Remote Evaluation in Universal Design Using Video Conferencing Systems During the COVID-19 Pandemic⋆

Joschua Thomas Simon-Liedtke[1][0000−0003−4809−1688], Way Kiat Bong[2][0000−0003−3714−123X], Trenton Schulz[1][0000−0001−6217−758X], and Kristin Skeide Fuglerud[1][0000−0002−5648−0264]

[1] Norwegian Computing Center, P.O. Box 114 Blindern, 0314 Oslo, Norway
{joschua,kristins,trenton}@nr.no
https://www.nr.no/en/
[2] University of Oslo, P.O. Box 1072 Blindern, 0316 Oslo
waykb@uio.no

Abstract. Usability and accessibility evaluations with diverse users are an essential part of an iterative universal design process for digital solutions. The COVID-19 pandemic has made it difficult to run traditional local evaluations due to social distancing restrictions to reduce infections. Remote synchronous and asynchronous evaluation methods may be a solution if it can be used by various user groups, including people with impairments. Incorporating video conferencing systems into traditional remote evaluations can be a valuable supplement to — or, given the current situation, alternative to — traditional usability and accessibility evaluations. We present a protocol for remote formative usability evaluations designed with accessibility in mind. The protocol explains how to prepare for an evaluation and its technical setup, how to conduct synchronous evaluations using video conferencing systems, and how to debrief and analyze the collected data. We tested this protocol in a pilot study for an ongoing project where we ran both synchronous and asynchronous remote evaluations. In our pilot, the synchronous evaluation using video conferencing systems provided opportunities for richer qualitative data than the asynchronous evaluation as we ran it. The findings from the pilot study indicate that the protocol is feasible and can be used when having participants with diverse abilities or impairments. We also provide suggestions for others that wish to adapt the protocol.

Keywords: Universal design · Usability · Accessibility · Remote evaluation · Video conferencing · Pandemic · COVID-19

1   Introduction

Universal design (UD) is a design process that makes products and environments usable and accessible to the greatest extent possible by all members of society regardless of level of ability [23]. One of UD's main goals is to promote equality and ensure full participation in society for individuals with impairments [23]. UD has been shown to give improved products and services, increased market and customer satisfaction, enhanced community relations and reputation, improved internal processes, increased financial effects, and avoid legal costs and damages [12]. Studies have shown that cross-disciplinarity and participation from user organizations and diverse user groups in the planning, implementation and evaluation stage of the development process is important for achieving a universally designed solution, i.e. a solution that is usable and accessible to as many people as possible [2,3,10,11,13,14,23,26,27].

The COVID-19 pandemic has made the process of involving users with impairments more challenging due to obstacles and challenges related to, for example, health concerns, lockdown measures, and travel restrictions. At the same time, the pandemic has been advancing digitization of the society, and its workflows that may offer new opportunities for people with impairments. Remote evaluations as discussed in several studies [19,27,28] may be a solution for evaluating systems during this challenging time. More people have had to start using a video conferencing solution such as Microsoft Teams, Zoom, or Google Meet on a regular basis. Several of these solutions promote their accessibility features. This could mean that remote evaluation using a video conferencing solution could offer an even more convenient and accessible evaluation setting for people with impairments. Much of the research within remote evaluation, however, was done before these current solutions were available and in common use.

We have investigated possibilities, challenges, and limitations of video conferencing systems in the remote evaluation of usability and accessibility of web applications and present our findings in this paper. First, we review related literature concerned with remote usability and accessibility evaluation. Second, we propose a protocol for using video conferencing systems in remote usability and accessibility evaluation. Third, we implement the protocol in a remote pilot study assessing a web application, CAPABLE, where we tested our methodology with users from different user groups. Fourth, we discuss the observations from the pilot study and its implications. Finally, we conclude with recommendations for running remote usability and accessibility evaluations and our future works.

2   Related work

A guide for selecting an appropriate strategy for evaluation of design artifacts are outlined by Venable et al. including iterative technical and formative evaluations [37]. One way to include people with impairments in the UD process is by conducting *formative* and *summative* evaluations involving participants with impairments at various stages of the design process. The goal of a formative

evaluation is to determine which aspects of the design work well (or not), why, and to propose improvements to the design. The goal of a summative evaluation is to give an overall assessment of a product or service [9].

Both kinds of evaluations reveal barriers, needs, and bottlenecks for different user groups and increase the understanding of these issues by the developers [27,28]. The evaluations are often performed as *local (user) evaluations* where the participant and the evaluator are at the same location, e.g. in a usability lab. These approaches can be resource intensive since they require a dedicated location, an appointment, and can represent an additional burden for people with impairments needing to travel to an unfamiliar location [28]. Some of these burdens could be reduced by having the evaluation at the participant's home or another familiar location of their liking in a *field study* [28]. Here, the evaluators do the travelling, and equipment may need to be transported for the evaluation process. This reduces some of the control of a lab environment with standard equipment and software but expands the diversity of people that can participate.

Consequently, a *remote evaluation*, where the participant and evaluator are not in the same location, may provide a convenient and efficient alternative for both parties [28]. One study with normal-sighted participants found that participants often perceive remote evaluations as more comfortable and convenient than travelling to a local evaluation at a usability lab [6]. At the same time, this study found that participants had difficulties concentrating on the task at hand, and that the collected issue descriptions were less rich. These observations have been confirmed by other studies that showed that remote user evaluation of usability and accessibility give similar quantitative results as a local evaluation, but that qualitative data was less rich. One reason for the lack of qualitative richness was due to participants not recording observed issues thoroughly enough, and the inability for evaluators to observe the participants and ask question where needed [6,7,27].

Remote evaluation methods can be further split into methods where the evaluator and the participant are separated by space alone, i.e., *synchronous*, or by space and time, i.e., *asynchronous* [8]. Studies [6,7,27] showed that synchronous remote evaluation give similar quantitative results as a local evaluation, but the qualitative data is less descriptive for asynchronous evaluations. One study [27] suggested that newer video conferencing systems could increase the richness of the qualitative data compared to asynchronous remote evaluation methods. Another study with blind and visually impaired users found that most of the participants preferred a local test to a synchronous remote test [21]. The main reason for this preference was that the preparations and setup of the remote synchronous test were too demanding. The tests were conducted on an internet phone program, and the authors noted that a web-based application might have been less technically demanding and might have required less setup.

People around the world have met increasing challenges during the COVID-19 pandemic related to health concerns, local lockdown measures, or travel restrictions. These challenges create demands and needs to running evaluations in a remote environment. Some people with impairments may be more vulnerable

in a pandemic. For example, a study examining the effects of COVID-19 on people with visual impairment reported that nearly 37% of the 937 respondents have underlying health problems that make them more vulnerable to COVID-19 [30]. Comorbidities such as diabetes I/II, asthma, heart disease, chronic obstructive pulmonary disease (COPD), immunosuppression, arthritis, cancer, and kidney dysfunction in this sample of visual impaired people exacerbated their risk of getting a severe case of COVID-19. Thus, local or field evaluations can be a potential source for infection under the on-going global COVID-19 pandemic. Besides the issues mentioned previously, it may be difficult for people with vision impairments to keep a safe, comfortable distance when they cannot see where other people are. Consequently, safety could be added as an additional benefit of remote evaluation.

Moreover, lockdowns, travel restrictions and other governmental measures for mitigating the pandemic have encouraged people to stay more at home and not have visitors. This has made it difficult for people, with or without impairments, to (1) be willing to travel to a usability lab to participate in an evaluation or (2) have an evaluator visit them for a field study at home.

In contrast, the advancement of the digitization of the workplace and the society has offered new opportunities for remote evaluation. Many types of work have switched to digital forms like home schooling or home office [29] offering new opportunities also in the field of remote evaluation. Similarly, many other tasks that before were carried out in person have moved to digital video conferencing systems such as Microsoft Teams, Zoom and Google Meet during the COVID-19 pandemic [29]. This means that many people, including people with impairments, that had previously not used video conferences have gained experience with these collaboration tools. This and the fact that manufacturers of video conferencing systems claim to have accessibility features [22,40] might open up new opportunities for remote evaluation in UD.

Although video conferencing systems provide promising opportunities for remote user evaluation, we are not aware of any studies examining how these current systems can be beneficial for the remote evaluation of usability and accessibility involving people with impairments. Similarly, there is a need to study how these evaluations compare to more traditional asynchronous remote approaches.

3   Protocol for remote evaluation of usability and accessibility

We propose a protocol for using video conferencing systems for remotely evaluating the usability and accessibility of a web-based application targeting desktop and mobile devices. The starting point for the protocol is in the Human Risk & Effectiveness evaluation strategy [37] that emphasizes iterative technical and formative evaluations early in the process before progressing to a more rigorous summative evaluation to study the effectiveness, utility, and benefit of the artifact. Specifically, a Human Risk & Effectiveness strategy is recommended when

one of the major risks of a design is social or user oriented, and it is a goal to rigorously establish that benefit will continue in real situations and over the long run [37]. This strategy includes iterative testing, starting with formative evaluation in an increasingly more naturalistic setting. The protocol presented here is informed by existing research on asynchronous and synchronous formative evaluation methods [6,7,27,28], but revised for the use video conferencing systems like Microsoft Teams and Zoom, which were not included in previous studies on remote usability and accessibility evaluations [6,27].

3.1 Preparations

Before conducting usability and accessibility testing including people with impairments, it is advisable to remove as many accessibility barriers as possible by ensuring conformance to accessibility guidelines, such as W3C WCAG 2.0 or 2.1 [31,38,39]. This is important to avoid wasting participants' and evaluators' time. Besides, research indicates that conformance to these guidelines will improve usability for users with and without impairments [33,34].

The next step is to decide on the goal of the remote evaluation. What should be found out during the evaluation, how should it be examined, and who (broadly) will be participating? This usually results in several tasks for the potential participants to do.

While defining the tasks, one should also look at the different video conferencing solutions that are available and decide what will be used to make sure that the solutions can support our study. Some options to consider can be the accessibility of the system (both with assistive technology and for potential participants to have access to it); the capabilities of the system (e.g., can it record video? audio? share displays?); and how widely used is the system among the target group that will be tested? These considerations are most important when doing a synchronous evaluation. An asynchronous evaluation will normally have less requirements. One should check information from the video system provider and consult with experts in the impairment community to make sure that participants with a specific impairment can use the system. Additionally, it may be necessary to offer more than one system option to reach all participants. When offering multiple options, one should investigate that there are equivalent functions in all system options to get the data that is needed, and that participants can use at least one of the system options.

Another item to consider under preparations is how the participant can give informed consent and how this can be recorded. At the same time, one should consider if and how to distribute compensation to the participants. It is common to compensate participants in usability studies with money or an item, such as a gift card, to acknowledge their effort and time participating in usability studies [28,36]. For research, the method and value of compensation should be appropriate for the subject population and the research activities. Since the evaluation happens remotely, it may be necessary to investigate alternative methods for compensation such as gift cards that can be delivered digitally, using a mobile payment system, or sending something in the mail. A digital delivery can help

with an instant gratification for participation, but it is not strictly necessary. Regardless of what is chosen, it is important that compensation is accessible for the participant. Different places and studies will have different requirements. So, ensure that your evaluation's set-up can satisfy them.

Although obvious, it bears repeating that all the documents that are eventually sent to the participants need to be accessible. As a minimum requirement, the materials should conform to accessibility guidelines such as W3C WCAG. Additionally, one can check that the documents work with different types of assistive devices like screen readers, etc.

3.2   Recruiting participants

As with other evaluations, one must consider how recruitment should be done. User organizations have traditionally been a good way to get in contact with people with impairments. When contacting the user organization, offer the participants two options including synchronous remote evaluation using video conferencing solutions or asynchronous where participants can do it by themselves when they want. Then, the organization may be able to contact additional chapters and talk to people that might not normally have been a candidate earlier, e.g. because of geographical or mobility barriers.

Once participants are recruited, the next steps depend on whether one is planning on running an asynchronous or synchronous evaluation. For asynchronous evaluations, participants receive two documents. The first document includes information about the evaluation, an informed consent form, and information about how to perform the evaluation. The second document provides all the tasks the participant should attempt to complete and questions to answer. If there are specific kinds of feedback that the asynchronous participant should give, be sure to document how it should be given. One should also schedule a debriefing interview when the participant has finished the activities.

Participants for a synchronous evaluation need to be contacted ahead of time to find an adequate time slot for a video conference. Part of the scheduling includes finding out if a participant is familiar with any of the predetermined video conferencing systems or if additional guidance is necessary. One can, for example, offer to send instructions if necessary. It is advisable to choose the option with which the participant is most familiar, especially if they are using an assistive device. The reason for letting participants choose is to consider ease of use and competency, so any lack of skill in using the video conferencing system would interfere as little as possible with the evaluation task at hand. Participants should focus on problems related to the web application and not struggle with unfamiliar video conferencing systems. During this time the information about the evaluation and informed consent form can be sent to them to read and sign.

3.3   Running synchronous evaluations

The participants doing an asynchronous evaluation do the tasks on their own and have a debriefing later. This section focuses on a synchronous evaluation, although the debriefing for an asynchronous evaluation may have similar steps.

Before starting the session with the participant, make sure that the audio and video are working on your end of the video system. This will make it easier to isolate issues if things work on your end. Once the session has started, welcome the participant, and make sure that the audio and video works for the participant as well.

If everything is working, the next step is to start with a short briefing session in which one gives a short summary of the study and its goals, the informed consent form, and the participant's rights. It is important to check if the participant has given consent as a reply to the initial e-mail. If they have yet to give consent, give the participant time to read the informed consent form, offer to read it aloud to them, or provide an acceptable summary. Moreover, the participants should be given time to close any other applications, such as e-mail before sharing their screens etc., to protect the participant's privacy. That way, they are not distracted during the evaluation as well [21]. It bears repeating that it is important that consent is given before asking the participant to share their screen or any recording takes place.

After consent is given, screen sharing recording is started, it is simply a matter of going through the tasks and collecting the data. For a synchronous remote evaluation, this is not much different than a local evaluation.

3.4   Debriefing the participant

An asynchronous or synchronous evaluation should end with a debriefing where the evaluator thanks the participant and the participant can give any final overall impressions. This may also provide an opportunity for collecting additional qualitative information to unclear issues found, especially for asynchronous evaluations. This may also be one possible point to provide compensation if it was not provided at the start of the session.

At this point, the evaluation session is over. Analyzing the data depends on the data collected and the goals that were set at the beginning of the evaluation.

4   Pilot study implementing the protocol

To see if our protocol is feasible, we used the protocol to pilot an evaluation of a web application that is part of the ongoing CAPABLEproject [17,18]. The aim of the CAPABLEproject is to create a universally designed digital tool that empowers citizens in Norway to actively use their clinical and personal health information to manage their health. The target users are all citizens, including elderly people above 65 years old, citizens who have an impairment, using more than three prescribed medications, or having potential nutritional risks. As one

of the earliest iterations in our evaluation strategy for CAPABLE, we conducted a pilot study using our proposed protocol. Feedback from this pilot study is used to adjust the protocol. At the same time, we use discovered usability and accessibility issues in CAPABLEto improve the artifact for a later feasibility study involving more users. The current remote testing is naturalistic in the sense that it is conducted in a place selected by the users and on their own devices, but it does not include real data and a real organizational setting.

In addition to developing the protocol, we are interested in comparing asynchronous and synchronous evaluations, and investigate the benefits and challenges of including video conferencing to the synchronous remote evaluation. We therefore use our pilot study to prepare investigations of the following questions: (1) What advantages and disadvantages do the participant and evaluator experience with synchronous evaluation with video conferencing compared to asynchronous evaluation? (2) Do both approaches give similar quantitative or qualitative results? (3) Does synchronous remote evaluation with video conferencing improve the subjective qualitative experience compared to traditional methods like asynchronous methods? (4) Finally, are our protocol guidelines enough for the smooth implementation of remote evaluation using video conferencing systems?

4.1   Study preparation

We identified the central functionality of the CAPABLEweb application and defined tasks around the functionality that the participants would later be asked to complete. Typical tasks were logging into the application, retrieving information, registering, and editing data, etc.

We investigated using several video conferencing systems including Microsoft Teams, Zoom, Google Meet, TeamViewer, and Skype. After trying out all of these with a focus on accessibility and relevant functionality (e.g. screen sharing and video recording), we eventually settled on Microsoft Teams and Zoom as they seemed to have the biggest reach for the groups from which we were recruiting, and with whom we could easily schedule meetings. We also wrote detailed instructions for how to use these systems for our study and ensured that the documents that would eventually be sent to participants were accessible, also for participants using assistive technology. We, for example, made all documents comply to the WCAG 2.1 standard, and tested the readability of the documents for screen readers.

For compensation, we devised a method using a mobile payment service. Initially, we were targeting digital gift cards options. However, it turned out to be difficult to find digital gift card suppliers where the distribution, information and use of the gift card was sufficiently accessible and universally designed. Moreover, we considered issues such as the accessibility of the information about where to use the gift card, how to get information about the value on the gift card, and whether it had sufficient usage possibilities, both geographically and digitally. Instead of using gift cards, we chose to work out a payment routine together with the administration of our research organization using a mobile payment service

called Vipps that has broad usage in Norway and has won an award for UD [25]. This routine considered privacy, tax rules, and the accessibility and flexibility of payment method to suit the diverse needs of our potential participants. Participants in the pilot were not compensated as they were participating in the pilot as part of their regular jobs. For the final remote usability and accessibility study, participants will receive 300 NOK (around €30) for their effort.

4.2   Recruitment for the pilot study

Participants were recruited through three user organizations: Norwegian Association for the Blind and Partially Sighted (NABP), rheumatics (NRF), and people with heart and lung diseases (LHL). These organizations are partners in the CAPABLEproject and have earlier been involved in needs and requirements elicitation [17,18]. The participants recruited in this pilot study were staff persons working in the user organizations. They are user representatives at the system level, with knowledge of the needs of the group they represent. They were divided into a synchronous and an asynchronous reference group.

Participants were asked about their preferred video conferencing systems. For synchronous evaluations, we also provided a document with technical details about how to start the chosen video conferencing system including screen sharing if needed. With the provided information, we then set up a meeting on the preferred platform and sent the participants additional information about the study and an informed consent form. Participants for the asynchronous evaluation also received additional documents as detailed below.

At the time of writing this article, we have recruited four participants for the pilot. Three participants had the synchronous remote evaluation and one performed the asynchronous evaluation. One of the synchronous participants was a proficient screen reader user. All participants reported above average experience with ICT.

4.3   Running the synchronous evaluation

After the briefing session described in the protocol above, we verified that the participant's informed consent had been delivered beforehand or provided the necessary information to obtain informed consent in the briefing. We asked for consent to the study in general, sharing the display, and recording the meeting. Only the first two were necessary for us to proceed. The video recording was not necessary but was meant to aid our analysis after the session. After receiving consent from the participants, we asked them to activate screen sharing, and started recording the meeting if consent was given.

Finally, we explained the think-aloud protocol [4] and how to identify usability and accessibility issues. As suggested by Nielsen [24], we provided a video example for a think aloud session if participants were unfamiliar with the technique.

We started to conduct the evaluation by giving the participant the tasks using CAPABLE. The screen sharing with the video conferencing systems made

it possible to observe what the user was doing on-screen, at the same time as we could see their facial impressions through the camera when the camera was activated.

To encourage the participant to talk and minimize influence over the participants' focus and workflow, we used acknowledgment tokens [4] rather than asking too many questions while they were completing the tasks. At the end of each task, we asked the participants about their experience, and investigated more details of problems or challenges they might have encountered during the task. Additionally, we inquired participants about their opinions about pitfalls, challenges, or possible improvements especially in those cases where the participant was a specialist in using assistive devices. We also asked them about their perceived difficulty of the task using the Single Ease Question (SEQ) [32].

4.4   Running the asynchronous evaluation

After agreeing to participate, the participants in asynchronous evaluation were provided two documents. The first document described how they should perform the evaluations, from providing us their informed consent to answering the questions regarding to testing tasks and arranging for the debriefing session. We provided instructions and guidance, specifically about what kind of feedback they should provide in the answers.

In the second document, all the testing tasks and questions that the participants were required to answer were presented. These questions were the same as in the think-aloud protocol in the synchronous evaluation. The aim was to have the qualitative data collected in asynchronous evaluation as comparable as possible to those in synchronous evaluation. Following the testing tasks and questions, we had questionnaires about the CAPABLEprototype and the evaluation process.

We ensured that these documents were accessible and universally designed, considering our goal to include diverse users, including participants using assistive technologies in our remote evaluations.

4.5   Debriefing

Finally, we conducted a short debriefing session with synchronous and asynchronous participants. For the synchronous participants, this session followed right after completing the tasks, while we set up a separate meeting through video conferencing or telephone for the asynchronous participant. During the debriefing session we collected additional data by interviewing participants and with the help of questionnaires. Here, we asked about the participant's experience with the artifact and the evaluation process with respect to comfort, ease, memory, concentration, convenience, preference, etc. [6,27]. We also included a SUS questionnaire [5] in Norwegian, and a questionnaire with opinions about the evaluation process. Finally, we asked the participants about their demographic

background information related to age and gender, ICT experience, and technical details, such as types and version of assistive device if used, operative system and browser.

4.6   Analysis

The analysis for each participant was based on notes we made of the observations during the sessions, and on the video recording in those cases the participants consented to it. For each participant, two researchers extracted information to account for the evaluator effect [16]. The relevant problems were summarized in a list with description of the identified usability and accessibility issue. Each issue was furthermore categorized as either cosmetic, minor, or critical [1].

We also summarized and evaluated issues related to the evaluation protocol, such as the order and number of tasks, the design of the tasks, and questions for the final evaluation. Although the time used to conduct the evaluation varied between participants, we found that the number of tasks and questions was quite realistic to conduct within the allocated time, which we had set to an hour and a half. Since we found that we needed around 15 minutes for the debriefing, we stopped when there was 15 minutes left of the agreed time, even if the participant had not completed all the tasks yet. Therefore, it was also important to put the most important tasks in the beginning of the evaluation. In earlier versions of the protocol, we also recorded the time for each participant. However, since the time difference between participants with and without assistive devices might vary, at the same time as different participants talk sometimes more or less during evaluation, depending on their personality and/or interaction with the evaluators, time is not included as a measure for the analysis.

5   Results and discussion

We completed several synchronous and one asynchronous evaluation using our protocol. Participants uncovered several usability and accessibility issues. We do not report on these issues since the CAPABLEweb application is in an early stage of development and this was a pilot study. In the following we focus on our experiences with the remote evaluation protocol.

All but one participant completed all the tasks in the time allocated (ninety minutes). In the one case were the participant did not finish, we truncated the remaining tasks so that we had time to complete debriefing interview instead. We observed that using assistive devices like screen readers takes more time, which agreed with our experience from previous evaluations. The asynchronous participant reported significantly less usability and accessibility issues than any of the synchronous participants.

In the following paragraphs, we discuss general impressions of the evaluation procedure including flexibility and social aspects, technical preferences and challenges, richness of the obtained data, and the usability and accessibility of the evaluation protocol. We compare both the synchronous with the asynchronous evaluation, and the fact that we are using video conferencing systems.

5.1   Flexibility in performing the evaluation

We observed that participants prefer different evaluation types in terms of synchronous or asynchronous evaluation, and different video conferencing systems.

On the one hand, it was faster to recruit participants for the synchronous evaluation than for the asynchronous evaluation in the pilot study. One reason could be the social aspect of the synchronous evaluation, and the familiarity with the video conferencing systems during the COVID-19 pandemic. All our participants were working from home due to recommendations and regulations from their employer and the local and national government. Thus, meetings on Microsoft Teams and Zoom were familiar for all participants. Besides, participants might have considered the synchronous evaluation easier to perform since the process did not require them to read instructions or write elaborate answers. Another reason could have been that participants recruited for the pilot were doing the evaluation as part of their job during work hours, and it might have been easier to work a synchronous evaluation in as part of the workday.

On the other hand, the flexibility in terms of time provided to participants to complete the evaluation by themselves was one of the assumed advantages for the asynchronous remote evaluation. Hartson et al. [15] pointed out that asynchronous participants can decide where, when, and how they want to do the evaluation. During the recruitment, two participants reported that they had other tasks during the day besides their regular work to do. The asynchronous evaluation was likely more flexible for them since they could perform the evaluation whenever they wanted. Asynchronous evaluation may also have appealed to participants who feel uncomfortable talking to or being observed by a stranger, have social or performance anxiety, etc. Providing the option for an asynchronous evaluation can remove pressure in the social setting and remove competitive aspects in future evaluations.

Moreover, participants preferred different video conferencing systems in the synchronous evaluations. Some participants, for example, chose Zoom and some Microsoft Teams. Likely because of having the option to choose, we did not observe any significant usability or accessibility issues related to participants use of the video conferencing systems. All the participants successfully communicated over audio and video and shared their screen with the evaluator.

One advantage of the proposed protocol is the possibility to reach a wider spectrum of participants. One important aspect in UD is the inclusion of diverse user groups, and remote evaluation can contribute in achieving that. Diversity in user groups means the users have different abilities and/or impairments, socio-demographic background, preferences, etc. Using remote evaluations, participants in a bigger geographical area can be reached, and diverse user groups can be included [35]. Using video conferencing systems can make these remote evaluations feel more natural as the participants can see the evaluator and vice versa. Our investigated systems allowed screen sharing, such that we could easily follow the participants on the screen. When a usability or accessibility issue was detected, we could easily identify them without having to rely on descriptions by the participant alone.

However, it is crucial to keep in mind the downsides of conducting remote evaluations using video conferencing systems, i.e. the challenges that might be faced by participants who have low or no ICT skills in using these systems. Our participants were quite skilled in their usage of video conferencing systems but many people may not have the experience in using video conferencing systems as they do not have access to or need for them in their everyday life. We, therefore, suggest to provide detailed instructions on how to use the video conferencing systems to assist those who need them, and to supplement with local evaluations when possible.

5.2 System preferences

All the participants chose to use video conferencing on a laptop or desktop PC (hereafter simply "desktop system") instead of a mobile phone or tablet (hereafter simply "mobile system"). This choice was made even though phone and tablet were listed before PC as possible devices during recruitment.

There may have been several reasons for the participants in the pilot preferring to use a desktop system. One could be that the people recruited for the pilot where responding to a meeting during working hours and used the system they normally use at that time. In addition, while most phones now carry cameras and microphones, many laptops do as well, and participants may have had the video conference set-up already to work on their desktop systems.

Another reason could be the amount of time set off to do the study. We set up an appointment that could last possibly an hour and a half. People may have been more comfortable doing this length of a meeting sitting at a desktop system rather than having to hold a device during that time. In addition, using a desktop system also frees up both hands to do things with the interface or better use an assistive technology.

Using video conferencing together with a website may have imposed additional barriers for the participants. That is, the participants might have been aware that they would not only have to complete the tasks, but also set up the video connection to the evaluators. This might have been more demanding on a mobile screen compared to the space available on a desktop screen. Screen sharing in the mobile app versions of these systems includes additional challenges such as not being able to use the camera or having to move the video window to work with the rest of the user interface.

Regardless, further research should examine if this was purely by chance or deliberate, and differences in running a remote evaluation with video conferencing software on mobile versus desktop systems. If there is a preference for one system over another, it may be necessary to find ways to have people choose a particular solution to make sure that different systems are sufficiently covered.

The fact that participants likely use their own equipment does provide the advantage that they are also likely using their preferred assistive technology configured correctly for them —something that may be difficult or require additional time to provide in a lab setting with lab-provided equipment.

5.3   Technical challenges and pitfalls

During our pilot study, we noticed possible technical challenges and pitfalls related to the camera setup and internet connectivity.

Before running the pilot, we discussed the possibility of recording both the participant and the screen separately. We ultimately chose only the shared screen as we felt that it was a fair balance between richness of information and requiring additional set-up of the users. This is somewhat mitigated by the fact that most video conferencing systems record both the screen and a thumbnail version of the participants' face when participants choose to have their camera on. However, there could have been richer data recordings of the participants' faces (for example, by filming the participant using additional technologies).

Adding additional cameras requires extra set-up by the participant and knowledge of the software. This might have put additional stress on the participants. In our pilot, some participants in the synchronous evaluations seemed already slightly stressed in the beginning of the evaluation as they needed to perform the evaluation tasks while setting up and managing the conference systems on their own. We could improve this in future protocols by providing a checklist before the evaluation. In addition, it is always good to emphasize (or reemphasize) that the artifact is being evaluated and not the participant in these cases [4]. Using supportive statements and maintaining a calm voice can also relieve some stress from the participant.

Moreover, McLaughlin et al. [20] emphasized the importance of considering suitable hardware and software in a study evaluating medical devices remotely. Connectivity and internet access should also be considered. One issue to keep in mind when running synchronous remote evaluations in this way is that it depends on the participant and the evaluator having a stable internet connection with sufficient bandwidth and corresponding equipment (i.e. a microphone, a camera, and a computing device). While this may be taken for granted in some parts of the world, it is not universal. This may lead to excluding segments of people that should be part of an evaluation. Alternate methods like local evaluations or field studies have to be used in these cases. In the pilot conducted in Norway, we did not observe any interruptions related to the Internet connection during the pilot study.

5.4   Obtaining informed consent

One issue that arose during the pilot was how to properly record that someone has given informed consent. Participants would normally sign a paper for documentation in local evaluation studies. Having participants send a certified letter with a signed form via the postal service, however, was impractical and would have defeated the purpose of the participants' convenience of remote evaluation.

Ideally, the participants in the synchronous and asynchronous evaluation would have signed, taken a picture of, and sent back a respective form right after the initial e-mail. However, the participants did not always follow the instruction of providing informed consent before the appointment for the evaluation began.

To keep the process simple, we chose to record verbal consent on video where participants consented to both the terms of the study and the video recording. In cases where participants only consented to the study itself, we asked them to send a text message by phone or an e-mail with their consent.

5.5 Richness of data in synchronous and asynchronous evaluations

Given the small participant sample in the pilot, we cannot draw any general conclusions about the difference in issues between synchronous and asynchronous evaluations. We can, however, discuss what we experienced as increased richness of the data we could collect with screen sharing and recording of video conferencing systems.

From our pilot, the participants in asynchronous and synchronous evaluations could complete the tasks, and we were able to collect data from both types of evaluations. This is positive and means that our documents for the evaluations were written well enough. Generally, we were able to collect richer, more detailed data from the synchronous evaluations than the asynchronous evaluation. Our documents for the asynchronous evaluations allowed participants to write detailed answers to the questions. However, in practice the asynchronous participant would write that there were some problems logging in, but not go into details since the participant eventually were able to log in. Contrast this with a synchronous evaluation where an evaluator could watch the participant struggle with the login interface and detect nuances in the participant's voice, actions, and reactions. Previous research already indicated that the tone of voice in phone evaluations was enough to sense frustration [6], and we did notice that facial expressions could add even more information if a participant allowed sharing of their video camera image. The evaluator could then investigate the participants' thoughts or feelings in synchronous evaluations, which often lead to the discovery of additional usability and accessibility issues.

The recording of the synchronous evaluations provided additional rich data since it was possible to look back and investigate some details more deeply than under the initial evaluation. Also, the shared screen made it easier for evaluators to follow the participants on the screen and back track any possible challenges. This *can* be easier with newer video conferencing systems since recording and screen sharing has become an integrated part of most systems. In lab or field studies, evaluators would either have to install additional software, or be next to or behind a participant (i.e., shoulder surfing), which could feel uncomfortable and intrusive.

That asynchronous evaluations provide fewer and less rich data than synchronous evaluations is consistent with previous findings [27]. One way of improving the data for asynchronous evaluations could be asking the asynchronous participants to also record their sessions and send the recording to the evaluators. This does put more burden on the asynchronous participant and evaluators as it may require installing and learning additional software, having storage space for the recording, and needing a secure way to transfer and store the recording. On the other hand, some video conferencing systems provide the ability to record

and share screen in a session of just one person, so the actual recording may become easier over time.

5.6   Use of questionnaires

During our evaluation, we noticed the importance of the distinction between questionnaires and open-ended questions. In this pilot study, we used a questionnaire about the CAPABLEprototype (perceived safety and security, perceived usefulness, and utility) and the remote evaluation process. In addition to usability and accessibility, answers to such questions are important for the future success of the remote evaluation protocol. However, reflecting upon our current findings, it was unclear whether the participants' answers were based on the current version of the CAPABLE, or a potential future improved CAPABLEsolution. It may also have been difficult, especially for a formative evaluation, to answer these questions by assigning a value. It may make more sense to probe for more qualitative data around these perceptions in future evaluations. Therefore, we will improve the questions in the protocol to clarify these aspects. We will, for example, include more open-ended questions as a good approach to obtain more insights from the participants because they can talk more about their experience and clarify their answers to the questionnaire.

5.7   Limitations

This study was a pilot study with limited reach. An obvious limitation is the small amount of participants, and that not all of the participants were representative end users, i.e. elderly people above 65 years old, citizens who are having an impairment, or using more than three prescribed medications, or having potential nutritional risks. Moreover, we only had one participant using one type of assistive device, i.e. a screen reader. We have not yet investigated the protocol for other assistive devices or other impairments. In addition, the participants in this study were competent users of video conferencing systems. Thus, they might face less challenges when performing the remote evaluations. During future evaluations, we will investigate performances of participants with lower ICT competence or experience, and on other devices like mobiles, etc.

Despite the limitation and some identified issues for improvement, we have verified that the protocol is feasible for future remote evaluations. The documents used and sent to the participants in the remote evaluations were prepared with keeping accessibility and UD in mind. We also identified some usability and accessibility problems of the CAPABLEsolution from the participants' feedback. These issues have been reflected accordingly, to provide a more accessible and inclusive remote evaluation experience to the end users, and a more universally designed CAPABLEfor them in the future.

6   Future work and conclusion

In this study, we aimed to investigate the use of video conferencing systems in the remote evaluation of usability and accessibility of web applications by exploring possibilities, challenges, and limitations through our pilot study. Following the protocol, we conclude that conducting remote evaluation of usability and accessibility of web applications is feasible for a wider study involving more participants with different impairments using different assistive devices. The discussed protocol was designed with accessibility and UD in mind, and our findings indicate that it was accessible and universally designed enough to be used for a proficient screen reader user.

We discussed advantages of video conferencing systems in the remote evaluation. The participants could perform the usability and accessibility evaluations in their most natural settings of surroundings and devices. Flexibility is given to participants in terms of choosing to take part in a synchronous or asynchronous evaluation. Those who choose synchronous evaluation, have the freedom to use the video conferencing system with which they are most familiar. For those who are busy during working hours or any other reason, they can opt for the asynchronous evaluation and conduct the evaluation by themselves whenever they want. We observed general advantages of the synchronous evaluation related to the quantity and richness of the observed data. At the same time, video conferencing systems can add an additional layer of richness by allowing the evaluators to follow the user on-screen and being able to read facial impressions and body language. The possibility to record both screen actions and camera images that many video conferencing systems provide can help the analysis afterwards even more. Moreover, conducting remote evaluation with assistive technology users using video conferencing systems can be an advantage when the assistive technology users do not have to bring their assistive devices with them.

Finally, we managed to identify issues related to the evaluation protocol through the pilot study. These issues have been addressed accordingly to ensure iterations of the usability and accessibility evaluations can be conducted in a more feasible way. In addition, a better participants' experience can be provided when the remote evaluation process itself is more accessible and inclusive. Moreover, we identified both strengths and weaknesses in the design of CAPABLEin relations to usability, accessibility, and UD through the remote evaluations. Since we only conducted a pilot study, we did not have sufficient participants for a quantitative analysis of the findings. Future work should therefore focus on conducting iterations of usability and accessible evaluation using CAPABLEwith more diverse participants and improving CAPABLEbased on the participants' feedback.

References

1. Andreasen, M.S., Nielsen, H.V., Schrøder, S.O., Stage, J.: What happened to re-mote usability testing? an empirical study of three methods. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 1405–1414. CHI '07, Association for Computing Machinery, San Jose, California, USA (Apr 2007). https://doi.org/10.1145/1240624.1240838, https://doi.org/10.1145/1240624.1240838

2. Begnum, M.E.N.: Universal Design of ICT: A Historical Journey from Special-ized Adaptations Towards Designing for Diversity. In: Antona, M., Stephanidis, C. (eds.) Universal Access in Human-Computer Interaction. Design Approaches and Supporting Technologies. pp. 3–18. Lecture Notes in Computer Science, Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-49282-3_1

3. Bonacin, R., Dos Reis, J.C., Baranauskas, M.C.C.: Universal Participatory Design: Achievements and Challenges. SBC Journal on Interactive Systems **10**(1), 2–16 (2019)

4. Boren, T., Ramey, J.: Thinking aloud: reconciling theory and practice. IEEE Transactions on Professional Communication **43**(3), 261–278 (Sep 2000). https://doi.org/10.1109/47.867942

5. Brooke, J.: SUS - A quick and dirty usability scale. In: Jordan, P.W., Thomas, B., Werdmeester, B.A., McClelland, I.L. (eds.) Usability Evaluation in Industry, pp. 189–194. Tayler & Francis, London, UK, first edn. (1996)

6. Brush, A.B., Ames, M., Davis, J.: A comparison of synchronous remote and local usability studies for an expert interface. In: CHI '04 Extended Abstracts on Human Factors in Computing Systems. pp. 1179–1182. CHI EA '04, Association for Computing Machinery, Vienna, Austria (Apr 2004). https://doi.org/10.1145/985921.986018, https://doi.org/10.1145/985921.986018

7. Bruun, A., Gull, P., Hofmeister, L., Stage, J.: Let your users do the testing: a comparison of three remote asynchronous usability testing methods. In: Proceed-ings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 1619–1628. CHI '09, Association for Computing Machinery, Boston, MA, USA (Apr 2009). https://doi.org/10.1145/1518701.1518948, https://doi.org/10.1145/1518701.1518948

8. Castillo, J.C., Hartson, H.R., Hix, D.: Remote usability evaluation: can users report their own critical incidents? In: CHI 98 Conference Summary on Human Factors in Computing Systems. pp. 253–254. CHI '98, Association for Computing Machinery, Los Angeles, California, USA (Apr 1998). https://doi.org/10.1145/286498.286736, https://doi.org/10.1145/286498.286736

9. Experience, W.L.i.R.B.U.: Formative vs. Summative Evaluations (Jul 2019), https://www.nngroup.com/articles/formative-vs-summative-evaluations/, last checked on February, 2nd 2021.

10. Fischer, B., Peine, A., Östlund, B.: The Importance of User Involvement: A Sys-tematic Review of Involving Older Users in Technology Design. The Gerontolo-gist **60**(7), e513–e523 (Sep 2020). https://doi.org/10.1093/geront/gnz163, https://academic.oup.com/gerontologist/article/60/7/e513/5644100

11. Fuglerud, K.S.: Inclusive design of ICT: The challenge of diver-sity. Ph.D. thesis, University of Oslo, Faculty of humanities (2014). https://doi.org/10.13140/2.1.4471.5844, http://publications.nr.no/1418159224/Fuglerud{_}2014{_}PhD-Thesis{_}Inclusive-design-of-ICT.pdf

12. Fuglerud, K.S., Halbach, T., Tjøstheim, I.: Cost-benefit analysis of universal design (Jan 2015)
13. Giannoumis, G.A., Stein, M.A.: Conceptualizing Universal Design for the Information Society through a Universal Human Rights Lens. International Human Rights Law Review **8**(1), 38–66 (Jun 2019). https://doi.org/10.1163/22131035-00801006, https://brill.com/view/journals/hrlr/8/1/article-p38_38.xml
14. Halbach, T., Fuglerud, K.S.: On assessing the costs and benefits of universal design of ict. Studies in health technology and informatics **229**, 662–672 (2016)
15. Hartson, H.R., Castillo, J.C., Kelso, J., Neale, W.C.: Remote evaluation: the network as an extension of the usability laboratory. In: Proceedings of the SIGCHI conference on human factors in computing systems. pp. 228–235 (1996)
16. Hertzum, M., Jacobsen, N.E.: The Evaluator Effect: A Chilling Fact About Usability Evaluation Methods. International Journal of Human–Computer Interaction **13**(4), 421–443 (Dec 2001). https://doi.org/10.1207/S15327590IJHC1304_05, https://doi.org/10.1207/S15327590IJHC1304_05
17. Janson, A.L., Moen, A., Fuglerud, K.S.: Design of the CAPABLE prototype: preliminary results of citizen expectations. In: Norwegian Centre for E-health Research og EHiN Research (ed.) Linköping Electronic Conference Proceedings. pp. 181–181. Linköping Electronic Conference Proceedings, No (2019), http://www.ep.liu.se/ecp/161/ecp19161.pdf
18. Janson, A.L., Moen, A., Fuglerud, K.S.: Design of the capable health empowerment tool: Citizens' needs and expectations. Studies in health technology and informatics **270**, 926–930 (2020)
19. McFadden, E., Hager, D.R., Elie, C.J., Blackwell, J.M.: Remote usability evaluation: Overview and case studies. International journal of human-computer interaction **14**(3-4), 489–502 (2002)
20. McLaughlin, A.C., DeLucia, P.R., Drews, F.A., Vaughn-Cooke, M., Kumar, A., Nesbitt, R.R., Cluff, K.: Evaluating medical devices remotely: Current methods and potential innovations. Human Factors **62**(7), 1041–1060 (2020)
21. Miao, M., Pham, H.A., Friebe, J., Weber, G.: Contrasting usability evaluation methods with blind users. Universal Access in the Information Society **15**(1), 63–76 (Mar 2016). https://doi.org/10.1007/s10209-014-0378-8, https://doi.org/10.1007/s10209-014-0378-8
22. Microsoft: Accessibility overview of Microsoft Teams. https://support.microsoft.com/en-us/office/accessibility-overview-of-microsoft-teams-2d4009e7-1300-4766-87e8-7a217496c3d5 (2020)
23. Miljøverndepartementet: T-1468 B/E Universell utforming (Nov 2007), https://www.regjeringen.no/no/dokumenter/t-1468-universell-utforming/id493083/
24. Nielsen, J.: Demonstrate Thinking Aloud by Showing Users a Video (2014)
25. Øyvann, S.: Vipps har vunnet Innovasjonsprisen for universell utforming — Computerworld (jan 2021), https://www.cw.no/artikkel/utvikling/vipps-har-vunnet-innovasjonsprisen-universell-utforming
26. Persson, H., Ohlsson, K., Petersén, S., Jonsäll, A.: Unexploited Resources in Interaction Design for Universal Access: People with Impairments a s a Resource for Interaction Designers. In: Stephanidis, C. (ed.) Universal Access in Human-Computer Interaction. Addressing Diversity. pp. 145–153. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02707-9_16
27. Petrie, H., Hamilton, F., King, N., Pavan, P.: Remote usability evaluations with disabled people. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 1133–1141. CHI '06, Association for Computing Machinery,

Montréal, Québec, Canada (Apr 2006). https://doi.org/10.1145/1124772.1124942, https://doi.org/10.1145/1124772.1124942

28. Power, C., Petrie, H.: Working With Participants. In: Yesilada, Y., Harper, S. (eds.) Web Accessibility, pp. 153–168. Human–Computer Interaction Series, Springer, London (2019). https://doi.org/10.1007/978-1-4471-7440-0_9, https://doi.org/10.1007/978-1-4471-7440-0_9

29. Richter, A.: Locked-down digital work. International Journal of Information Management **55**, 102157 (Dec 2020). https://doi.org/10.1016/j.ijinfomgt.2020.102157, http://www.sciencedirect.com/science/article/pii/S0268401220308422

30. Rickly, J., Halpern, N., Hansen, M., McCabe, S., Fellenor, J.: Covid-19: The effects of isolation and social distancing on people with vision impairment (Sep 2020). https://doi.org/10.17639/nott.7074, https://rdmc.nottingham.ac.uk/handle/internal/8608, last checked on January 28th, 2021.

31. Røssvoll, T., Fuglerud, K.S.: Best Practice for Efficient Development of Inclusive ICT. In: Stephanidis, C., Antona, M. (eds.) Universal Access in Human-Computer Interaction. Design Methods, Tools, and Interaction Techniques for eInclusion. vol. 8009, pp. 97–106. Springer Berlin Heidelberg (2013). https://doi.org/10.1007/978-3-642-39188-0_11, http://dx.doi.org/10.1007/978-3-642-39188-0{_}11

32. Sauro, J.: A Practical Guide to Measuring Usability: 72 Answers to the Most Common Questions about Quantifying the Usability of Websites and Software. Measuring Usability LCC, Denver (2010)

33. Schmutz, S., Sonderegger, A., Sauer, J.: Implementing Recommendations From Web Accessibility Guidelines: Would They Also Provide Benefits to Nondisabled Users. Human Factors **58**(4), 611–629 (Jun 2016). https://doi.org/10.1177/0018720816640962, https://doi.org/10.1177/0018720816640962

34. Schmutz, S., Sonderegger, A., Sauer, J.: Implementing Recommendations From Web Accessibility Guidelines: A Comparative Study of Nondisabled Users and Users With Visual Impairments. Human Factors **59**(6), 956–972 (Sep 2017). https://doi.org/10.1177/0018720817708397, https://doi.org/10.1177/0018720817708397

35. Schnepp, J., Shiver, B.: Improving deaf accessibility in remote usability testing. In: The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility. pp. 255–256. ASSETS '11, Association for Computing Machinery, Dundee, Scotland, UK (Oct 2011). https://doi.org/10.1145/2049536.2049594, https://doi.org/10.1145/2049536.2049594

36. Sova, D.H., Nielsen, J.: 234 Tips and Tricks for Recruiting Users as Participants in Usability Studies. Tech. rep., Nilsen Norman Group (jan 2003), https://www.nngroup.com/reports/how-to-recruit-participants-usability-studies/

37. Venable, J., Pries-Heje, J., Baskerville, R.: FEDS: A Framework for Evaluation in Design Science Research. European Journal of Information Systems **25**(1), 77–89 (2016). https://doi.org/10.1057/ejis.2014.36

38. World Wide Web Consortium (W3C): Web content accessibility guidelines (wcag) 2.0. https://www.w3.org/TR/WCAG20/ (Dec 2008), https://www.w3.org/TR/WCAG20/, last checked on 2020-09-17.

39. World Wide Web Consortium (W3C): Web content accessibility guidelines (wcag) 2.1. https://www.w3.org/TR/WCAG21/ (Jun 2018), https://www.w3.org/TR/WCAG21/, last checked on 2020-09-17.

40. Zoom   Video   Communications: Meetings   for   Everyone.
    https://zoom.us/accessibility (2019)