective experience of users, considering various factors such as perception, emotions, cognition, and context. In this respect, previous studies [7, 8, 10] have explored the concept of ecological validity in experimental design. However, with the growing body of research it has become increasingly difficult to obtain an overview of existing guidelines and the various aspects to consider when aiming to design ecologically valid studies. Furthermore, the concept of ecological validity currently fails to provide objective criteria for experimental design which additionally exacerbate the challenge [18]. Therefore, despite the increasing attention and need for designing ecologically valid studies, there is currently limited understanding of how to design ecologically valid studies and which dimensions are particularly important in this respect.

The overall aim of this study is therefore to synthesize and provide a systematic overview of existing guidelines for conducting ecologically valid QoE user studies. In addition, based on these guidelines, we propose a basic visualisation model that QoE researchers can adopt to map the ecological validity of their user study in terms of various dimensions. More concretely, the study is based on a *systematic literature review* (SLR) and involved:

- *Mapping of Relevant Studies:* First, we mapped existing studies that provide guidelines for conducting ecologically valid studies within the fields of QoE and HCI.
- *Grouping of the Extracted Guidelines:* Next, based on relevance the extracted guidelines were coded and categorized into different dimensions and corresponding categories.
- *Providing a Basic Visualisation Model for Assessing Ecological Validity:* Based on the identified dimensions and the corresponding categories the guidelines were grouped into, we provide a basic visualisation model for assessing the ecological validity of a given study design.

The article is further structured as follows. Section 2 provides background and insight into related work. Next, Sect. 3 describes the research methodology opted. The results are then presented in Sect. 4, and Sect. 5 provides a discussion

of the main findings. Finally, the conclusion is presented in Sect. 6.

Background

The evaluation of different types of validity, i.e., how accurate (in terms of study design, execution and analysis) are the results of a certain study, is a key pillar in scientific research [9]. Validity should be distinguished from the aspect of reliability i.e., how stable and consistent is what has been measured and would the same results be obtained if the same procedure and conditions would be repeated? Reliability is therefore more strongly linked to reproducibility, and can be threatened by various aspects, e.g., participant or observer bias or different types of error [9]. For example, in a classical ACR (Absolute Category Rating) study measuring compression, we have two main variables: the opinion score O and the compression ratio X. While validity refers to the extent to which O reflects the user's QoE, reliability reflects the consistency and stability of the results O in relation to X.

Studies based on an experimental paradigm typically have a strong focus on aspects that are related to their internal validity. If it is plausible that the experimental manipulation or treatment in a study caused the observed change in the dependent variable(s), the internal validity is usually considered high [9]. However, there are many possible threats (such as, participant bias) to this internal validity that need to be taken into account [19].

After conducting a lab-based experimental study, a logical question to address is to which extent the obtained results are generalisable. The external validity or generalisability of laboratory studies is typically very low, as the results apply to the specific test and lab conditions that were used [9]. In the context of QoE studies, which are heavily guided by recommendations issued by the International Telecommunications Union (ITU) (see [20] for an overview of relevant recommendations per application area), the lack of ecological validity has repeatedly been put on the agenda [13, 21–23]. In response to this observation, new methodological approaches have been suggested and tried out over the years.

For instance, in [24], living lab approaches were proposed as a way to bridge the gap between the lab and users' natural environment. While such "in the wild" studies were conducted in the last decade, they come with distinct challenges (e.g., privacy and user consent, low samples and participant drop-out, noisy data and challenging data analysis,...). Other approaches went fully away from the experimental test design and focused for instance on evaluating QoE in a home setting and thereby better integrating the daily life context into the evaluation of QoE [25, 26]. While these studies were considered as highly disruptive and novel, they did not trigger the intended methodological renewal and did not result in new recommendations for QoE studies.

Yet another body of work focused more on improving the ecological validity of experimental studies in the lab, for instance by using more immersive test paradigms which ensure participants view realistic content without any repetition [23]. The latter avoids e.g., to repeat content or to use meaningless test stimuli in order to avoid that users loose focus and get bored. Other strategies that have been proposed in this respect include the use of dedicated measures to keep participants' attention high (e.g., by adding an engaging task) [22, 23]. A more recent effort [14] used a gamification-based approach to increase the ecological validity of subjective lab studies. More concretely, the authors designed a 5-minute mixed-reality escape room game to investigate several dimensions (i.e., a realistic stimulus, a tailored response measure and a realistic research setting) of ecological validity in the context of Mixed Reality interaction technology QoE [14].

Finally, as a more hybrid approach, crowdsourcing has gained momentum over the last decade [27, 28]. While crowdsourcing primarily can help to make subjective testing more resource-efficient and can help to overcome some of the typical challenges associated with lab testing (e.g., typically lower number of participants and recruitment challenges, lack of population diversity and bias in participant profiles) [29], it also comes with a set of challenges that need to be properly addressed in the study design and data analysis. For instance, as in field studies, the experimenter has a much lower degree of control over the participants than in a lab setting [29]. Further, reliability of workers, strategies to detect unreliable participants and appealing task design are of crucial importance [29, 30].

However, despite these and other approaches that have been proposed in the literature and that address the challenges related to ecological validity of QoE studies, there is still a lack of shared guidelines and recommendations on how to conduct subjective studies in a more ecologically-valid way [22]. In addition, to the best of our knowledge, there is no common instrument to report on the ecological validity of a study. Such an instrument would allow to situate and classify individual studies in terms of their ecological validity on different dimensions and could help with the interpretation of the results, identification of blind spots in the literature and triggering of more awareness around internal vs. external validity trade-offs. The work presented in this paper aims to make a contribution in this overall direction.

Methods

Systematic literature reviews (SLRs) clarify the state of existing research and the corresponding implications that should be drawn [31, 32]. As defined by Fink (2005) [33]

a systematic literature review is: "a systematic, explicit, [comprehensive,] and reproducible method for identifying, evaluating, and synthesizing the existing body of completed and recorded work produced by researchers, scholars, and practitioners" (pp. 3, 17).

To gain a better overview and understanding of relevant studies that have provided insight into conducting more ecologically valid studies, we performed a systematic literature review of existing research. By following clearly defined protocols and predefined criteria to search through existing research we were able to identify 55 recommendations and rationales provided in the literature to perform more ecologically valid studies. The entire procedure opted including databases and search terms, selection process, and analysis is described in the following sections.

Databases and search terms

The papers included in this review were retrieved on 27 August 2021 from four main databases: ACM, IEEE, Web of Science, and Scopus. The search string used to retrieve the studies from the various databases was constructed iteratively, as it was identified that a variety of terminologies are used to denote the same concept. Considering that the aim of this review was to identify recommendations to conduct more ecologically valid studies within the domain of QoE and User Experience, multiple keywords related to the distinct categories of *recommendations*, *ecological validity*, *research setting*, and *domain* were used. Despite limiting the domain to QoE and HCI, a wide variety of recommendations from studies corresponding to various sub and super domains were identified. The final query including boolean operators (AND, OR) and truncations (denoted by an asterisk) can be found in Appendix 1. The syntax of the query was further adapted to the specific format of each database.

Selection process

In addition to the search string, specific inclusion and exclusion criteria were used as a part of the selection process (see Table 1).

The database search identified a total of 782 studies. From the initial corpus, removing duplicates, non-English articles, non-peer reviewed articles, extended abstracts/short papers, books, dissertations, and articles older than 10 years (101 studies) resulted in a total of 681 studies. These 681 studies further went though the second phase of title and abstract screening to identify articles providing recommendations for conducting ecologically valid studies either as key or supplementary contributions, thereby eliminating 587 articles. The remaining 94 studies were assessed in full-text, resulting in a total of 12 studies that were included in the review. For example, studies which claimed to provide relevant insight in the abstract but which did not follow through in providing concrete recommendations in the body of the text were eliminated during this phase. Additional studies were also eliminated based on the inclusion/exclusion criteria in this phase. The final corpus was obtained by following the rigorous approach as presented in. Figure 1.

The 12 studies included in the final review were read in full by all the authors. As previously mentioned, studies that not only offer explicit recommendations but also those which provide any relevant insight into conducting ecologically valid studies were included in the review. For instance, Maki et al. (2013) [34] do not provide any explicit recommendations. Rather, the article provides insight on differences identified between conducting studies within the lab and in a public setting. This article was further included in the review because of its relevance. Table 2 presents the list of final studies included in the review and the corresponding scope of the studies.

Analysis

Data on the studies' scope, aim, recommendations, and rationale for the recommendations were extracted. These findings are presented in Tables 3, 4, 5, 6, 7, 8, 9, 10 within the Results section. The recommendations and rationales provided in the individual studies were iteratively coded for themes using an open coding approach by the first author. With each review article, codes were consistently created, removed, renamed, and rearranged to accommodate relevance and readability.

Table 1 Inclusion/exclusion criteria

Inclusion criteria	Exclusion criteria
Articles providing recommendations to conduct ecologically valid studies as either key or supplementary contribution of the article	Non-english articles
	Non-peer reviewed articles
	Extended abstracts
	Short papers
	Books and dissertations
	Articles older than 10 years

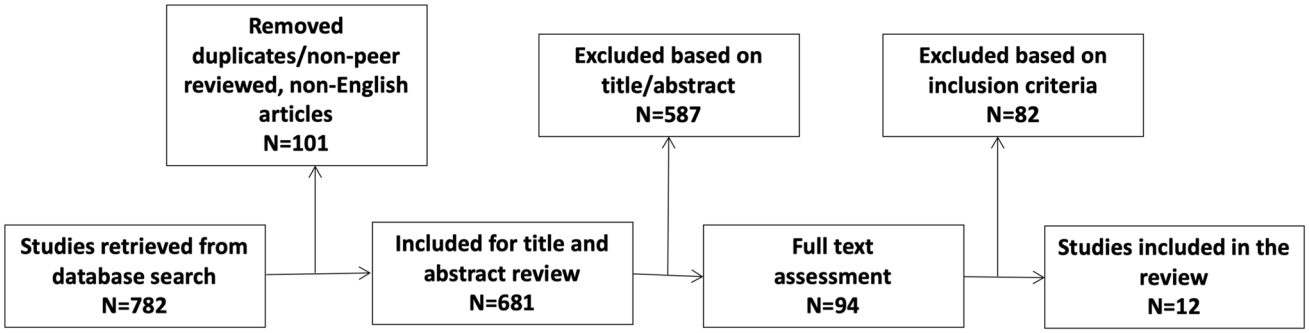

Fig. 1 Flow chart of the study selection

Table 2 List of studies included in the review

Study	Scope
Lew et al. (2011) [35]	HCI
Maki et al. (2013) [34]	QoE
Staelens et al. (2014) [36]	QoE
Maguire and Renaud (2015) [37]	Security
Robitza et al. (2015) [23]	QoE
Robitza et al. (2016) [38]	QoE
Robitza and Raake (2016) [39]	QoE
Mottelsen and Hornbæk (2017) [40]	HCI
Blum et al. (2018) [41]	Vibrotactile haptics
Labonte-LeMoyne et al. (2018) [42]	Physiological computing
Dole and Ju (2019) [10]	HCI
van Berkel et al. (2020) [43]	HCI

Results

Based on the iterative coding that was performed (as described in Sect. 3), the identified guidelines for conducting ecologically valid studies were classified into 8 categories: Environment, Technology, Content, Participant Recruitment, User Behaviour, Study Design, Task and Data Collection. These categories were grouped within three main dimensions: Setting, Users, and Research Methodology. The three-dimensional classification of the guidelines comprising the various categories are illustrated in Fig. 2.

The following sections provide an overview of the extracted guidelines. The various categories of guidelines are presented within the three main overarching dimensions, i.e., Setting, Users, and Research Methodology. Furthermore, the 8 different categories of guidelines are presented within separate Tables 3, 4, 5, 6, 7, 8, 9, 10). While the extracted guidelines are from studies within different scopes (e.g., HCI, QoE, Security, physical computing and vibrotactile haptics), the guidelines are generic and applicable to a variety of studies.

Setting

This dimension comprises of all guidelines corresponding to the study set up and comprises of categories: Environment, Technology, and Content.

Environment

The category of environment groups the guidelines that refer to the environmental aspects of a study (see Table 3).

Maki et al. (2013) [34] and Labonte-LeMoyne et al. (2018) [42] suggest guidelines related to the environmental aspect of conducting a study. As compared to conditional lab environments, the real world is filled with background noise, such as people talking for instance. Similarly is the case for variable illumination in the real world as compared to lab setups as pointed out by Maki et al. (2013) [34]. Labonte-LeMoyne et al. (2018) on the other hand suggest that in-lab simulated environments can be more conclusive than field studies, taking into consideration the complexities and cost of natural settings versus that of a simulated environment. Lew et al. (2011) [35] state the need for realism in the the overall setting, as the authors claim that lack of realism can indicate to the

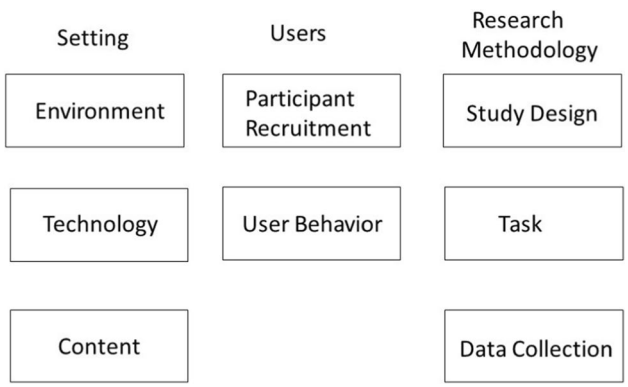

Fig. 2 Classification of Guidelines

Table 3 Guidelines for environment

Study	Scope	Environment Recommendation	Rationale
Maki et al. (2013) [34]	QoE	Consider background disturbance (people discussing, working etc.)	As compared to conditioned background in lab experiments
Maki et al. (2013) [34]	QoE	Consider variable illumination	As compared to controlled illumination in lab experiments
Maki et al. (2013) [34]	QoE	Consider the audio disturbance caused by other people	As compared to no audio disturbance in labs
Maki et al. (2013) [34]	QoE	Consider that participants are surrounded by other people	As compared to Full privacy in labs
Labonte-LeMoyne et al. (2018) [42]	Physiological computing	in-lab simulated environments can be more conclusive than going in situ	Considering the complexities and costs of naturalistic vs. simulated environments
Lew et al. (2011) [35]	HCI	Conduct studies in a realistic setting	Participants provide inauthentic responses which are only observed in lab settings
Staelens et al. (2014) [36]	QoE	Everyday-life context should be integrated in subjective video quality experiment. E.g., as in the current study by going to the people's home and letting them watch the video sequences on the tablet device in their typical home environment	Subjects are placed in a more ecologically valid testing environment
Dole and Ju (2019) [10]	HCI	Include key structural conditions from the real world into the lab	Doing so will produce mundane and likely experimental realism in the minds of participants
van Berkel et al. (2020) [43]	HCI	Consider the larger context in which the device is evaluated and construct user goals around this to construct scenarios which encompass device use from start to finish	Presenting participants with rigid tasks containing a single start- and end-point unrepresentative of a real-world experience is a pitfall
van Berkel et al. (2020) [43]	HCI	Ensuring contextual realism while accounting for patient safety and study design requirements	Evaluating the device in isolation without consideration of the environment in which it will be used is a common pitfall

participants that they are being tested. The authors also indicate the need to conduct studies in a more realistic setting as increasing this parameter of realism on all levels provides participants with a more authentic and life-like perspective thereby increasing the overall ecological validity of a given study.Furthermore, lack of realism in the setting may persuade participants to provide inauthentic responses or behaviors that are not observed outside the laboratory setting. Several of the studies highlight the need for contextual realism [10, 36, 43]. More specifically, Staelens et al. (2014) [36] suggest the need to integrate everyday life context such as going to participants' home, or including key structural conditions from the real world to the lab setting [10]. Similarly, van Berkel et al. (2020) [43] suggest taking into consideration the larger context in which device or technology is evaluated, since unrepresentative real world experiences represent a pitfall in terms of conducting ecologically valid studies.

Technology

The guidelines corresponding to the technological aspect of conducting studies are grouped within this category as seen in Table 4.

Based on studies conducted in a lab and a public setting (i.e., an exhibition hall), Maki et al. (2013) [34] offer insight into the difference in screens used for both types of studies. The authors indicate that while in a lab environment an external monitor is used, in the real world participants would use their own laptop screens which should be taken into consideration. Maguire and Renaud (2015) [37] further suggest ensuring that the technology/application being tested exists beyond the evaluation period, as users may come to depend on it or the ecology supporting it. Different technologies, even different models of technology have a slight difference in simple variables such as speed, accuracy and throughput which should be taken into consideration to also

Table 4 Guidelines for technology

Study	Scope	Technology Recommendation	Rationale
Maki et al. (2013) [34]	QoE	Consider the Laptop's internal monitor	As compared to External monitors in labs
Maguire and Renaud (2015) [37]	Security	The application should exist beyond the evaluation period	participants could come to depend on the application or the ecology supporting it
Mottelsen and Hornbæk (2017) [40]	HCI	Expect simpler dependent variables, e.g., speed, accuracy, throughput to vary with technology, but complex phenomena (such as body ownership, presence) to depend more on internal control	-
van Berkel et al. (2020) [43]	Usability	Beware of the degree to which the completeness of the software and the simulated (patient) data represent the breadth of real-world use cases	Implementation gaps which force unspoken assumptions are a pitfall
van Berkel et al. (2020) [43]	Usability	Consider the effect of prototype fidelity on ecological validity of the study	Overlooking the effect of small differences between prototype and medical grade hardware on user interaction process is a pitfall
Labonte-LeMoyne et al. (2018) [42]	Physiological computing	Projects in-the-wild require ad-hoc modifications or adaptations to the equipment	To allow for personalized setup
Blum et al. (2018) [41]	Vibrotactile haptics	Laboratory studies can use fragile, custom-built haptic devices with little regard for portability or power constraints	Moving to in-the-wild use, however, requires hardware that remains expressive but is also portable and robust enough to provide reliable data throughout the experiment
Blum et al. (2018) [41]	Vibrotactile haptics	To maximize flexibility, it may be necessary to either heavily modify commercial hardware (e.g., by integrating a larger battery), or design and build experimental hardware despite the additional work and drawbacks	Even though commercial haptic products can provide off-the-shelf robustness for in-the-wild use, they tend to limit the customizability of, and access to, the underlying haptic hardware
Blum et al. (2018) [41]	Vibrotactile haptics	In-the-wild experiments often use low fidelity vibro-tactile actuators due to their combination of perceptual intensity, power efficiency, low cost, and compact size	for applications where ensuring perception is critical
Blum et al. (2018) [41]	Vibrotactile haptics	To minimize power consumption, choose a system that uses more efficient actuation mechanisms	Use of the haptic engine also consumes power, and using the engine too much may create a noticeable drain on battery life
Blum et al. (2018) [41]	Vibrotactile haptics	Obtain the greatest perceptual impact from the delivered stimulus under a given power budget	Use of the haptic engine also consumes power, and using the engine too much may create a noticeable drain on battery life
Lew et al. (2011) [35]	HCI	Consider realism in appearance of the interface	Lack of realism in interfaces signals participants that they are being tested

Table 5 Guidelines for content

Study	Scope	Content Recommendation	Rationale
Robitza et al. (2015) [23]	QoE	Ensure subjects don't have to see the same content twice	Proven to be a valid means to make subjective studies more enjoyable, and ecologically valid
Robitza and Raake (2016) [39]	QoE	Let people freely decide what to watch	increases their intrinsic motivation
Robitza and Raake (2016) [39]	QoE	Ask about the content itself instead of the quality	Increases extrinsic motivation to focus on content instead of quality
Staelens et al. (2014) [36]	QoE	Show longer duration of audio/visual content. E.g., in the current study the videos were 2 mins long	Content duration influences users QoE
Lew et al. (2011) [35]	HCI	Consider realism in the content of an interface	Despite using realistic technology, users may feel the system is fake if it doesn't generate appropriate content
Robitza et al. (2015) [23]	QoE	Provide realistic stimuli, both in terms of content and length, and overall enjoyable viewing experiences	It is more realistic and more likely to trigger realistic user responses

Table 6 Guidelines for participant recruitment

Study	Scope	Participant recruitment Recommendation	Rationale
Mottelsen and Hornbæk (2017) [40]	HCI	Pre-screen participants for the technology accessible to them	to avoid recruiting unqualified (based on inclusion/exclusion criteria) people
Mottelsen and Hornbæk (2017) [40]	HCI	Validate the integrity of participants for instance using verifiable control questions, context photos, or user performance	–
Mottelsen and Hornbæk (2017) [40]	HCI	Expect roughly (only) half of the participants to complete the study, so increase the sample size correspondingly	Based on experience
van Berkel et al. (2020) [43]	Usability	Consider the representativeness of the participant sample to the intended end-user group	Assumption that an evaluation with developers can provide a proxy for the end users is a common pitfall
van Berkel et al. (2020) [43]	Usability	Consider that the deep insights offered by realistic interaction between clinician and patient should be offset against the risk introduced to patients by their involvement in a study	Involving patients has numerous challenges in clinical settings (such as safety and between subject comparison)

have more internal control and to avoid potential confounding variabes [40].

Furthermore, while it is common to evaluate prototypes, van Berkel et al. (2020) [43] indicate that implementation gaps force unspoken assumptions among users and are a pitfall to avoid. They underline the importance of considering the effect of prototype fidelity on the ecological validity of a study.

Furthermore, with regard to in-the-wild studies, differences and modification of the technology are required to adapt to the more dynamic environment as compared to a stationary lab setting [41, 42].

Labonte-LeMoyne et al. (2018) [42], state that to allow for personalized setup, in-the-wild studies require ad hoc modifications or adaptations. Similarly, Blum et al. (2018) [41] state that while lab studies can use fragile devices with little regard for portability or power, in-the-wild studies however will require using more expensive hardware that is more robust and portable to allow collecting reliable data. While commercial (haptic) products may provide necessary robustness, there is a limit to the customizability and access to the underlying hardware, hence to maximize flexibility it may be necessary to heavily modify commercial products or even custom design and build experimental hardware despite the added effort [41]. Blum et al. (2018) further state that with respect to vibrotactile haptics, in-the-wild studies benefit from using low fidelity vibrotactile actuators which combine perceptual intensity, power efficiency, low cost, and portability.

Table 7 Guidelines for user behavior

		User behavior	
Study	Scope	Recommendation	Rationale
Maguire and Renaud (2015) [37]	Security	The application should offer benefits and avoid perverse incentives that could influence behaviour	Many assessments rely on incentives that potentially influence frequency of use and acceptance
Robitza et al. (2016) [38]	QoE	Beware that demand characteristics (DC) are strongly present in behavioral tests, leading to subject apprehensiveness and distorted results	if participants are told the experiment's purpose, this could already unconsciously bias their preconceptions
Robitza et al. (2016) [38]	QoE	Influence of demand characteristics would be exacerbated by instructions telling people which actions are possible	It would appear that those actions are wanted by the experimenter
Robitza et al. (2016) [38]	QoE	Hidden and passive measurements will make participants less aware of the measurement context	Reduces DC
Robitza et al. (2016) [38]	QoE	Reducing the involvement of the experimenter may also help	Reduces DC
Robitza et al. (2016) [38]	QoE	Use written and automated test instructions	Reduces any personal influence
Dole and Ju (2019) [10]	HCI	Avoid the introduction of unwanted between subjects variables	Variables like arousal and attention have far-reaching consequences, and should therefore be kept as consistent as possible between participants
Robitza et al. (2016) [38]	QoE	Ask participants about what role they thought they played in the study	To see how they judged their awareness of the research hypothesis, and whether they acted apprehensively
Robitza and Raake (2016) [39]	QoE	Reveal the real purpose of the test at the end of the experiment and access behavior in response to quality degradations	To avoid problems with demand characteristics. Change subjects' mindset from spotting quality degradations

Next, with respect to power consumption Blum et al. (2019) [41] suggest minimizing power consumption and using more efficient actuation mechanics and try to obtain larger perceptual impact under a power budget as haptic engines consume power and using the engine too much may drain battery life. Furthermore, Lew et al. (2011) [35] states the need for realism in the appearance of an interface. The authors claim that lack of realism, such as the touch and feel of a technology can indicate to the participants that they are being tested.

Content

The category of content is specifically related to QoE studies comprising audio/visual content as shown in Table 5.

The guidelines presented in Table 5 provide insight specifically related to the content used in the studies. Robitza et al. (2015) [23] state that to ensure ecological validity and to make subjective viewing experience more enjoyable participants should not have to see the same content twice. Similarly, Robitza and Raake (2016) [39] also suggest allowing participants to freely decide what to watch in order to increase their intrinsic motivation. The authors additionally recommend asking about the content itself rather than the quality as that will increase extrinsic motivation to allow participants to focus on the content instead of quality. Additionally, Staelens et al.(2014) [36] suggest showing longer duration of audio/visual content(e.g., 2 min) since content duration influences the users' QoE. Lew et al. (2011) [35] states the need for realism in the content provided. The authors claim that lack of realism, such as fake content can indicate to the participants that they are being tested. Specifically in the context of QoE, Robitza et al. (2015) [23] suggest providing realistic content, length and a realistic overall viewing experience.

Users

This dimension comprises of all guidelines corresponding to the end users or the participants included in studies and comprises of the categories: participant recruitment and user behaviour.

Participant recruitment

The category of participant recruitment groups all the guidelines related to the recruitment of participants for studies as shown in Table 6.

Mottelsen and Hornbæk (2017) [40] and van Berkel et al. (2020) [43] provide insight regarding the recruitment of participants for studies. Firstly, to avoid recruiting unqualified people, Mottelsen and Hornbæk (2017) [40] suggest pre screening participants with regard to the technology that is

Table 8 Guidelines for study design

Study	Scope	Study Design Recommendation	Rationale
Mottelsen and Hornbæk (2017) [40]	HCI	15 min seems like the maximum tolerable duration for keeping the pose required to use the VR cardboard system	-
Mottelsen and Hornbæk (2017) [40]	HCI	Design experiments well-suited for both standing and sitting	Accommodate for different needs and preferences
van Berkel et al. (2020) [43]	Usability	Consider the amount and availability of training material that can realistically be expected to be taken in by the user of the system prior and during use	Assumption that end-users will have (completely) read, understood, and remembered the device's manual or use instructions is a pitfall
Dole and Ju (2019) [10]	HCI	Ecological validity is a goal but not a destination	It is important to remember the impossibility of declaring that ecological validity has been achieved

Table 9 Task related guidelines

Study	Scope	Task Recommendation	Rationale
Lew et al. (2011) [35]	HCI	Consider realism in Task	Personal relevance of a task is key to understanding UX
Maguire and Renaud (2015) [37]	Security	An authentication mechanism must be coupled with a realistic primary task	An unrealistic coupling is likely to lead consumers to either abandon the application or resort to undesirable coping mechanism
Maguire and Renaud (2015) [37]	Security	An authentication mechanism must have a clear and transparent goal	Users should be empowered with a clear understanding of the consequences of authentication

Table 10 Guidelines for data collection

Study	Scope	Data collection Recommendation	Rationale
Lew et al. (2011) [35]	HCI	Use background logging methods to record a variety of measures unobtrusively	This is more realistic as it measure what actions users actually perform
Lew et al. (2011) [35]	HCI	Use multi-method approaches combining behavioural logging with questionnaires	This can alleviate the problem of mono-method bias
Robitza and Raake (2016) [39]	QoE	Record the interaction of the user with the system in the background	This is unobtrusive
Robitza et al. (2016) [38]	QoE	Crowdsourcing, passive large-scale measurements and longitudinal studies with friendly users can be opted instead of, or in addition to lab studies	To allow for collecting data in a more natural environment
Lebonte-LeMoyne (2018) [42]	Physiological computing	Plan for data loss	Increased ecological validity may lead to data loss
Dole and Ju (2019) [10]	HCI	Measure participants' presence in the simulation	It is participants' presence in a simulation that indexes its face validity

accessible to them. Next, the authors state that it is necessary to validate the integrity of participants for instances by means of verifiable control questions, context photos or user performance [40]. Furthermore, based on their own experience, the authors express that one should roughly expect that only half the participants may complete the study and therefore suggest planning for such dropouts and uncertainties (Which is more common in longitudinal studies).

Similarly, van Berkel et al. (2020) [43] argue that it is essential to consider the representativeness of participant sample to the intended user group, as including proxies for the intended end users is a common pitfall. Furthermore, based on their usability study in a medical context, the authors also recommend rethinking the necessity of involving patients as patient involvement can introduce numerous challenges such as safety and between subject comparison. Specifically with respect to usability evaluations, patient involvement can augment the realism of scenarios and use cases.

User behavior

The category of user behavior groups all the guidelines related to the behavior of users as shown in Table 7.

Maguire and Renaud (2015) [37] indicate that many studies largely rely on incentives which influence the frequency of use and acceptance among users. Therefore, the authors suggest ensuring that the application being tested offers benefits to the users such that other (extrinsic) incentives do not influence user behaviour. Similarly, Robitza et al. (2016) [38] highlight the presence of demand characteristics (DC), i.e., where participants form an interpretation of the study's purpose and subconsciously change their behavior to fit the interpretation [44]. The authors instruct about DC and that it can lead to distorted results. Hence it is recommended to avoid informing participants about the purpose of the study, as well as instructing participants about possible actions to take, as such actions can amplify DC. In this regard, it is also suggested to use hidden and passive measurements to make participants less aware of the measurement context as well as reducing the overall involvement of the experimenter to reduce DC. Furthermore, it is suggested to use written and automated test instructions to reduce any personal influence of the experimenter. Additionally, Dole and Ju (2019) [10] recommend minimizing the inclusion of subject variables, such as arousal and attention, due to their significant impact. Hence, they propose maintaining a high level of consistency in subject variables across participants.

Further, in order to evaluate how participants judged their understanding of the research hypothesis and whether they acted apprehensively, Robitza et al. (2016) [38] recommend asking participants about what role they thought they played in the study. In the context of audio/video QoE, Robitza an Raake (2016) [39] suggest revealing the real purpose of the study only at the end in order to access user behavior in response to quality so as to change participants' mindsets from spotting quality degration.

Research methodology

The dimension of research methodology comprises of the categories and the corresponding guidelines related to the research process, and includes the categories: study design, data collection and task.

Study design

The study design category groups all the guidelines related to design of research studies as shown in Table 8.

With regard to VR technology, Mottelsen and Hornbæk (2017) [40] suggest limiting VR studies to a maximum of 15 min as that has been identified as the maximum tolerable duration. Furthermore, the authors suggest designing studies which are suitable for both standing and sitting positions to accommodate the needs and preferences of the participants. van Berkel et al. (2020) [43] indicate that it is a pitfall to assume that end users will have understood and remembered how to use a device and its instructions, hence it is recommended to consider the amount and availability of training (material) that users can take in prior to and during the use of a system. Finally, Dole and Ju (2019) [10] highlight that ecological validity is a goal but not a destination as it is impossible to declare the achievement of ecological validity.

Task

The category of task groups all the guidelines related to the actual tasks performed in studies as shown in Table 9.

A task involves a range of activities that participants are expected to perform during studies. Lew et al. (2011) [35] indicates the need for realistic tasks. The authors claim that a lack of realism, such as the touch and feel of technology and fake content can indicate to the participants that they are being tested. Maguire and Renaud (2015) [37] further state the need to provide realistic tasks with clear and transparent goals as clear understanding is necessary and that a lack of realism in this regard can lead participants to abandon the application or task.

Data collection

This category groups the guidelines specific to the process of data collection as shown in Table 10.

Lew et al. (2011) [35] suggest using background logging methods to record a variety of measures unobtrusively including file revision histories, screen sharing, integrated videorecording, and mouse movement tracking. The authors state that such an approach is more realistic as it measures what actions users actually perform in contrast to questionnaires which only provide self-reported exposure to the causal variable in question [45].

Similarly, Robitza and Raake (2016) [39] recommend recording the interaction of users with the system in the backgound in an unobtrusive manner. Further, it is also recommended to use crowdsourcing, passive large-scale

measurements and longitudinal studies with users to allow for data collection in a natural environment [38]. In addition, Lebonte-LeMoyne (2018) suggest that it is necessary to plan for data loss in terms of time and participants recruitment as this may be common when increasing ecological validity. Additionally, with more relevance for VR technology, Dole and Ju (2019) [10] suggest measuring participants' presence in the simulation as an index to face validity.

Discussion

The systematic categorization of the guidelines provide insights and opportunity for kick-starting a discussion in the QoE community on a number of topics:

- The Flip Side of Ecological Validity
- Implications for QoE Research
- A Basic Visualisation Model for Assessing Ecological Validity

The flip side of ecological validity

While there is currently an emphasis on increasing the overall ecological validity of studies in various fields, there are, however, undesirable outcomes in doing so specifically in the context of QoE and conducting user studies. While all the articles reviewed in this study provided insights into the how and why of conducting more ecologically valid studies, some of the articles [35, 42] also highlighted the flip side of increasing the ecological validity which is something to remember while designing studies.

While numerous guidelines that highlight the necessity of realism on various levels, Lew et al. (2011) [35] indicate that increasing realism can introduce more noise into the collected data, and that the results obtained on one particular day may not generalize to other days. However, the authors further provide suggestions to overcome these issues by measuring behavioural metrics and data logging unobtrusively in the background, and opting a multi-method approach combining behavioural logging with questionnaires (as presented in the results Sect. 4). Furthermore, with regard to realism, Lebonte-LaMoyne et al. (2018) [42] state that by allowing participants to act naturally they do not repeat the same action multiple times. This could however also be problematic since without repetition there may arise uncertainty as to whether the responses are good representations of the stimulus in the real world. Similar to Lew et al. (2011), the authors also indicate that increasing ecological validity will often lead to increased data loss, hence Lebonte-LeMoyne et al. (2018) [42] indicate that while it is necessary to design more ecologically valid studies, it is also crucial to remember the flip side of increasing ecological validity and to plan studies accordingly (e.g., considering time and participant recruitment). The recommendation provided by Dole and Ju [10], namely to keep in mind that: *Ecological validity is a goal but not a destination*, is indeed an important one. Future work in this direction should therefore strive for more awareness and more transparency regarding a study's ecological validity and the implications of various decisions and choices on its results and implications. However, the existing challenge of designing more ecologically valid studies is further exacerbated as it is currently difficult to declare the degree of ecological validity in a transparent and standardized way, due to the absence of a standardized or shared instrument to support this.

The current review provides a categorized overview of the dimensions and components that influence the ecological validity of a study. In the next section, we briefly discuss a number of key considerations and implications when trying to apply these guidelines and dimensions in the context of QoE research.

Implications for QoE research

While the identified guidelines and their rationale were extracted from scientific articles with a broader scope than only QoE, many of them could inspire the modification and reconsideration of existing test designs and methodological approaches in the field of QoE. While deeper analysis in this respect goes beyond the scope of this paper and is left to future work, we briefly discuss the sub-dimensions and their guidelines in terms of potential implications for QoE studies.

In terms of the dimension *Setting*, the following sub-dimensions were presented:

- **Environment**: Guidelines that were extracted from the literature in this respect dealt with for instance the need to conduct tests with realistic background noise (e.g., voices, children around, talking colleagues around during a conference call,...) and illumination settings. Incorporating such environmental aspects to a larger extent in QoE studies could increase the ecological validity on this dimension.
- **Technology**: In terms of technology, numerous aspects were mentioned: for instance, the fidelity of a prototype or mock-up (e.g., a test website or landing page of a streaming service) should be carefully considered. Further, when using participants' own devices (e.g., field testing, crowdsourcing), their specific capabilities and properties should be captured and taken into account as good as possible. Further, in the context of e.g., living lab-based approaches and the use of passive monitoring tools and data collection approaches,

power/energy consumption should be considered, so that the measurement approach does not drain the battery significantly. Realistic distortions are another aspect of ecological validity. This aspect of the study is especially difficult to address. The Internet works at least "good" in most situations, so if a test is close to a real-life scenario, there should be nearly zero distortions and distortions in the content should be realistic and introduced in such a way that it resembles what might happen in a real-world setting. Nevertheless, in such a scenario, the measurements may become less interesting from a research point of view. Here, a potential approach could be to motivate the type of distortion that is considered, e.g., by asking real users or through observation.

- **Content**: For content, the guidelines closely related to the earlier mentioned realism: the content should ideally be meaningful and appealing to the test subjects. Preference should be given to tests designs in which users can e.g., pick the content themselves. Generally, repetitions should be avoided and as mentioned earlier, also the duration of the test stimuli / content should be meaningful.

In terms of the dimension *User*, the following sub-dimensions were presented:

- **Participant**: With regards to participants, the main recommendation is to clearly define what the intended user group or population is, to use appropriate screening techniques and to be transparent and aware of bias that might be introduced due to pragmatism in recruitment. While the use of screening techniques (e.g., visual acuity, colour blindness) is common in subjective studies, there is still a way to go when it comes to user diversity and awareness around the implications of using non-representative user populations in user studies (e.g., using university students only). The increased use of crowdsourcing approaches has already helped to overcome existing challenges to a certain extent, but such approaches and platforms also may contain inherent bias (e.g., user profiles that are not represented).
- **User behavior**: The guidelines and rationales dealing with user behavior boil down to reducing the demand characteristics and—via various measures—to try to have a set-up which allows users to behave as naturally as possible. These include for instance: using passive measurements and collecting indirect user feedback, not disclosing the real purpose of a study, to use only written and automated test instructions, to try to assure that there is an intrinsic motivator in the test design (e.g., gamification) to avoid that people behave in a certain way to get another (monetary) incentive. One additional recommendation that could easily be implemented in most cases is to ask participants in the debriefing session how they experienced and interpreted their role in the experiment or study.

In terms of the dimension *Methodology*, the following sub-dimensions were presented:

- **Study Design**: To start with, it should be kept in mind that ecological validity is a goal that contains multiple dimensions, but not a destination in itself. It always depends on the type of study and purpose to which extent ecological validity is of key importance overall, or e.g., only in terms of certain dimensions for which high ecological validity is required. A few of the extracted guidelines might be relevant to reflect upon from the perspective of QoE studies. First of all, it was suggested important to have a certain flexibility to adjust to different preferences (e.g., play a game standing or sitting). Further, it was also suggested to not overestimate users' capabilities in remembering everything in e.g., a detailed training session. The study and training session (if relevant) should be designed such that unnatural fatigue is avoided.
- **Task**: To increase the ecological validity of the task, the extracted guidelines indicate that the task should be realistic, clear and have transparent goals. Further, one of the guidelines suggests a redesign of some of the commonly used experimental designs to ensure that the secondary task (e.g., evaluate quality) is linked to a meaningful and realistic primary task (e.g., completing an assignment during a telemeeting, ...). What this more precisely implies for common subjective test designs largely depends on the purpose of using a certain service and application area / type of service, but making quality evaluation a secondary task could help to increase ecological validity.
- **Data Collection**: Finally, in terms of data collection, the implications of the extracted guidelines can be summarized by a need to have robust data collection approaches, which are preferable based on a combination of methods and measures. This means that preference should be given to multi-method approaches, using different types of measures, including behavioral and passive measures and a reduced reliance on self-report measures. One effective approach to understanding user perceptions and behaviors in real-world settings as suggested by lew et al. (2011) [35] is by using background logging methods to record a variety of measures unobtrusively. By capturing user interactions in real-time and real-world contexts, the data collected is more likely to represent

users' genuine experiences in a non-intrusive manner. In such cases, data can be collected in the background through customized mobile applications running in the background and logging various user interactions and contextual information (e.g., MobileDNA). Other passive monitoring tools that run on users' devices and collect data without requiring active user involvement can be considered. These tools can track web browsing behavior, application usage, system logs, or network performance. For instance tools such as Wireshark for network monitoring or analytics platforms like Google Analytics to gather web-based user data could be used. Furthermore, wearable devices equipped with sensors can be considered to collect physiological or contextual data. These devices can track heart rate, sleep patterns, location, or environmental factors. Ensure that the devices are non-intrusive, comfortable, and have sufficient battery life for extended data collection periods. To gain a more comprehensive understanding of user experiences, a combination of background logging with user surveys or interviews could be beneficial. These qualitative methods can capture subjective perceptions, preferences, and emotions that may not be evident from logged data alone. However, as mentioned above, some of these guidelines and the approaches may result in ethical dilemmas and may even meet user resistance. Hence, ecological validity is a concern that needs to be balanced with other concerns.

A basic visualisation model for assessing ecological validity

The identified guidelines and the corresponding categorization of these extracted guidelines allowed for proposing a basic visualisation model for assessing the ecological validity of a study in terms of the identified sub-dimensions. While providing a final and complete model for assessment is beyond the scope of this paper, we do provide a basic visualisation model which we briefly introduce. This model will be further detailed and expanded in our future work.

The proposed model (see Fig. 3) involves three main dimensions: Setting, User and Research Methodology. Furthermore, as previously presented in the Results Sect. 4, each dimension comprises of corresponding categories. The Setting dimension comprises of categories: Environment, Technology, and Content. Next, the User dimension consists of categories: Participant Recruitment, and User behavior. The third dimension, i.e., Research Methodology included categories: Study design, Task and Data collection. In this regard, the proposed model captures the significant dimensions and the corresponding categories that should be taken into consideration when designing an ecologically valid study.

The visualisation is denoted by a dark shade in the center growing lighter towards the edge, thereby indicating the increasing degree of ecological validity from the center of the model to the periphery. Hence, a study evaluated and positioned towards the center has lower ecological validity

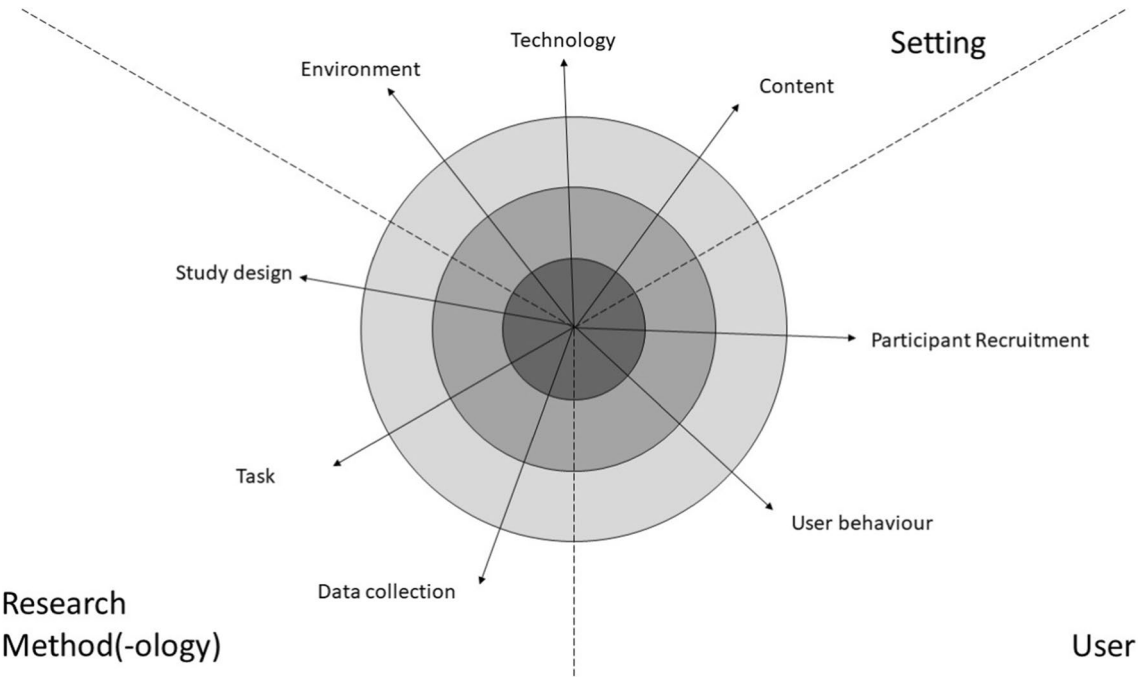

Fig. 3 Model for assessing ecological validity

in comparison to a study that is positioned towards the periphery of the model.

For e.g., in a user study employing low-fidelity prototypes, the *Technology* category within the *Setting* dimension would most likely be plotted towards the center. Similarly, if a study would employ a mid-fidelity prototype such as wire frames, the ecological validity of the technology might be plotted towards the middle of the model, while a high-fidelity prototype or real service would be evaluated as having a high ecological validity and would be plotted towards the periphery. In the same example, the task might be realistic and engaging and evaluated as highly ecologically valid, but the participant sample might consist of younger university students while the actual target group are elderly. As a result, the ecological validity of the participant-dimension would be evaluated as low and plotted towards the center.

In this regard, if all the other categories within the setting dimension are also high, then consequently the overall evaluation of the *Setting* dimension would also be high, so on and so forth the level of ecological validity pertaining to the individual dimensions and different studies can be evaluated in this manner. While the proposed basic visualisation is just a first step towards creating more awareness about the ecological validity of a study as a whole and in terms of specific sub-dimensions, it can be a useful tool to assess a study design.

Conclusion

The overall aim of the study was to obtain an overview of existing guidelines (within the fields of QoE and HCI) for conducting ecologically valid studies, and thereby contribute towards a better understanding of how to design more ecologically valid user studies. Through a systematic review of existing literature, a total of 12 articles were included from an initial 782 retrieved references. The review resulted in the extraction of 55 guidelines which were further grouped into 8 categories and 3 overarching dimensions.

The broad range of categories and the overarching dimensions capture the essential aspects involved in user studies, and the corresponding guidelines to ensure ecological validity. Simultaneously, the review shines light on the flip side of ecological validity and the undesirable outcomes involved in merely aiming for high ecological validity. The perspective validates the argument that *ecological validity is a goal but not a destination* [10], in the sense that one must be mindful of the adverse consequences involved and, thereby decide on the different aspects of designing a study accordingly. Further, we started a discussion on implications for QoE research and outlined a number of topics for future work. Additionally, the review also provides a basic visualisation model as a guide for the design of ecologically valid studies and for evaluating different dimensions of ecological validity.

By focusing on the importance of ecological validity, the intention of this article is not to downplay the importance of lab-based studies or over-emphasize the importance of in-situ studies. Rather, we believe that future research should address more in detail how and when research approaches characterized by high ecological validity (and correspondingly, low internal validity) and those characterized by low ecological validity (and normally high internal validity) can best complement each other in order to better understand the key factors influencing QoE for various types of applications, user segments, settings. Further, we believe that more transparency around the (sub)dimensions of ecological validity with respect to a particular study or set of studies is necessary. Follow-up work is therefore needed to provide a detailed checklist and description of the relevant dimensions and sub-dimensions to consider in a study design.

Appendix 1

The final full query used: ((recommend* OR guid* OR "best practice*" OR lesson* OR "lessons learnt") AND ("Ecological* valid*" OR "external validity") AND ("experiment*" OR "lab setting" OR "controlled environment") AND ("QoE" OR "Quality of Experience" OR "user experience" OR "UX"))

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

Funding Open access funding provided by NTNU Norwegian University of Science and Technology (incl St. Olavs Hospital - Trondheim University Hospital).

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Raake A, Egger S (2014) Quality and quality of experience. In: Quality of experience, pp 11–33. Springer, Berlin
2. Hassenzahl M, Wiklund-Engblom A, Bengs A, Hägglund S, Diefenbach S (2015) Experience-oriented and product-oriented evaluation: psychological need fulfillment, positive affect, and product perception. Int J Human Comput Interaction 31(8):530–544
3. McCarthy J, Wright P (2007) Technology as experience. MIT press, London
4. Reiter U, Brunnström K, De Moor K, Larabi M-C, Pereira M, Pinheiro A, You J, Zgank A (2014) Factors influencing quality of experience. In: Quality of experience, pp 55–72. Springer, Berlin
5. Baraković Husić J, Baraković S, Cero E, Slamnik N, Oćuz M, Dedović A, Zupčić O (2020) Quality of experience for unified communications: a survey. Int J Netw Manage 30(3):2083. https://onlinelibrary.wiley.com/doi/pdf/10.1002/nem.2083. https://doi.org/10.1002/nem.2083. e2083 NEM-18-0124
6. Wechsung I, De Moor K (2014) Quality of experience versus user experience. In: Quality of experience, pp 35–54. Springer, Berlin
7. Viola I, Subramanyam S, Li J, Cesar P (2022) On the impact of vr assessment on the quality of experience of highly realistic digital humans. Quality User Exp 7(1):1–32
8. Hodzic K, Cosovic M, Mrdovic S, Quinlan JJ, Raca D (2022) Realistic video sequences for subjective qoe analysis. arXiv preprint arXiv:2204.06829
9. Robson C, McCartan K (2016) Real world research. Wiley Global Education, Chichester
10. Dole L, Ju W (2019) Face and ecological validity in simulations: lessons from search-and-rescue hri. In: Proceedings of the 2019 CHI conference on human factors in computing systems, pp 1–8
11. Borchert K, Schwind A, Hirth M, Hosfeld T (2019) In vivo or in vitro? Influence of the study design on crowdsourced video QoE. In: 2019 11th International conference on quality of multimedia experience, QoMEX 2019. Institute of Electrical and Electronics Engineers Inc.
12. Masi AD, Wac K (2019) Predicting quality of experience of popular mobile applications from a living lab study. In: 2019 11th International conference on quality of multimedia experience, QoMEX 2019
13. Robitza W, Garcia MN, Raake A (2015) At home in the lab: assessing audiovisual quality of HTTP-based adaptive streaming with an immersive test paradigm. In: 2015 Seventh international workshop on quality of multimedia experience (QoMEX), pp 1–6. IEEE. http://ieeexplore.ieee.org/document/7148122/
14. Pérez P, González-Sosa E, Kachach R, Pereira F, Villegas Á (2021) Ecological validity through gamification: an experiment with a mixed reality escape room. In: 2021 IEEE international conference on artificial intelligence and virtual reality (AIVR), pp 179–183. https://doi.org/10.1109/AIVR52153.2021.00040
15. Dole L, Ju W (2019) Face and ecological validity in simulations:lessons from search-and-rescue hri. In: Proceedings of the 2019 CHI conference on human factors in computing systems. CHI '19, pp 1–8. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3290605.3300681
16. Deniaud C, Mestre D, Honnet V, Jeanne B (2014) The concept of "presence" used as a measure for ecological validity in driving simulators. In: Proceedings of the 2014 European conference on cognitive ergonomics. ECCE '14. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2637248.2637270
17. Vona F, Pieri L, Patti A, Tafaro S, Saccoccio S, Garzotto F, Romano D (2022) Explore 360$_{o}$ vr to improve the ecological validity of screening tests on cognitive functions. In: Proceedings of the 2022 international conference on advanced visual interfaces. AVI 2022. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3531073.3531171
18. Lewkowicz DJ (2001) The concept of ecological validity: What are its limitations and is it bad to be invalid? Infancy 2(4):437–450
19. Campbell DT, Cook TD (1979) Quasi-experimentation. Houghton Mifflin, Boston
20. Möller S, Raake A (2014) Quality of experience: advanced concepts, applications and methods. Springer, Berlin
21. Fiedler M, Möller S, Reichl P, Xie M (2018) QoE Vadisl (Dagstuhl Perspectives Workshop 16472). Dagstuhl Manifestos 7(1):30–51. https://doi.org/10.4230/DagMan.7.1.30
22. De Moor K, Fiedler M, Reichl P, Varela M (2015) Quality of experience: from assessment to application (Dagstuhl Seminar 15022). Dagstuhl Rep 5(1):57–95. https://doi.org/10.4230/DagRep.5.1.57
23. Robitza W, Garcia MN, Raake A (2015) At home in the lab: Assessing audiovisual quality of http-based adaptive streaming with an immersive test paradigm. In: 2015 Seventh international workshop on quality of multimedia experience (QoMEX), pp 1–6. IEEE
24. De Moor K, Ketyko I, Joseph W, Deryckere T, De Marez L, Martens L, Verleye G (2010) Proposed framework for evaluating quality of experience in a mobile, testbed-oriented living lab setting. Mobile Netw Appl 15(3):378–391
25. 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
26. Van den Broeck W, Jacobs A, Staelens N (2012) Integrating the everyday-life context in subjective video quality experiments. In: 2012 Fourth international workshop on quality of multimedia experience, pp 19–24. IEEE
27. Schmidt S, Naderi B, Sabet SS, Zadtootaghaj S, Möller S (2020) Assessing interactive gaming quality of experience using a crowdsourcing approach. In: 2020 Twelfth international conference on quality of multimedia experience (QoMEX), pp 1–6. https://doi.org/10.1109/QoMEX48832.2020.9123122
28. Seufert A, Wamser F, Yarish D, Macdonald H, Hoßfeld T (2021) Qoe models in the wild: Comparing video qoe models using a crowdsourced data set. In: 2021 13th International Conference on Quality of Multimedia Experience (QoMEX), pp. 55–60. https://doi.org/10.1109/QoMEX51781.2021.9465422
29. Hoßfeld T, Keimel C (2014) In: Möller, S., Raake, A. (eds.) Crowdsourcing in QoE evaluation, pp 315–327. Springer, Berlin. https://doi.org/10.1007/978-3-319-02681-7_21
30. Hoßfeld T, Keimel C, Hirth M, Gardlo B, Habigt J, Diepold K, Tran-Gia P (2013) Best practices for qoe crowdtesting: Qoe assessment with crowdsourcing. IEEE Trans Multimedia 16(2):541–558
31. Xiao Y, Watson M (2019) Guidance on conducting a systematic literature review. J Plan Educ Res 39(1):93–112
32. Okoli C (2015) A guide to conducting a standalone systematic literature review. Commun Assoc Inf Syst, 37
33. Fink A (2019) Conducting research literature reviews: from the internet to paper. Sage publications
34. Mäki T, Varela M, Laulajainen J-P (2013) Estimating the effect of context on qoe of audiovisual services: Laboratory vs. public places. In: 2013 International conference on smart communications in network technologies (SaCoNeT), vol 3, pp 1–6. IEEE
35. Lew L, Nguyen T, Messing S, Westwood S (2011) Of course i wouldn't do that in real life: advancing the arguments for increasing realism in hci experiments. In: CHI'11 extended abstracts on human factors in computing systems, pp 419–428
36. Staelens N, De Meulenaere J, Claeys M, Van Wallendael G, Van den Broeck W, De Cock J, Van de Walle R, Demeester P, De Turck

F (2014) Subjective quality assessment of longer duration video sequences delivered over http adaptive streaming to tablet devices. IEEE Trans Broadcast 60(4):707–714
37. Maguire J, Renaud K (2015) Alternative authentication in the wild. In: 2015 Workshop on socio-technical aspects in security and trust, pp 32–39. IEEE
38. Robitza W, Kara PA, Martini MG, Raake A (2016) On the experimental biases in user behavior and qoe assessment in the lab. In: 2016 IEEE globecom workshops (GC Wkshps), pp 1–6. IEEE
39. Robitza W, Raake A (2016) (re-) actions speak louder than words? a novel test method for tracking user behavior in web video services. In: 2016 Eighth international conference on quality of multimedia experience (QoMEX), pp 1–6. IEEE
40. Mottelson A, Hornbæk K (2017) Virtual reality studies outside the laboratory. In: Proceedings of the 23rd Acm symposium on virtual reality software and technology, pp 1–10
41. Blum JR, Fortin PE, Al Taha F, Alirezaee P, Demers M, Weill-Duflos A, Cooperstock JR (2019) Getting your hands dirty outside the lab: a practical primer for conducting wearable vibrotactile haptics research. IEEE Trans Haptics 12(3):232–246
42. Labonté-LeMoyne É, Courtemanche F, Fredette M, Léger P-M (2018) How wild is too wild: Lessons learned and recommendations for ecological validity in physiological computing research. In: PhyCS, pp 123–130
43. van Berkel N, Clarkson MJ, Xiao G, Dursun E, Allam M, Davidson BR, Blandford A (2020) Dimensions of ecological validity for usability evaluations in clinical settings. J Biomed Inf 110:103553
44. Orne MT (2009) Demand characteristics and the concept of quasi-controls. Artifacts in behavioral research: robert Rosenthal and Ralph L. Rosnow's classic books 110:110–137
45. Ansolabehere S, Iyengar S, Simon A, Valentino N (1994) Does attack advertising demobilize the electorate? Am Political Sci Rev 88(4):829–838