Taylor & Francis
Taylor & Francis Group

Behaviour & Information Technology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tbit20

Choosing between remote and face-to-face features in a test setting: methodology used in two usability and user experience case studies

Marja Liinasuo, Timo Kuula, Vladimir Goriachev & Kaj Helin

BEHAVIOUR & INFORMATION TECHNOLOGY
https://doi.org/10.1080/0144929X.2023.2165453

Choosing between remote and face-to-face features in a test setting: methodology used in two usability and user experience case studies

Marja Liinasuo, Timo Kuula, Vladimir Goriachev and Kaj Helin

VTT Technical Research Centre of Finland Ltd, Cognitive Production Industry, Espoo, Finland

ABSTRACT
Remote test settings have become more common due to COVID-19. This paper presents two user tests focusing on the usability and user experience of an augmented reality-related solution. We describe the proceeding of the tests from the perspective of what has been conducted face-to-face and remotely. Thereafter, the appropriateness of the used test methodology is evaluated based on (i) the acquired results, (ii) the ease of using and understanding the methods and (iii) the test atmosphere. The physical presence of a person providing technical support to the test participant proved vital for augmented reality-related testing; the location of other test organisers appears more indifferent. Finally, we present the perspectives to contemplate when choosing between face-to-face and remote features in the test setting.

ARTICLE HISTORY
Received 11 August 2022
Accepted 2 January 2023

KEYWORDS
Augmented reality; testing methodology; usability; user experience

1. Introduction

Scientific testing or evaluation usually takes place in dedicated facilities such as in a laboratory or somewhere else, in some more or less public (opposed to private) locations. Information collection is also being done on other premises. Remote testing is defined as testing in a situation where the test leader or moderator is separated in space and/or time from test participants (following loosely the definition by Andreasen et al. 2007). The idea of conducting usability testing remotely emerged already in the early 1990s (Hammontree, Weiler, and Nayak 1994). The increased number of commonly available software for collaborative activities such as videoconferencing or visual collaboration platforms have nowadays enabled remote testing on a large scale. Recently, COVID-19 has forced scientific testing to be done remotely.

Testing can be conducted in various locations. For instance, remote sensory testing has been successfully carried out at the assessor's home or workplace (Dinnella et al. 2022), instead of a laboratory where it is usually located. This has required the live online supervision of the test leader. Holland et al. (2020) found that people with chronic lung disease conducted an exercise test successfully with supervision at home or supported by remote administration.

Remote testing can set requirements for the test participants. Remote sensory testing proved to be feasible with trained panellists and was suggested to be useful also with consumers (Dinnella et al. 2022). This probably means that the testing procedure in question was not too complicated and could be facilitated also via a videoconference or the like, with a limited view of sight and without the possibility of guiding in a hands-on manner.

Test subjects may be sensitive by nature. Somewhat surprisingly, literature and practice support the usage of telemedicine assessments for patients with cognitive impairment, even if guidance is lacking (Geddes et al. 2020). Probably the need for such testing combined with the availability of appropriate technical devices and the acceptance by patients and caregivers have promoted the concept of remote testing in this context. The main deficiency seems to be the actual testing palette – the specific norms for remote testing must be set and the validity of such tests must be assessed.

Testing may include artefacts, which require professional handling. In testing focusing on sensing, sample (e.g. food) characteristics represent artefacts, which limit the possibilities for remote testing as samples should be handled and shipped without any hazard for participant safety and without biasing effects on the sample itself (Dinnella et al. 2022). In human-robot interaction (HRI) studies, a robot, a complex artefact, is sometimes controlled remotely.

CONTACT Marja Liinasuo marja.liinasuo@vtt.fi VTT Technical Research Centre of Finland Ltd, Cognitive Production Industry, Espoo 02044, Finland

Studying is challenging if physical proximity cannot be enabled (Gittens 2021).

The testing procedure can also be difficult, hampering remote testing. All tests of lung functioning are not appropriate to be conducted at home as not all tests document accurate desaturation with walking (Holland et al. 2020). Consequently, it was recommended in that study that patients at risk of desaturation should be prioritised for centre-based testing when possible.

On the other hand, remote testing can provide benefits not existing during 'normal' testing. For instance, conducting consumer testing remotely, close to the consumer, can be more persuasive for the test participant candidate as there is then no need to travel to testing facilities (e.g. Dinnella et al. 2022).

New guidelines have been called for in remote testing in many domains, such as education (Sando, Medina, and Whalen 2021) or clinical assessment related to cognitive impairment (Geddes et al. 2020). Appropriate guidelines are easier to produce after having gathered enough experience in remote testing. That way challenges are identified and a way to deal with them can be invented. Present recommendations seem to fall in a wide area, starting from appropriate behaviour of all stakeholders (Sando, Medina, and Whalen 2021) to high-level principles such as privacy and autonomy, finding the importance of domain-specific instruments (Geddes et al. 2020).

This paper presents two methodologically similar usability and user experience tests with Microsoft HoloLens 2 mixed reality smart glasses. Microsoft HoloLens 2 is built for interaction with three-dimensional (3D) models, a feature needed in both test applications. The ESA (European Space Agency) test (Helin et al. 2021) and the BIMprove (Improving Building Information Modelling by Realtime Tracing of Construction Processes) test (Liinasuo et al. 2022) were conducted partly remotely and partly as face-to-face. The roles in both tests are similarly located relative to the location of the test participant on test premises (physically present or remote). The actual test results are already published; this paper scrutinises the appropriateness of the user setting, focusing on the appropriateness of physical presence or remoteness of test organisers.

The purpose of the ESA test (Helin et al. 2021) was to conduct a preliminary user review of the first integrated prototype of the developed system. The ESA test was built on two ESA projects, AROGAN (Augmented Reality-based Orbit and ground ApplicatioNs) and VirWAIT (Virtual Workplace for AIT & PA Training and Operations Support). These projects provided an opportunity to develop a mixed-reality system for assembly, integration, testing and verification activities in the space domain (ESA 2022). In the test, the test participant, 'astronaut', carried out the test as if being in space, performing various activities according to the instructions provided by augmented reality. Test participant was asked to think aloud during testing and afterwards, questions were asked regarding the usability and user experience of the augmented reality-based instructions. Also, the SUS questionnaire was used.

We used in ESA testing AROGAN/VirWAIT AR system, which is Unity3D-based software supporting multi-platforms such as iPad and Microsoft HoloLens 2. As a whole, we used

(1) Microsoft HoloLEns 2 with AR player application
(2) mobiPV server for all Operations Data File (ODF) content and 3D models
(3) mobiPV web interface, which also allows the user to interact with the system.

The test procedure was prepared using the system's 'off-site' authoring environment, which used a CAD model of the SMILE system to define AR elements. The user was able to see information in 3D space related to each procedure step, e.g. 3D models and text information.

The motivation for the BIMprove test (Liinasuo et al. 2022) also was to acquire results for developing further the preliminary version of the augmented reality (AR) system. In the EU project BIMprove (BIMprove 2021), we developed a prototype of an augmented reality-based building model to be used in the construction industry by various professionals in the field. This model represented (imagery) a real building but in a digital form so that the user could have a closer look at the features of the real building by viewing or even moving in its' digital counterpart (the so-called digital twin of the real building). In the test, the usability and user experience in moving and manipulating the digital objects in the digital building were studied. Methods for collecting data were the same as in the ESA test (thinking aloud, interview and SUS questionnaire).

The software for the BIMprove test was Unity3D-based software for showing the building information model (BIM) as digital twin data at the construction site. In order to overlay digital twin information on top of the real building, additional pivot objects had to be created. The pivot object location matched the location of the fiducial markers in the real world. That way the digital twin and its real-world counterpart are located in the same coordinate system.

After the background presented in 'Introduction', the methods for testing in the AR context are described in

'Methods'. In 'Appropriateness of research methods' the appropriateness of the methods, especially from the perspectives of remote and physically present testing. In 'Discussion', the results are contemplated as such and from the perspective of scientific literature. Also, perspectives to consider, when choosing between face-to-face and remote test features, are provided. Finally, conclusions are drawn in the last section 'Conclusions'.

2. Methods

We performed testing mainly remotely in two separate projects but similarly. The methods of thinking aloud, observation, questionnaire and interview were used for studying usability and user experience of an AR solution. Tests were managed remotely, using Teams' videoconference (see BIMprove study (Liinasuo et al. 2022) for a detailed description of the used methods). Testing was preliminary by nature in both studies, to remove the most obvious usability flaws before the actual testing with end-users. The main difference between the tests is in the nature of the object of the testing. In the ESA study (Helin et al. 2021), instructions to be presented to astronauts in space, shown in augmented reality, were tested. The instructions on how to proceed with different tasks in space were delivered to the test participant ('astronaut') using AR, so no oral instructions were needed to guide in proceeding with the testing. The user interacted with the system with voice commands and gestures. In the BIMprove study (Liinasuo et al. 2022), the digital twin of an imagery building information model (BIM), shown in AR, was tested, to be used by various professionals in the construction phase. The testing session consisted of separate tasks to conduct with the building model in augmented reality. The user interacted with the system with gestures.

The term 'technical expert', used in the ESA study (Helin et al. 2021), is here replaced with 'test instructor', because it includes both the technical support during the testing session, realised in both studies and the guidance in the testing session and expert involvement in the post-test interview, the two latter activities of BIMprove study only.

2.1. Test participants

There were two test participants in the ESA test and four in the BIMprove test. In ESA, the test participants were ESA experts, and one of them was familiar with the application beforehand. In BIMprove, the test participants were researchers from various fields. One had supported the development of the AR application without seeing it and was highly familiar with the HMD (head-mounted display) device; one was not familiar with the application but highly familiar with the HMD device; and other participants were not familiar with the application nor with the HMD device.

2.2. Test organisers

In addition to the test participants, also other roles were needed in the test. They are coined here as test organisers. The test leader dictated when and how to proceed in the testing session and conducted the test interview. The role of the test instructor was to act as technical support, aiding the test participant in using the AR system and, additionally in the BIMprove test, informing the test participant about the next task. The test assistant wrote down the test participant's verbal expressions during the testing session (thinking aloud) and the responses to interview questions.

2.3. Proceeding with the test

Testing proceeded similarly in both tests. The only difference is in the detail in which test is executed, explained in the list below (in point 4). The test was performed separately for each participant. The whole testing session (points 2–5 in the following list) was audible and visible also to the test organisers not physically present in the test room (the test leader and the test assistant), through a videoconference. In the following, the test is described in more detail (Figure 1).

1. Before the test, we sent a usability questionnaire (System Usability Scale, SUS) to the test participants by e-mail.
2. The actual test session starts. The test leader introduces the agenda, test goals and methods to the test participant in the test room, utilising videoconference. The test assistant is also present in the videoconference but with a muted microphone. The test instructor is physically present with the test participant in the test room.
3. The test leader shows a video about how to use the AR application through a video conference. The test assistant is also present in the video conference with a muted microphone. Being located in the same room as the test participant, also the test instructor hears the proceeding of the instruction but has no role at this point.
4. The test is executed. The test participant expresses verbally his/her intentions, decisions,

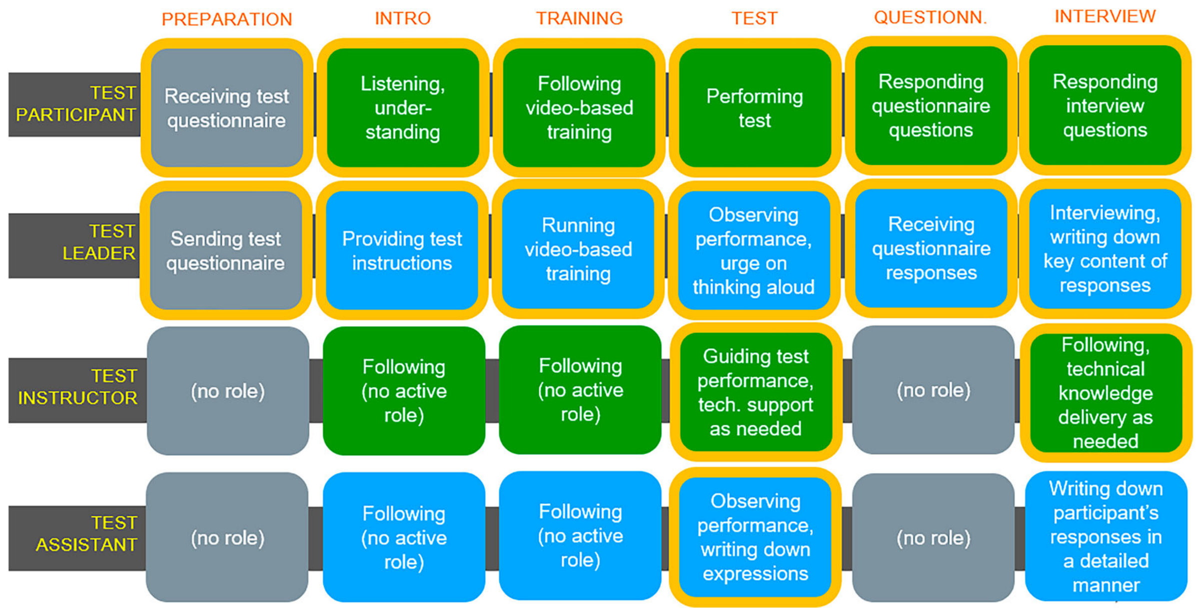

Figure 1. The roles participating in the testing are indicated on the left and the test phase is written upmost in each column. Green means presence in the test facilities, pertaining practically to the test participant and the test instructor only, blue means telepresence (videoconference), valid for the test leader and the test assistant throughout the testing and grey means that the location is indifferent from the perspective of the test methodology, of the test preparation and whenever the role has no task in the phase in question. Key roles in each phase, from the research perspective, are indicated with an outline (yellow).

contemplations and experiences during the testing of the system as instructed by the test leader in point 2 (the think-aloud method). The behaviour of the test participant is mediated by a video conference to the test leader and the test assistant. The task of leading the test is divided between the test leader and the test instructor. The test leader prompts the test participant to think aloud when it seems to be that thinking aloud is forgotten. The test instructor helps the participant in using the AR system when needed. The test assistant observes the test and writes down the test participant's verbal expressions and makes notes about his/her behaviour. Difference between ESA and BIMprove studies:

- In ESA (Helin et al. 2021), the AR system provides instructions on how to proceed with the test.
- In BIMprove (Liinasuo et al. 2022), the test instructor, physically present in the same room with the test participant, informs the test participant about the next task with the AR application, one task at a time.

5. The test participant fills in the usability questionnaire (SUS) after the test.
6. The test leader interviews about the usability and user experience of AR applications using a video conference. The test assistant, also present online, writes the responses down and asks additional questions as needed. The test instructor, hearing the interview, clarifies technical matters related to the interview questions and answers when needed.

3. Appropriateness of research methods

The appropriateness of methods is assessed by

- the number and quality of results, especially when applying a mixed-method approach as in the present paper (e.g. Abowitz and Michael Toole 2010),
- the ease of using (test organiser) the methods (Moody 2003) and, correspondingly, the understanding (test participant) on how to perform in the test and
- test atmosphere, i.e. user experience in the specific context of testing (e.g. Allam, Razak Che Hussin, and Mohamed Dahlan 2013) of the testing event.

The number and quality of results reflect the validity and effectiveness of the methodology. This is the most important criterion for evaluating the appropriateness of the methods used. The ease of using the method is important for test organisers as with easy-to-use methods the results can be pursued without focusing on how to use the chosen methodology. For the test participant, it is important that the method is understandable in such a way that it is easy to learn what kind of performance is expected in the test. Test atmosphere is important to test participants especially when the testing aims to elicit expressions without a predefined format; an awkward situation hardly tempts positive spontaneous expressions. For test organisers, a positive atmosphere is a benefit but not a necessity; instead, test organisers are the ones who can and should affect the test atmosphere positively. Test atmosphere is here included also because it can be assumed that remote testing brings about a different atmosphere than a test in which all parties are physically present.

For the number and quality of results, see the related publications (Helin et al. 2021), (Liinasuo et al. 2022). In both tests, a lot of relevant and interesting results were obtained, supporting the further development of AR solutions. The choice of methods was successful from that perspective. The ease of using the methods is known of test personnel and the understanding of the methods is assumed, based on test participants' behaviour during the test situation. Test atmosphere is evaluated based on the test personnel's, i.e. authors, experiences and the ease with which test participants appeared to express themselves during the test. In the following, method appropriateness is described from the two latter perspectives (ease of use/understanding the methods and atmosphere).

The use of an existing questionnaire (SUS) was easy for the test organisers. The SUS questions are brief and simple and are designed to be responded to without further advice, so the location for filling in has no effect on results or the ease of understanding what to do. Furthermore, the filling in the questionnaire is always done in solitude, so the exact location in which it is performed is not important, as long as the tools are appropriate (e.g. if a laptop is used, there needs to be space and electricity to use it, etc.). The way the questionnaire is transferred to the test participant and after filling in, to the test leader, hardly affects test results either; the sending and receiving material by e-mail is easy. The remote location of receiving, filling in and sending the questionnaire may make the situation appear distant from the perspective of the test participant. However, this process is rapid and hardly affects the testing atmosphere.

The introduction to the test was conducted by a videoconference. Brief instructions were shown in a pre-recorded video, showing what kind of virtual objects there are in augmented reality and how they can be controlled. The test participants could ask questions in a similar way as in a face-to-face event; there is no reason to assume that the understanding of instructions or the test results would have been different if the information were provided in the same location where the test participant was. The use of a video for training is easy for the test leader and it ensures that the instructions are similar and provided with the same care to all test participants. Perhaps the situation would have been more pleasant if all parties having a role in the event had been physically present in the same location, but as the introduction and training took together only about 15 min, it probably did not have any significant effect on the atmosphere.

In the actual test, technical support was provided by a person physically present but the lead of the test, urging the participant to think aloud, contributed to the test through video conferencing. The test assistant also could see the behaviour of the test participant and hear the participant's verbal expressions utilising the video conference. The ease of guiding these tasks was not affected by the difference in the physical location. The experience was a bit less intensive for the remote test organisers, but it did not seem to affect the ease of using the testing methods.

It seemed to be highly necessary to have the test instructor providing technical support next to the test participant; it would have been impossible to provide the same, strong support remotely as was conducted physically present. Support was needed in the gesture-based control of HoloLens. It was hard for an inexperienced test subject to know what kind of gestures to be performed in various situations and what the correct physical gesture looks like in augmented reality. When the test instructor was physically close to the test participant, HoloLens caught the gestures of the test instructor as if the gestures were performed by the test participant. In this situation, the test participant was able to both check what the gesture looks like in the real world and by looking through the head-mounted display, to see how it appears in augmented reality. Thus, the test instructor could help by showing the correct gesture in the specific control situation – what type of gesture, how fast to move, how far to stretch – to perform the required control action. When the test participant cannot proceed with testing due to insufficient knowledge of what to do, the test can be permanently interrupted. Based on this, the physical presence of the test instructor can be claimed as having been vital for this testing. The test participant may also feel the physical presence of the test instructor is more supportive than a remote presence. The task of supporting someone in AR is more demanding than performing it in VR. AR is not shared with other people, so to know the situation of the test participant in augmented reality, the test instructor needs to be able to conclude the situation and perhaps also the location of the test participant by his/her verbal description only.

The repeated request to think aloud, instead, can probably be uttered next to the test participant or remotely without difference in results, ease of use, understanding of what to do and test atmosphere. In a test with AR involved, the test participant is absorbed in the augmented world with holograms (virtual objects) and the urges and questions from the test leader are received from the 'outside world' anyway. The exact physical location of the test leader was probably somewhat indifferent to the test participant during the testing session. The test participants were rich and appeared spontaneous in their verbal expressions, indicating that the atmosphere was pleasant enough, resulting in responses that were sincere and extensive and showed the way to improve the tested applications.

Interviewing remotely was easy and even if the remote setting may add emotional distance between the interviewer and the interviewee, it hardly affected the quality or number of responses. The test leader shared the display with the interview structure through video conference and wrote down the keywords of the test participant's response, showing the written-down responses for the general-level questions also to the test participant. This is how both key parties, the test leader and the test participant, shared the key information of the situation and it was verified that the main content of the response was understood correctly. It probably added the trustworthiness of the situation as experienced by the test participant as in this case, the participant got immediate feedback on how the response was understood and recorded. This kind of sharing is easy to conduct in a video conference but happens rarely in a face-to-face interview as conceived by the authors of this paper. In hindsight, the shared writing of the responses could have been extended to all questions. Thus, the remote location of the test leader was possibly beneficial in this respect.

The test instructor was also needed in the interview. In this study, the test instructor was also the developer of the AR solution, whereas the test leader was a professional human factors expert. This being the case, technical questions could be answered and false assumptions corrected in the interview immediately by the test instructor. This made the interview more discussion-like, probably making the interview more motivating to the test participant and enabling the elicitation of more ideas as test responses.

The task of writing down interview answers, the task of the test assistant in this study, requires constant concentration on hearing and understanding what is being said so the location of the test assistant is indifferent from the perspective of the successfulness of performing that task, as long as the interview answers are clearly audible.

4. Discussion

4.1. Our results and the related literature

In this paper, we contemplate the efficiency of remote testing vs testing with a physical presence, based on the results and experiences of our AR-related studies described in this paper. We classify the testing as remote in these studies as the test lead was physically separated from the test participant. The setting can also be called remote moderated (Giroux et al. 2021) or human-moderated remote user testing (Vasalou et al. 2004) as irrespective of physical separation, the test lead and the test participant communicated directly during the test. However, as the test instructor co-located with the test participant, the setting can also be coined as a hybrid. Next, the effect of the location – remote or face-to-face – is contemplated, method by method.

Questionnaire-related activities – sending/receiving, filling in, sending back – are not location sensitive. The more complex the questionnaire, however, the possibility to intervene may become more relevant, for instance, with filling-in related issues.

When testing the AR solution, the location of the test leader and the test assistant appeared relatively indifferent but technical support was needed, provided by the physically present test instructor, for using the AR system. Obviously, assistance is needed when complex devices are used in testing without thorough training or experience beforehand. Furthermore, the technology-clarifying comments of the test instructor facilitated the interview, ensuring that incorrect assumptions were rectified and unclear matters clarified. This way the interview provided information also to the test participant, making the situation probably more meaningful to him/her as well. Perhaps the presence of the test instructor also provided the needed human closeness. Also, the sharing of the main interview results in a videoconference possibly added to the trustworthiness of the test for the test participant.

All in all, a hybrid test setting was proved appropriate in these studies, key elements being the physical presence of the test instructor providing technical support during the test, the relative indifference of physical location of other testing organisers and the possibility to share main results during the interview through videoconference.

On usability testing, already in 1994 Hammontree, Weiler, and Nayak (1994) suggested technology, which enables remote testing. Since then, supporting means such as protocols (e.g. Vasalou et al. 2004), accessibility (e.g. Oncins 2021) and tools (Giroux et al. 2021; Thompson, Rozanski, and Haake 2004; Chen and

Zhang 2015, April) for remote testing have been contemplated and tested. Due to COVID, remote user testing has become even more common. As methodological tools are being continuously developed and because the benefits of remote testing have become more familiar to researchers, it is possible that remote and hybrid testing will become even more common, also after COVID. In the literature, hybrid testing is referred to in rather complex test settings (e.g. Vasseur et al. 2021a; Reinhorn et al. 2004). Seems to be that the concept is not used in cognitive ergonomics, even if that kind of testing would have been conducted. Perhaps hybrid tests are categorised as belonging to remote testing, such as when performing consumer testing close to the consumer (Dinnella et al. 2022). It can also be that such studies are seldom performed or published.

There is no general advantage of one setting over another (McCrum and William 2016) so researchers need to be able to make informed decisions among the possible options. Based on the literature, the successfulness of different test settings seems to be mainly evaluated by the number and quality of results as well as the time the testing requires (e.g. Andreasen et al. 2007; Abowitz and Michael Toole 2010; Sauer et al. 2019; Johnson, Scheitle, and Ecklund 2019) although some more qualitative perspectives, such as the peace in responding to questionnaire questions outside laboratory (Andreasen et al. 2007), are mentioned. Our paper provides a broader perspective.

Table 1. Tentative perspectives for choosing between a face-to-face and a remote test setting

	TEST LEADER and other test organisers	TEST PARTICIPANT
RESULTS	Are the expected number and/or the quality of results reachable in the face-to-face/remote setting?	N/A
METHODOLOGY USABILITY	Is the method usable when applied in the face-to-face/remote setting?	Is testing procedure understandable so the test can proceed fluently in the face-to-face/remote setting?
METHODOLOGY BENEFITS	Are there result-related benefits if the test is conducted face-to-face/remotely?	Are there participant-related benefits (pleasantness, interestingness …) if the test is conducted face-to-face/remotely?
ARTEFACT (SOFTWARE, HMD, OR OTHER PHYSICAL TESTING ITEM)	Is an artefact needed, to be used by the test leader, and are there limitations when used face-to-face/remotely?	Is an artefact needed, to be used by test participants, and is it highly useful when used face-to-face/remotely?
TEST LOCATION	Is it beneficial to have the test face-to-face/remotely for methodological or result-related purposes?	It is beneficial to have the test face-to-face/remotely for comfort-related reasons?
TEST ATMOSPHERE	Is the test leading comfortable enough to manage face-to-face/remotely?	Does mood remain positive during testing when conducted face-to-face/remotely?

4.2. Relevant perspectives for choosing an appropriate test setting

Our approach can be even broadened to perspectives that assist in deciding the setting for some specific study. Basic assumptions, not handled here, are that the research is planned professionally: it is expected that the research question is meaningful, the research method is valid for its purpose and the test setting is ethical. Our perspectives are partly overlapping but still present important matters to contemplate separately. In the following, the perspectives are briefly presented (Table 1).

The possibility to gain extensive results with good quality is the most important criterion for choosing the face-to-face or remote setting. It is probably hard to evaluate, however, without focusing on factors contributing to reaching the results, explained in the following.

The methodology should be usable. Firstly, from the perspective of test organisers, the method(s) to be used in testing should exist in the intended setting (face-to-face or remote); this cannot be taken for granted (e.g. Geddes et al. 2020; Sando, Medina, and Whalen 2021). Secondly, the method(s) should also be used in the setting in question; for instance, video recording does not always provide as accurate an image as needed. From the perspective of the test participant, the counterpart for the method is the resulting testing procedure; the procedure should be easy to understand and follow so that test can proceed fluently.

It is also possible that some method brings results answering better the research question(s) in some specific setting (face-to-face or remote). This is an aspect relevant to test organisers. From the perspective of the test participant, a positive experience may be stronger with some specific method setting, adding, for example, the pleasantness or interest in the test.

Sometimes software or some other artefact is needed in testing. It can be an artefact, elementary for the test, such as a head-mounted display in a test related to augmented reality or a supporting artefact needed in some specific setting, like video conferencing with good connections is needed as an artefact for interviewing in remote settings. Test supervision can be performed, at least in some cases, equally well both face-to-face and remotely (Dinnella et al. 2022; Holland et al. 2020). Additional artefact(s), if needed, should be available and usable for test organisers. If the test participant should use some artefact (e.g. see Dinnella et al. 2022), it also should be available and highly usable.

Some specific test locations may be preferred for methodological or other purposes (the viewpoint of test organisers) or comfort-related reasons (a matter affecting test participants). For instance, the study of human-robot interaction requires that the test participants are in the same location as the robot, enabling physical proximity (Gittens 2021).

Finally, the perspective of the test atmosphere is included as the positive mood of the test participant has value as such. It is not ethical to cause discomfort to test participants and participants' discomfort may also bias test results (see, e.g. Vasseur et al. 2021b). A positive atmosphere is affected not only by the attitude – respect and positivity – the test leader shows towards the test participant but also, say, by the visibility (facial) feedback provided by the test leader during testing (Andreasen et al. 2007). Thus, the physical presence or absence in the testing event affects the test atmosphere from the test participant's perspective. However, when contemplating the factors affecting the atmosphere for the part of the test participant, not only the test organiser location but also other things matter, such as how interesting or demanding the test participant perceives the test and the used methodology. Furthermore, from the test organiser perspective, test atmosphere, i.e. the level of comfort, may highly depend on the ease of using the test methods as a whole, the physical location being only one aspect of it. Hence, the test atmosphere is not an independent factor but is entangled with other perspectives and the subject of the test as well.

5. Conclusions

The combination of face-to-face and remote test settings proved to be successful for our AR-related studies (Helin et al. 2021; Liinasuo et al. 2022). The importance of having a technical expert physically present and in an active role, when demanding technology is used in testing, was identified. The location of other test organisers, remote or in physical proximity, did not appear as important in our case. Videoconferencing provided valuable means to share the main interview results immediately online.

The choice among different test settings can be supported by contemplating the possibilities from several perspectives. A suggestion for the perspectives is presented in this paper, emphasising the roles of both the test leader and test participant. Cognitive-ergonomics-based literature seems to emphasise only the existence of remote and face-to-face settings. Our results show that a more fine-tuned, hybrid approach can be fruitful when designing a test.

The results would have been more convincing if a comparison test had been conducted, for instance, with a completely remote or complete onsite setting. Also, the appropriateness of the methods should be systematically assessed using feedback from all stakeholders instead of partly concluding it afterwards, as was performed in this study. This kind of study remains to be done in the future.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by Horizon 2020 Framework Programme: [Grant Number grant agreement No 958450].

ORCID

Marja Liinasuo http://orcid.org/0000-0003-0070-1926
Timo Kuula http://orcid.org/0000-0002-5108-7199
Vladimir Goriachev http://orcid.org/0000-0003-3804-5039
Kaj Helin http://orcid.org/0000-0002-5951-0932

References

Abowitz, Deborah A., and T. Michael Toole. 2010. "Mixed Method Research: Fundamental Issues of Design, Validity, and Reliability in Construction Research." *Journal of Construction Engineering and Management* 136 (1): 108–116. doi:10.1061/(ASCE)CO.1943-7862.0000026.

Allam, A. H., Ab Razak Che Hussin, and Halina Mohamed Dahlan. 2013. "User Experience: Challenges and Opportunities." *Journal of Information Systems Research and Innovation* 3 (1): 28–36.

Andreasen, Morten Sieker, Henrik Villemann Nielsen, Simon Ormholt Schrøder, and Jan Stage. 2007. "What Happened to Remote Usability Testing? An Empirical Study of Three Methods." Proceedings of the SIGCHI conference on human factors in computing systems, 1405–1414. San Jose, CA, USA.

BIMprove. 2021. "Website for the EU Project BIMprove." Accessed January 28, 2022. https://cordis.europa.eu/project/id/958450.

Chen, Kevin, and Haoqi Zhang. 2015. "Remote Paper Prototype Testing." Proceedings of the 33rd annual ACM conference on human factors in computing systems, 77–80. Seoul, Korea.

Dinnella, Caterina, Lapo Pierguidi, Sara Spinelli, Monica Borgogno, Tullia GallinaToschi, Stefano Predieri, Giliana Lavezzi, et al. 2022. "Remote Testing: Sensory Test During Covid-19 Pandemic and Beyond." *Food Quality* *and Preference* 96: 104437. https://www.sciencedirect.com/science/article/pii/S0950329321003190. doi:10.1016/j.foodqual.2021.104437.

ESA. 2022. "Website of the European Space Agency." Accessed April 20, 2022. https://www.esa.int.

Geddes, Maiya R., Megan E. O'Connell, John D. Fisk, Serge Gauthier, Richard Camicioli, Zahinoor Ismail, and Alzheimer Society of Canada Task Force on Dementia Care Best Practices for COVID-19. 2020. "Remote Cognitive and Behavioral Assessment: Report of the Alzheimer Society of Canada Task Force on Dementia Care Best Practices for COVID-19." *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 12 (1): e12111. doi:10.1002/dad2.12111.

Giroux, Felix, Pierre-Majorique Léger, David Brieu gne, Francois Courtemanche, Frédérique Bouvier, Shang-Lin Chen, Salima Tazi, et al. 2021. "Guidelines for Collecting Automatic Facial Expression Detection Data Synchronized with a Dynamic Stimulus in Remote Moderated User Tests." International conference on human-computer interaction, 243–254. Virtual event.

Gittens, Curtis L. 2021. "Remote HRI: A Methodology for Maintaining COVID-19 Physical Distancing and Human Interaction Requirements in HRI Studies." *Information Systems Frontiers*, 1–16. https://doi.org/10.1007/s10796-021-10162-4 https://link.springer.com/article/10.1007s10796-021-10162-4.

Hammontree, Monty, Paul Weiler, and Nandini Nayak. 1994. "Remote Usability Testing." *interactions* 1 (3): 21–25. doi:10.1145/182966.182969.

Helin, Kaj, Jaakko Karjalainen, Timo Kuula, Paul Kiernan, Gianluca Casarosa, and David Martinez Oliveira. 2021. "The Remote User Test of Mixed Reality System for AIT/AIV in the Space Domain." In *In EuroXR 2021: Proceedings of the Virtual EuroXR Conference*, edited by K. Helin, J. Perret, and S. Arlati, 11–13. VTT Technical Research Centre of Finland Ltd. https://sarjaweb.vtt.fi/pdf/technology/2021/T395.pdf

Holland, Anne E., Carla Malaguti, Mariana Hoffman, Aroub Lahham, Angela T. Burge, Leona Dowman, Anthony K. May, et al. 2020. "Home-based or Remote Exercise Testing in Chronic Respiratory Disease, During the COVID-19 Pandemic and Beyond: A Rapid Review." *Chronic Respiratory Disease* 17: 1–18. doi:10.1177/1479973120952418.

Johnson, David R., Christopher P. Scheitle, and Elaine Howard Ecklund. 2019. "Beyond the In-Person Interview? How Interview Quality Varies Across In-Person, Telephone, and Skype Interviews." *Social Science Computer Review* 39 (6): 1142–1158. doi:10.1177/0894439319893612.

Liinasuo, Marja, Timo Kuula, Kaj Helin, Vladimir Goriachev, and Matthias Aust. 2022. "Building's Digital Twin Model Shown as First-person and Third-person Views." CIB WBC2022 conference on 27–30 June 2022, Melbourne, Australia.

McCrum, D. P., and M. S. William. 2016. "An Overview of Seismic Hybrid Testing of Engineering Structures." *Engineering Structures* 118: 240–261. doi:10.1016/j.engstruct.2016.03.039.

Moody, Daniel L. 2003. "The Method Evaluation Model: A Theoretical Model for Validating Information Systems Design Methods." ECIS 2003 proceedings, 79. http://aisel.aisnet.org/ecis2003/79

Oncins, Estella. 2021. "Accessibility in Online User Testing." *Journal of Audiovisual Translation* 4 (2): 6–22. https://doi.org/10.47476/jat.v4i2.2021.176.

Reinhorn, Andrei M., Mettupalayam V. Sivaselvan, Zach Liand, and Xiaoyun Shao. 2004. "Real-time Dynamic Hybrid Testing of Structural Systems." 13th World conference on earthquake engineering, Vancouver, B.C., Canada, August 1–6, 2004, paper No. 1644.

Sando, Karen, Melissa S. Medina, and Karen Whalen. 2021. "The Need for New Guidelines and Training for Remote/Online Testing and Proctoring Due to COVID-19." *American Journal of Pharmaceutical Education. Commentary*. https://www.ajpe.org/content/ajpe/early/2021/01/28/ajpe8545.full.pdf.

Sauer, Juergen, Andreas Sonderegger, Klau Heyden, Jasmin Biller, Julia Klotz, and Andreas Uebelbacher. 2019. "Extra-laboratorial Usability Tests: An Empirical Comparison of Remote and Classical Field Testing with lab Testing." *Applied Ergonomics* 74: 85–96. doi:10.1016/j.apergo.2018.08.011.

Thompson, Kathrine E., Evelyn P. Rozanski, and Anne R. Haake. 2004. "Here, There, Anywhere: Remote Usability Testing that Works." Proceedings of the 5th conference on information technology education, 132–137. Salt Lake City UT USA.

Vasalou, Asimina, Brian David Ng, Peter Wiemer-Hastings, and Lidia Oshlyansky. 2004. "Human-moderated Remote User Testing: Protocols and Applications." 8th ERCIM workshop, user interfaces for all, Wien, Austria, Vol. 19.

Vasseur, Aurélie, Pierrre-Majorique Léger, Francois Courtemanche, Vanessa Elise Labonta-Lemoyne, Audrey Valiquette Geroges, David Brieugne, Emma Rucco, Constantinos Coursaris, Marc Fredette, and Sylvain Sénécal. 2021a. "Collection for UX Evaluation: A Pilot Project." International conference on human-computer interaction, 255–267. Virtual Event.

Vasseur, Aurélie, Pierre-Majorique Léger, Francois Courtemanche, Elise Labonte-Lemoyne, Vanessa Georges, Audrey Valiquette, David Brieugne, et al. 2021b. "Distributed Remote Psychophysiological Data Collection for UX Evaluation: A Pilot Project." International conference on human-computer interaction, 255–267. Virtual Event.