Post-Print version
Desolda, G., Lanzilotti, R., Caivano, D., Costabile, M. F., & Buono, P. (2023). Asynchronous Remote Usability Tests Using Web-Based Tools Versus Laboratory Usability Tests: An Experimental Study.
IEEE Transactions on Human-Machine Systems, 53(4), 731–742. https://doi.org/10.1109/THMS.2023.3282225

Asynchronous remote usability tests using web-based tools vs laboratory usability tests: an experimental study

Giuseppe Desolda, Rosa Lanzilotti, Danilo Caivano, Maria Francesca Costabile (*Life Senior Member, IEEE),*
Paolo Buono (*Member, IEEE),*
Computer Science Department, University of Bari Aldo Moro, Via Orabona 4, 70125 Bari, Italy

***Abstract*—Remote usability testing is performed by evaluators who are in different physical locations from the participants (synchronous remote testing) and possibly operating at different times (asynchronous remote testing). The tools developed in recent years to support remote tests exploit web technology based on HTML5 and JavaScript ES6 and thus enable previously unexplored scenarios. However, studies providing evidence on the benefits or drawbacks of utilizing recent web-based tools have not yet been reported in the literature. This article sheds some light on the impact of such tools on asynchronous remote usability testing of websites by reporting an experimental study with 100 participants and 15 evaluators to compare real-time laboratory tests with asynchronous remote tests. The study investigates 1) how the metrics results of asynchronous remote usability tests performed through a web-based tool differ from those of usability tests conducted in real-time laboratory settings, and 2) how the experience of participants differs in the two types of tests. The lessons learned in the study are instrumental in informing the design of future tools. Some results of particular interest indicate that the web technology used by the tool for asynchronous remote testing affects task execution times and participants' satisfaction. Another indication is that slow internet connections must be managed in asynchronous remote testing; slow connections introduce delays when transferring large amounts of collected data, which, together with the lack of human support, make participants of asynchronous remote tests more prone to feel negative emotions.**

***Index Terms* — Website usability evaluation, web-based asynchronous remote tests, comparison study.**

I. Introduction and Motivation

With the ever-increasing proliferation of interactive systems since the 1980s, usability has continued to be a key factor of software quality and is considered an important component of the wider concept of user experience (UX). Several usability evaluation techniques have been proposed (see, e.g., [1, 2]). One of the most successful is usability testing, which can be performed in more or less controlled environments. According to various authors, usability testing is a valid technique in terms of the number and quality of detected usability problems [3, 4]. However, this technique is often neglected, primarily because professional developers think it is very resource-demanding, they do not have adequate expertise to perform it, or there is limited automation in the evaluation process [5, 6]. Other important concerns about usability testing are the factors that might influence the outcomes of the testing. Salvendy proposed a formal model with four chief factors [7], previously identified in [8]: user characteristics, task scenarios, product properties, and testing environment. The latter relates to the important issue of the best location for conducting the test, i.e., a laboratory or a natural setting.

The possibilities of internet technology at the beginning of the 1990s pushed researchers to investigate remote usability testing (see, e.g., [9, 10]), i.e., usability testing performed by evaluators who are at a spatial distance from the participants and possibly at a time distance. There are two types of remote evaluations: synchronous and asynchronous. In the synchronous type, also called "live" or "collaborative", the participants and evaluators operate at the same time – they are in different locations but are connected thanks to screen-sharing software (to see the user's screen), telephone or ad hoc software for audio/video communication (e.g., Skype) [11]. All the data are automatically gathered and stored by dedicated tools. One of the main differences from laboratory (or in-lab) tests is that the users participate in the study from their natural environments, using their personal computers and tools. In the asynchronous type, the participants and evaluators are separated in both space and time, since the participants perform the test when it is convenient for them, without any supervision by or live communication with the evaluator(s).

Participants can be easily recruited for remote testing since they perform the study tasks in their natural settings (e.g., offices, home), which also increases the ecological validity [12]. In addition, remote testing permits notable budget reductions, since there are no costs for lab renting and no travel expenses for the participants. More culturally diverse users can participate in the test, while keeping the cost of the study low. Thanks to several comparison studies, there is empirical evidence that remote testing results are generally comparable to those of in-lab testing (see, e.g., [12-16]).

Since the beginning of this millennium, several software tools for performing remote usability testing of websites have been developed (e.g., see [17-20]). The technology exploited by the early tools was limited; for example, they required the installation of specific clients (which limited their adoption across operating systems (OSs)) as well as access to the website source code; they did not permit audio or video recordings, and comments of the participants were collected by using simple forms [15, 21-24].

Technology advances in recent years, in particular the advent of HTML5 (introduced in 2014) and JavaScript ES6

(introduced in 2015), have enabled previously unexplored scenarios in remote usability testing. Currently, it is possible to develop more powerful tools for remote usability testing of websites that, for example, work across browsers, are OS-independent, do not require access to the website source code, and permit screen recording and user-interaction tracking as well as the use of peripheral devices such as webcams and microphones to capture the face and speech of the participants during tests [25-29]). The functionality provided by these recent tools permits collecting much more quantitative (e.g., questionnaire answers, task success/time) and qualitative data (e.g., audio/video recording), which may reveal further usability issues.

The COVID-19 pandemic has highlighted the importance of tools for remote work and, in the case of usability studies, the value of tools for asynchronous remote testing. Companies, researchers and practitioners have intensified the adoption of web-based tools for remote testing in their daily practices. This trend will likely continue to increase.

While the pros and cons of remote usability testing tools developed up to 2014 are known, to the best of our knowledge, there is no evidence on the benefits or drawbacks of the use of recent web-based tools, i.e., those developed after 2014. The novel contribution of this article is that it sheds some light on the impact of such tools on asynchronous remote usability testing of websites. We selected eGLU-Box PA, a tool that is representative of these recent tools, to perform an experimental study with 100 participants and 15 evaluators to compare lab-based usability tests with asynchronous remote tests. The study aims to investigate 1) how the metrics results of asynchronous remote usability tests performed through a web-based tool differ from those of usability tests conducted in real-time laboratory settings and 2) how the emotions, workload, and overall experience of participants differ in the two types of tests. The study outcomes provide lessons that are instrumental in informing the design of advanced tools supporting asynchronous remote testing. In particular, regarding objective 1), the study revealed that in asynchronous remote testing, task execution times are slightly slower, the task success rate is not affected, and the participants' satisfaction about the evaluated website is lower. Regarding objective 2), it was found that emotions of participants to asynchronous remote tests are more negative; one reason is the lack of human support, and this confirms a criticality already revealed by some previous studies, e.g., [30]. Moreover, the overall participants' experience is further worsened if they do not have a good internet connection because long delays are incurred by the tool in storing the large amount of collected data.

The article is organized as follows. Section 2 briefly reports related work on remote usability testing, describing some tools that support it, as well as studies conducted thus far that compare remote usability testing with other testing approaches. Section 3 describes the study performed, discusses the results and summarizes the limitations of the study. Section 4 reports the lessons learned, and Section 5 provides the conclusions.

II. Remote usability testing: tools and comparison studies

Remote usability testing was defined in 1996 to limit some usability testing drawbacks [31]. In this section, the features of the tools used in performing remote usability testing are illustrated and studies that compare different testing approaches are discussed.

A. Tools for remote usability testing

Some of the first software tools for the remote usability testing of websites, which have been developed since 1995, are described in [17-20, 32, 33]. They were used primarily to allow participants to test websites from their locations at times when it was convenient for them. These tools offer similar functionalities: they aggregate data collected from several test sessions and log the participant interaction to obtain data, such as the paths users take and entry and exit pages, to provide evaluators with data to discover the obstacles users encounter in performing tasks; they also automatically compute metrics such as the task success rate and the average time spent on a task.

Three main approaches were adopted for implementing these tools, i.e., as a *Web browser*, as a *web application*, and as a *proxy*. The *Web browser* tools consist of using a specific client-side application, i.e., a Web browser, that study participants and evaluators must install on their personal computers (e.g., [34]). The main disadvantages are that participants must install a specific browser and sometimes ad hoc hardware and/or software. This is not always a quick and easy task and strongly limits the use of the browser on different operating systems or in contexts where it is not possible to install software without administration grants.

Web application tools are web applications integrated into the websites to be evaluated and allow tasks and surveys to be administered and user behavior to be monitored (e.g., [35]). In addition to the inability to record data from webcams, microphones and screen capture, another main limitation is the lack of full access to the website code, which strongly reduces the possibility of evaluating any website.

Proxy tools are based on a proxy server that collects logs of the interactions between clients and the tested website [32]. These tools do not permit configuring a controlled study with the possibility of administering tasks and surveys, and it is not possible to collect qualitative data.

The advent of HTML5 in 2014 and JavaScript EM6 in 2015 led to a new generation of tools for remote usability testing. Examples are *Loop11* [25], *Lookback* [26], *Userlytics* [27], *UserTesting* [28] and *eGLU-Box PA* [36, 37]. These tools are characterized by similar novel and useful features that might contribute to overcoming some limitations of the previous generation of tools. Indeed, these tools permit to accomplish the following:

1) evaluate any website without accessing its source code;

2) use the tool regardless of the specific operating system since a common web browser is sufficient;

3) use the tool without the need to install software, which might require administration grants;

4) execute the tool with various devices such as a personal computer, a tablet, and a smartphone;

5) collect very rich data during the tests (user logs, task time, success rate, and audio and video of participants' interactions).

B. Usability test comparison studies

The effects of using different usability testing approaches are discussed in studies that compare conventional lab-based testing and remote testing of websites. These studies often used academic prototype tools for remote usability testing [15, 16] or general-purpose solutions for video-audio conferencing, such as WebEx or Microsoft NetMeeting (see, for example, [14, 23]).

The dependent variables most used in such studies are task success rate, task execution time, number and severity of usability problems, and participant satisfaction. Focusing on remote usability testing of websites, which is the main interest of this article, no significant difference was reported between remote testing and lab-based testing in most studies (see, for example, [15, 22-24]), indicating that in general, one testing approach is not better than the other. However, a few studies partially contradict this finding. Specifically, [14, 15] report that remote testing allows the detection of more problems than lab testing, and [15, 16] show that remote testing identifies content-related usability issues more easily than device-related issues. In [24], participants performing remote testing seemed slightly less motivated, as shown by the shorter time spent on task execution and the greater likelihood of giving up on a task. Conversely, in [21, 23], asynchronous testing was considerably more time-consuming and identified fewer usability problems. The differences in these studies may depend on the products being tested, the users' characteristics, the tasks being performed, and/or the testing environments, which are acknowledged as the main factors affecting testing [8, 38].

The study in [39] is unique since it compares the test performed with a prototype of a 3D virtual usability testing laboratory built using the Open Wonderland toolkit, with a more conventional lab-based test and with a synchronous remote test using the WebEx platform. The three testing approaches agreed in terms of task execution time and the number and severity of problems. However, there was a significant difference in the workload experienced by both test participants and evaluators, with the conventional lab condition requiring the lowest workload and the virtual lab and the remote conditions requiring similar higher workloads. The participants experienced greater involvement and a more immersive experience in the virtual lab condition than in the remote condition, while no significant difference was found between the remote and conventional lab conditions.

A 2019 paper reports three experiments comparing synchronous and asynchronous remote usability testing to lab-based testing under various operational conditions (dual-task demands, poor product usability) and using various artifacts (website, computer-simulated mobile phone and fully operational smartphone) [12]. The results showed that there was no difference between remote testing and lab-based testing under favorable operational conditions. Some complex patterns emerged in less favorable conditions, i.e., when the testing method was combined with other factors such as dual-task demands and poor product usability. The overall result confirms no advantage of one testing approach to another. Notably, no web-based tools for synchronous and asynchronous remote usability testing were used in that study; the authors simply created a website that was tested remotely by asking the users to perform a set of preassigned tasks and fill in forms already provided by the website. Thus, despite being a very recent study, it does not provide any indication about advantages and drawbacks of web-based tools for synchronous and asynchronous remote usability testing.

In the analysis of the studies in the literature on remote usability tests, we did not find any studies addressing remote tests performed by using recent web-based tools (after 2014). As discussed in Section II.A, these modern tools (e.g., Loop11 [25], Lookback [26], Userlytics [27], UserTesting [28], and eGLU-Box PA [36, 37]) are characterized by integrated web-based environments that allow participants to perform a usability test by simply using a website; in general, users do not need to install new software on their personal computers, as required in the past [17-20, 32, 33]. Thanks to advances in web technologies, such tools automatically collect quantitative (e.g., participant logs, task time, success rate) and qualitative data (e.g., the recording of a webcam, microphone, and screen).

III. COMPARING REAL-TIME IN-LAB USABILITY TESTS AND WEB-BASED ASYNCHRONOUS REMOTE TESTS

This section reports the experimental study carried out to compare usability tests of websites performed real time in laboratory with asynchronous remote usability tests performed with a recently developed web-based tool. The study is motivated by the fact that the technological advances exploited by the new generation of web-based tools might influence the participants' performance and experience.

A. eGLU-Box PA: the tool for asynchronous remote testing

eGLU-Box PA is the tool used in this study. It is a web application that supports the asynchronous remote testing of websites and, by exploiting advanced web technologies, implements the novel features of the recent tools mentioned in Section II. Such tools are very similar in terms of not only the technology they are based on but also the ways the asynchronous remote tests are organized and executed, the variety of data they can collect, and the support they provide for data analysis. Thus, eGLU-Box PA was chosen for the study since it well represents the new generation of tools for asynchronous remote testing. It is a professional tool currently adopted by hundreds of web managers to perform remote usability tests and has been certified according to ISO/IEC 25010:2005 standard "Software engineering - Software product Quality Requirements and Evaluation" (see [36]).

In the experimental study, to make a careful comparison between the two experimental conditions (real-time in-lab vs. asynchronous remote), we needed to record qualitative data (through webcam, microphone and desktop) from the moment the participants logged-in the tool and collect detailed data.

These requirements go beyond what is generally considered during a usability test, and thus, they are not provided by the tools available on the market. This is another reason for selecting eGLU-Box PA in our study because we could access its source code and modify it accordingly to better manage the experiment.

It is worth mentioning that although eGLU-Box PA has been conceived in projects funded by the Italian government, whose goal was to improve the usability of public administration (PA) websites, it has been designed as a general-purpose tool that supports the evaluation of any website, in line with similar tools such as Loop11 [25], Lookback [26], Userlytics [27], and UserTesting [28]. Some authors of this article were involved in the design and development of eGLU-Box PA through an iterative human-centered design process. Several stakeholders, with and without Human-Computer Interaction (HCI) and Information Technology (IT) skills, were involved in the process. In the early phases of the design, 50 website managers were recruited. They were involved in activities such as workshops, interviews and questionnaires aimed at eliciting the tool requirements and evaluating the prototypes under development. Other web managers were asked to use other versions of eGLU-Box PA to perform usability tests of their websites. Some of the studies carried out to evaluate eGLU-Box PA at various stages are reported in [37, 40, 41], in which further details on the design, development, and use of this tool can also be found. Please note that a previous version of eGLU-Box PA was called UTAssistant.

To guarantee the development of a robust, safe, scalable and fast web application, eGLU-Box PA has been developed by adopting the Laravel framework[1] since it is one of the most popular solutions to develop professional web applications. Moreover, privacy and security are guaranteed following proper design patterns and solutions. Regarding privacy, a usability test can be set as anonym, meaning that no data stored in the database (task time/success and questionnaire answers) can be traced back to a specific user to ensure its completely anonymous participation. This feature has also been assessed during the ISO certification process. In addition, in case of recording audio and video and, more generally, to inform participants about the collection of sensible data, digital consent forms were provided and agreed to before starting the study. Regarding security, native browser APIs for recording audio and video are used by eGLU-Box PA, and HTTPS connections protect the transfer of the collected data from the participants' personal computers (e.g., audio-video recording) to the server.

Evaluators and test participants can access and use the eGLU-Box PA website from their personal computers or mobile devices wherever and whenever by using a web browser without installing specific software. The evaluators create a usability test; after the test, they visualize the data automatically gathered and analyzed by the platform. Specifically, to create a usability test, they are guided to define:

- a brief introductive text to welcome participants;
- the set of tasks (specifying the URL where the task starts and the URL where the task is considered completed);
- post-test questionnaires chosen among System Usability Scale (SUS), Usability Metric for User Experience (UMUX-Lite) and Net Promoter Score (NPS);
- custom questionnaires;
- the data to be automatically collected during the user interaction with the website under evaluation (i.e., desktop recording, video and/or audio recording);
- the participants to be invited.

The invited participants receive an email containing a URL to start the usability test. Each participant is guided by the platform step by step: first, the platform tests the personal computer peripherals that may be necessary to record the data selected by the evaluator (e.g., webcam and microphone); second, it administers the tasks to be carried out in order; finally, it administers the questionnaires selected by the evaluator(s), if any. During the test execution, eGLU-Box PA automatically stores all the collected data (task time/success, screen recording, microphone recording, webcam recording, questionnaire answers) on its web server. Usability test metrics are also automatically analyzed and made available to the evaluators for analysis. In particular, eGLU-Box PA summarizes *efficiency*, measured through task execution time; *effectiveness*, measured through task success of each task and of each user; and *satisfaction*, measured through the administered questionnaires and visualized by using proper graphs. eGLU-Box PA also allows the playback of all audio/video recordings and, if needed, the annotation of particular participants' actions or comments related to usability issues to facilitate a qualitative analysis; such videos also help to get what evaluators see in laboratory user studies. Finally, it produces a PDF report of all the previous results.

B. Research questions, study design and participants

The twofold objective of the study is expressed more formally in two research questions that this study aims to answer:

RQ1: Do the metrics results of asynchronous remote usability tests performed through web-based tools differ from those of usability tests conducted real time in laboratory?

RQ2: Does the experience of participants in asynchronous remote usability tests performed through web-based tools differ from the experience of participants in usability tests conducted real time in laboratory?

A between-subject design was adopted, with the test method as an independent variable and two between-subject conditions: *real-time in-lab test* condition (also called *lab test* for short in the rest of this article) and *asynchronous remote test* condition (also called *remote test* for short). It is worth noting that there is no intent to evaluate the eGLU-Box PA tool itself; it has been used since it well represents the class of web-based tools for asynchronous tests.

A total of 100 usability test participants (31 females, 69 males) were recruited through convenience sampling (see [42] for more on convenience sampling). Their mean age was 28.13

[1] https://www.peerbits.com/blog/laravel-most-popular-php-framework.html

y.o. (SD = 8.1, min = 19, max = 59); 4 of them had a middle school diploma, 86 had a high school diploma, and 10 had a university degree; 32 of them were university students, 54 were workers, 7 were housewives, and 7 were unemployed.

A total of 15 evaluators were also involved, 10 to conduct the lab test (mean age 44.3 years, SD = 4.4, min = 36, max = 49) and 5 to organize the remote test with eGLU-Box PA (mean age 42.8 years, SD = 4.7, min = 37, max = 48). The evaluators had a similar background and expertise in performing usability testing. They were graduate students in Computer Science who had already practiced usability tests during their course on HCI for their bachelor's degree and performed several usability tests for their thesis on HCI.

C. Tested websites and administered tasks

To increase the external validity, attention was devoted to websites of various categories. We started with the identification of popular website categories typically used by common users: video, news, travel, e-commerce, and public administration. The 10 websites eventually tested, 2 for each category, were selected by considering the websites most visited[2] and their usage by users without any prior knowledge since our target participants were purposely recruited without constraint. The selected websites are YouTube and Netflix (Video), Ansa and Repubblica (Italian News), Booking and Trip Advisor (Travel), Amazon and eBay (e-commerce), Italian Ministry of Internal Affairs and Italian Ministry of Defense (Public Administration). For each website, 5 tasks were chosen from typical tasks for the website category. For example, for Amazon and eBay the five tasks were: 1) Go to the log-in page and log in; 2) Search for iPhone X smartphone; 3) Access the wish list (list of favorites on eBay); 4) Search for products on offer; and 5) View the items in your shopping cart.

It is worth remarking that since the detection and analysis of usability problems were not considered in this study, mature websites in terms of usability were purposely selected.

D. Measurements and instruments

Quantitative and qualitative data were collected to answer the research questions. Since one of the aims of the study is to evaluate possible interferences of the web-based tool on the usability metrics results rather than possible differences in the number and type of usability issues that can be discovered in the two types of tests, the usual metrics *task success*, *task completion time* and *participant's satisfaction* were selected to answer RQ1. Satisfaction was measured by the SUS, NPS and UMUX-Lite questionnaires. The SUS measures system usability through 10 statements rated on a 5-point Likert scale [43]. NPS asks a single question: "How likely is it that you would recommend our company/product/service to a friend or colleague?" The answer ranges from 0 to 10 [44]. The UMUX-Lite is composed of only two items that use a 7-point scale and is targeted toward the ISO 9241 definition of usability (effectiveness, efficiency and satisfaction) [45].

RQ2 is a research question that represents one of the novelties of this work; it has not been considered in previous studies. To answer RQ2, the emotions, workload and overall experience of the participants during the test were considered. The *Affectiva SDK*[3] was used to capture and analyze, from the videos recorded by each participant webcam, the emotions felt by each participant during the test. Affectiva detects facial expressions in video frames (5 frames/second analyzed in our case, with video recorded at 25 fps), according to the Emotional Facial Action Coding System (EMFACS) model developed by Friesen & Ekman [46]. This model represents seven emotions: joy, anger, disgust, surprise, fear, sadness and contempt. For every frame, Affectiva computes the value of each emotion from 0 to 100, where 100 indicates maximum emotion intensity. It must be noted that advances in AI make visual emotion recognition (e.g., the Affectiva tool) as reliable as human coding [47, 48] and as precise as more advanced and invasive instruments such as facial electromyography [49].

The subjective workload of the overall testing procedure, without any reference to the tested websites, was measured by administering the NASA Task Load Index (NASA-TLX) questionnaire to the participants at the end of their two tests. Further qualitative measurements were collected through two open questions, administered after NASA-TLX, asking the participants to comment on what they liked most (first question) and what they liked least (second question) about the overall test procedure.

E. Study procedure

The 100 test participants were divided into two homogeneous groups with respect to gender, age, education and job: each group was assigned to a study condition, i.e., lab test or remote test. All participants had similar experiences with the websites chosen for the tests, and none of them had previous experience with asynchronous remote usability testing. Each participant tested two websites, each in a single session. A total of 200 tests (1000 tasks) were executed, 100 in the lab test (500 tasks) and 100 in the remote test (500 tasks). Each website was tested 20 – 10 times for each condition. Websites and task orders were counterbalanced according to a balanced Latin square design.

The procedure to conduct the tests was adapted to the study conditions. In the case of the lab test, 5 groups of 2 evaluators (one acted as an observer and the other as a facilitator) and 10 participants were randomly formed. Each group performed the tests in a quiet university room on a laptop with a 15-inch display with an external mouse. Each group scheduled two participants per day for a total of 5 days. Every participant followed the same procedure. First, the facilitator welcomed the participant, who was then introduced to the study purpose, informed on what to do and signed a consent form. Then, the facilitator provided the participant with a sheet reporting the five test tasks for the website and started to execute each task. At the end of all the tasks, the facilitator asked the participant to complete an online Google form that presented the SUS, NPS and UMUX-Lite questionnaires to be answered one after the other. Before testing the second website, following the same procedure, the participant was invited to relax for five minutes.

[2] https://trends.google.com/trends/

[3] http://developer.affectiva.com/

At the end of the second test, the facilitator asked the participant to complete another Google form with the NASA-TLX questionnaire and the two open questions on the pros and cons of the test procedure. The observer took notes during the procedure. OBS Studio was used during the procedure to record the participant webcam given the goal of RQ2 to analyze and compare participants' emotions.

In the case of the remote test, 5 groups of 1 evaluator and 10 participants were randomly formed. The tests to be performed were created by the evaluator on eGLU-Box PA, and an email was sent to the participants asking them to register on eGLU-Box PA. Afterward, an automatic email notified each participant that the tests could be performed. This email also reported the study purpose, the technical requirements for the test (personal computer, webcam, microphone, the use of a browser – such as Chrome, Firefox, Edge, Safari – and a stable internet connection) and the approximate time required for each of the two tests so that the participant could freely decide when to perform each test without interruption or disturbance. As in the lab test, the test procedure of each website was concluded by administering, through eGLU-Box PA, the SUS, NPS and UMUX-Lite questionnaires. At the end of the second test, the participant answered the NASA-TLX and the two open questions. The remote and in-lab tests took each participant approximately 25 minutes (10 minutes per website plus a 5-minute break after the first test).

F. Data Analysis

Welch's t test (also called unequal variances t test) was computed to analyze task times and the numerical values of the emotions resulting from the analysis of the video by Affectiva because of the violation of normal distribution (assessed with the Shapiro–Wilk test). An independent t test was computed to analyze the questionnaire results (SUS, NPS, UMUX-Lite, and NASA-TLX) since they did not violate a normal distribution (assessed with the Shapiro–Wilk test). Pearson chi-square was computed to analyze success results (dichotomic nominal values, failed or succeeded). An alpha level of .05 was used for all statistical tests. In the case of a significant difference, the effect size was also checked by calculating *Cohen's d_s* [50]. According to Cohen, the difference can be very small (Cohen's d_s 0.00 < 0.20), small (Cohen's d_s 0.20 < 0.50), medium (Cohen's d_s 0.50 < 0.80) or large (Cohen's d_s 0.80 or more).

Four HCI researchers analyzed the two open questions in a systematic qualitative interpretation using an inductive thematic analysis [51].

G. Study results

This section reports the results of the analyses performed on the data collected during the study. In the tables, the variables revealing significant differences are shown in gray cells.

1) Task success and task time

Regarding task success, in the lab test, 59 tasks failed and 441 succeeded, while in the remote test, 48 tasks failed and 452 succeeded. Pearson's chi-square test revealed that there was no statistically significant difference between the lab test and the remote test ($\chi(1) = 1.266$, $p = .306$).

The task time analysis indicated that the average time (in seconds) to complete the tasks in the lab ($\bar{x}$ = 36.51, SD = 28.91) was approximately 12% lower than that in the remote test ($\bar{x}$ = 41.28, SD = 42.03); Welch's t test showed that this difference was statistically significant (t(801.236)=-1.981, p=.048), with a very small effect (Cohen d_s = -0.132).

2) Satisfaction questionnaires

Table 1 reports the results of the questionnaires for the lab test and the remote test, as well as the results of the t tests. Only the SUS score related to the tested website was affected by the use of eGLU-Box PA, since this score was lower in the remote test, and the t test confirmed that the difference was statistically significant ($p = .049$) with a small effect (Cohen d_s = -0.28). According to Lewis and Sauro [52], SUS was decomposed into two factors, i.e., system learnability (SUS statements #4 and #10) and system usability (the other 8 statements). This allowed us to obtain more information. Indeed, the learnability of the tested website was lower in the case of the remote test, and this difference was statistically significant ($p = .000$) with a medium effect (Cohen d_s = -0.59), while no differences emerged in the case of usability ($p = .234$). Table 1 shows that there were no differences between the lab test and the remote test for either NPS ($p = .576$) or UMUX-Lite ($p = .303$).

3) Workload questionnaire

The NASA-TLX questionnaire was administered to all the participants at the end of the study procedure and asked them to answer with reference to the test procedure without considering the evaluated websites. The remote test had a higher NASA-TLX score than the lab test (see Table 2). The t test showed that this difference was statistically significant ($p = .008$) with a large effect (Cohen d_s = 0.96); in other words, the workload of the participants in the lab test was lower than that of the participants in the remote test (the lower the better). To gain more insight from this analysis, the six subscales of the NASA-TLX, i.e., Mental Demand, Physical Demand, Temporal Demand, Performance, Effort and Frustration (each scale ranges from 0 = low to 100 = high), were analyzed separately [53]. The details for each subscale, as well as the results of the t tests, are reported in Table 2: the higher workload of the remote test is mainly caused by the Performance ($p = .023$) and Frustration ($p = .003$) subscales, in both cases with a small effect.

4) Participants' facial expressions

Each participant's video was split into three parts, i.e., test introduction, task execution and questionnaire filling, to differentiate and analyze the participant's emotions during the three main phases of a test. A total of 600 videos (100 participants x 2 tests x 3 phases), for a total of more than 4000 minutes, were analyzed, and 30864 emotions were detected. Although Affectiva was set to analyze the user face every 5 frames, in some frames (e.g., due to rotated face, hands on the face, face out of the webcam view, or low light in the room), the face was not detected. Emotions having a value less than 1 were removed since we empirically observed that values below this threshold are affected by noise produced by Affectiva.

Table 1. Results of the administered questionnaires. In addition to the results for SUS, the results of the two SUS factors of System Learnability and System Usability are reported (indicated as SUS Learnability and SUS Usability, respectively). The variables revealing significant differences are shown in gray cells.

	SUS		SUS Learnability		SUS Usability		NPS		UMUX-Lite	
	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD
Remote	83.17	16.09	84.00	25.07	82.97	14.49	8.39	1.61	6.07	0.99
Lab	87.53	14.95	94.50	11.14	85.78	16.85	8.52	1.66	6.21	0.92
t test	t(198)= 1.981 p=.049*		t(198)= 3.828 p=.000*		t(198)= 1.193 p=.234		t(198)= 1.034 p=.303		t(198)= 0.560 p=.576	
Effect size	0.28		0.59							

Table 2. Results of NASA-TLX are in the first column; the other columns report the values of the six subscales. The variables revealing significant differences are shown in gray cells.

	NASA-TLX		Mental Demand		Physical Demand		Temporal Demand		Performance		Effort		Frustration	
	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD
Remote	21.60	11.59	26.8	18.95	14.20	9.12	20.40	14.83	20.80	12.03	26.50	21.80	19.10	16.14
Lab	17.67	8.89	22.40	16.15	16.00	11.45	17.40	10.69	16.60	13.79	21.70	20.84	13.70	8.48
t test	t(185.55)= 2.96 p=.008*		t(198)= 1.767 p=.079		t(198)= 1.229 p=.221		t(198)= 1.64 p=.102		t(198)= 2.29 p=.023*		t(198)= 1.59 p=.113		t(198)= 2.96 p=.003*	
Effect size	0.96								0.32				0.41	

The results showed that in the remote test, there was an effect on three user emotions, namely, surprise, fear, and sadness, whose values were higher (i.e., these emotions were more intense) than those in the lab test, even if a small effect emerged in all three cases (see Table 3a). As previously mentioned, the emotions felt by the participants were analyzed with reference to the three main phases of the usability test, i.e., introduction, task execution and questionnaire filling. During the introduction phase, there was only one difference, for fear ($p = .004$), with a large effect (Cohen $d_s = 0.81$), indicating that the participants felt more fear in the remote test (Table 3b).

Regarding the task execution phase, which can be considered the most important one, four emotions, namely, sadness and surprise (with a low effect) and fear and disgust (with a medium effect) (Table 3c), appeared in the remote test. Surprise, fear, and sadness were higher in the remote test, while disgust was lower. The final phase involved completing the questionnaires. The results, summarized in Table 3d, reveal that four emotions, namely, disgust and surprise (with a low effect) as well as fear and sadness (with a medium effect), were influenced in the remote test. However, disgust was higher in the remote test.

5) Participants' comments on the test procedure

The answers to the two open questions asking participants what they liked most and what they liked least about the overall test procedure were analyzed in a systematic qualitative interpretation using an inductive thematic analysis [51] by four researchers with senior experience in qualitative data analysis. Two of these researchers started the analysis independently. They systematically generated codes across the collected answers. Then, working together, they grouped the codes into potential themes informed by the open question goals, namely, the good and bad aspects of the test procedure the participants followed in the lab or remote test. A review analysis was carried out by the four researchers, who discussed whether the themes were properly related to the codes, generating a thematic 'map' of the analysis; they also refined the definitions and names of the themes.

All participants answered the two questions, but most of them did not give any specific comments. Some examples answers are "Nothing relevant", "No negative aspects" and "I liked everything". From the analysis of the more articulate answers, the following five themes were developed, namely, two for the lab test and three for the remote test. For each theme, significant participant quotes are reported, with the participant code given in square brackets.

Theme 1. *Clear and simple procedure for real-time in-lab tests*. Most participants in the real-time in-lab test provided very positive comments, primarily related to low effort, understandability and ease of execution of the procedure.

[P24] "The study procedure did not require much effort"

[P12] "The test procedure is clear and easy to understand"

[P4] "I found it really easy to execute all the activities like the tasks and questionnaire, regardless of the specific website difficulties"

Theme 2. *Under pressure during the real-time in-lab tests*. Some comments referred to the pressure caused by the presence of the experimenter.

[P15] "Sometimes the presence of the facilitator created a bit of embarrassment and made me feel under examination"

[P70] "Being assisted by the facilitator made me feel a little pressured"

Theme 3. *High usability of the tool for the asynchronous remote tests*. Even for the asynchronous remote tests, most comments were positive. The participants appreciated that the procedure was completely guided, the interaction with the system functions was easy, and the ability to perform the test whenever and wherever they preferred was very convenient.

Table 3. Emotions felt by participants: (a) during all phases of the usability tests; (b) during the introduction phase; (c) during the task execution phase; and (d) during the questionnaire filling phase. The variables revealing significant differences are shown in gray cells.

	Joy		Anger		Disgust		Surprise		Fear		Sadness		Contempt	
	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD
Remote	77.17	34.81	17.91	25.87	17.75	28.42	17.12	24.04	18.59	19.99	24.60	29.65	50.11	44.15
Lab	75.67	36.45	16.37	25.77	17.98	28.51	13.62	20.94	9.47	12.27	17.31	24.99	51.19	44.11
Welch's t test	t(239.38)=.458 p=.648		t(449.798)=.848 p=.397		t(1571.86)=-.237 p=.813		t(1659.15)=4.68 p=.000*		t(89.665)= 5.044 p=.000*		t(592.64)= 4.260 p=.000*		t(780.405)=-.484 p=.628	
Effect size							0.15		0.46		0.25			

(a) All phases

	Joy		Anger		Disgust		Surprise		Fear		Sadness		Contempt	
	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD
Remote	82.28	30.39	19.69	25.45	20.86	30.77	16.57	23.01	19.33	17.56	24.51	27.95	50.21	43.50
Lab	99.11	.59	20.50	38.44	15.99	23.65	12.27	17.45	5.21	4.16	30.31	37.99	44.18	13.97
Welch's t test	- few cases -		t(5.394)=-.051 p=.961		t(45.138)=1.134 p=.263		t(27.371)=1.157 p=.257		t(7.057)=4.256 p=.004*		t(5.348)=-.368 p=.727		t(9.967)=-.029 p=.977	
Effect size									0.81					

(b) Introduction phase

	Joy		Anger		Disgust		Surprise		Fear		Sadness		Contempt	
	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD
Remote	74.73	37.08	19.07	26.77	16.65	27.48	16.92	24.06	18.37	19.99	23.95	29.35	51.49	44.80
Lab	75.92	35.87	18.21	26.84	20.80	31.58	13.47	20.67	10.46	13.17	18.25	26.93	53.36	43.58
Welch's t test	t(149.138)=-.276 p=.783		t(155.556)=.305 p=.761		t(660.87)=-2.676 p=.008*		t(785.153)=3.34 p=.001*		t(65.580)=3.514 p=.001*		t(202.648)=2.201 p=.029*		t(360.236)=-.614 p=.539	
Effect size					0.46		0.15		0.41		0.20			

(c) Task execution phase

	Joy		Anger		Disgust		Surprise		Fear		Sadness		Contempt	
	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD
Remote	80.10	31.36	14.75	23.64	18.51	29.05	17.66	24.32	18.89	20.99	26.40	31.14	46.56	44.46
Lab	74.53	37.90	14.88	24.48	15.37	25.28	13.83	21.41	7.01	9.95	16.07	22.74	48.94	44.72
Welch's t test	t(92.278)=1.037 p=.302		t(312.969)=-.055 p=.956		t(974.948)=2.303 p=.021*		t(935.06)=3.28 p=.001*		t(20.229)=3.544 p=.002*		t(402.779)=3.855 p=.000*		t(471.301)=-.669 p=.504	
Effect size					0.11		0.16		0.58		0.36			

(d) Questionnaire filling phase

[P6] "The procedure for the test was very simple, and I found all the steps very clear"

[P66] "I really appreciated the ease of use of the tool to perform the test"

[P19] "I found it important to run the test when it was more comfortable for me, directly from my home"

Theme 4. *Feelings about the absence of a facilitator in the asynchronous remote tests*. A few participants highlighted the absence of a human facilitator during the test execution as a negative aspect.

[P44] "It would have been useful sometimes to ask an expert some questions about the task execution"

[P27] "I found it helpful to be able to run the test from home when it was more convenient, but running it without the support of a technician made me feel a bit lost because the procedures are not so familiar"

Theme 5. *Boring waiting times in the asynchronous remote tests*. At the end of each task execution, eGLU-Box PA sends all the multimedia files to the server. For a task 5 minutes long, the multimedia files (screen, webcam and microphone recording) are approximately 20 Mb large. In the case of regular or fast connections, the upload requires a few seconds; however, in the case of slower connections, it can require up to a couple of minutes.

[P25] "The waiting time at the end of each task was very annoying"

[P66] "I had to wait 3-4 minutes at the end of some tasks, and this slowed down the test too much"

H. Discussion

This section discusses the similarities and differences of asynchronous remote usability tests with respect to real-time in-lab tests, as emerged from the comparison study we performed. As reported in the following, it was found that the use of the web-based tool for asynchronous remote testing affects both the metrics results (RQ1) and the experience of the participants (RQ2).

Regarding the test metrics, task time and participant's satisfaction computed by SUS were affected. The statistical analysis highlighted that task time in the remote test was greater than that in the lab test. To provide an integrated environment in which the website to test can be opened, eGLU-Box PA

integrates an Apache reverse proxy that overcomes some technical constraints such as CORS[4]. When opening a webpage, this mechanism introduces a delay of approximately 0.5 seconds (we performed some measurements). Even if this delay seems very short and cannot be perceived by the users when opening a webpage, it becomes significant for task time when several web pages must be opened to complete a task. From our logs, we have seen that an average of 9 webpages were visited for each task; thus, we can consider an average delay of approximately 5 seconds per task. In the study, the average time to complete tasks in the lab was 36.51 seconds, while in the remote test, it was 41.28 seconds, i.e., approximately 12% slower. The t test revealed this difference to be statistically significant, even if the very small effect indicates that the difference is not critical. Our result highlights that the adoption of web technologies increases task execution time, in contrast to what emerged when previous technologies were used for performing asynchronous remote tests [15, 22-24]. Considering that the selected tasks on the tested websites do not involve the navigation of many pages, this small time influenced the overall task execution time, but it appeared acceptable to the participants.

The SUS score was significantly lower in the remote test ($p = .049$), showing that it was negatively affected, with a small effect. The analysis of the two SUS factors, learnability and usability, showed that this difference was mainly due to learnability, which was lower in the remote test ($p = .000$), with a medium effect. This is supported by the results of the other two questionnaires UMUX-lite and NPS, which do not include learnability. Indeed, UMUX-lite and NPS do not show any difference in their scores in the two conditions. The analysis of both the videos of the interactions and the participants' comments indicates that even if the participants perceived eGLU-Box PA to be usable (Theme 3 “High usability of the tool for remote test”), the double interaction with both the tool for asynchronous remote testing and the tested website overloads the participants. Regarding the overall experience of the participants when performing the test procedure (RQ2), the analysis of their emotions revealed that three emotions were higher in the remote test, i.e., fear, sadness and surprise. According to the Ekman model [46], fear is induced by perceived danger or threat; sadness is caused by feelings of disadvantage, loss, helplessness and disappointment; and surprise is the result of an unexpected event and can have either a positive or a negative valence. Because the other two dimensions are negative, we can safely assume that surprise has a negative valence. The participants’ comments may provide some reasons for explaining these emotions, in particular the ones summarized in Theme 4 “Feelings about the absence of a facilitator in the remote test”, which are in line with some suggestions in the literature (e.g., [39]). More indications come from the analysis of the three phases of the usability test, namely, introduction, task execution and questionnaire filling. In the introduction phase, only fear was higher in the remote test, with a high effect, which underlines a significant difference. During task execution, fear, sadness and surprise were higher in the remote test, while disgust was lower, with a medium effect. Disgust is felt when a person sees, touches, hears, or tastes something nasty or repulsive, experienced by taste, smell, touch, or vision. The lower values of disgust in the remote test might be due to the usability of eGLU-Box PA (Theme 3 “High usability of the tool for the remote test”), which makes the task accomplishment pleasant and thus less “disgusting”. The last phase of the test, questionnaire filling, is in line with the results of task execution for fear, sadness, surprise, but disgust was also higher in the remote test, possibly due to a negative impact of the request to complete the questionnaires. It is worth noting that in the lab test, the participants worked in the presence of the experimenter; thus, the lower values in the lab test of the above negative emotions may be due to the well-known tendency to fulfill the social expectations of the experimenter [12].

The workload measured through NASA-TLX was higher in the remote test ($p = .008$), and it had a large effect. The analysis of the six NASA-TLX subscales highlighted that performance and frustration determined a higher workload with a medium effect. The performance subdimension, measured by the item “How successful were you in accomplishing what you were asked to do?”, highlighted that the participants felt less successful in accomplishing the remote test. One reason could be that, in the lab test, the participants felt confident that the experimenter would advise them if they were not performing well. Concerning frustration, measured by the item “How insecure, discouraged, irritated, stressed, and annoyed were you?”, this result is perfectly in line with the detected emotions discussed above. A cause of this negative result could be the absence of a facilitator. Actually, there are mixed concerns about the presence of the experimenter during usability tests, as highlighted in studies in the literature. In our study, some participants remarked on the pressure they felt in the presence of the evaluator in the lab test (Theme 2 “Under pressure during the real-time in-lab tests”). Referring to both subdimensions, a concurrent cause could be the long waiting times at the end of each task, as indicated by some participants (Theme 5 “Boring waiting times in the asynchronous remote tests”).

I. Study limitations

We are aware that the performed study has the following limitations. Only one representative of the advanced tools for the asynchronous remote test, eGLU-Box PA, has been used. Future studies should be performed using other tools. External validity can be improved by considering additional websites and tasks. Also, despite the efforts made to balance gender, most of the participants who were eventually involved in the tests were male. Another limitation is the missing comparison between tools developed with modern web-based technology and old technologies, as well as between synchronous and asynchronous remote tests. Moreover, the number and severity of usability problems have not been addressed. Finally, due to the necessary anonymization of participants’ test data, it was

[4] https://www.w3.org/wiki/CORS

not possible to deeply analyze the extent to which slower internet connections affect the user experience, especially during the task execution phase that requires stable and fast connections.

IV. Lessons Learned

By exploiting advanced web technology, software tools for asynchronous remote testing add to the known benefits of asynchronous remote testing in gathering a greater amount of both qualitative and quantitative data. However, this study revealed that this introduces other problems. The lessons learned in the study are now presented to inform the design of future tools that should make asynchronous remote tests more effective. Although the study was performed with eGLU-Box PA, we discuss features that are common to most recently developed tools. The lessons learned are listed here and then briefly discussed:

- **The current web technology affects task execution time**
- **Slow internet connections must be managed**
- **Pay attention to possible interference with website usability**
- **The facilitator presence matters**

The current web technology affects task execution time. Recent tools are web applications running on a web browser. On the one hand, this choice simplifies both participant recruitment and test activities; on the other hand, it poses challenges related to internet security policies, i.e., CORS constraints[5] and opening an HTTP website under an HTTPS connection[5]. In eGLU-Box PA, this has been solved by using a reverse proxy, which is a state-of-the-art solution. Possible alternative solutions are used by other web-based tools [25-28], such as ad hoc browsers or browser plugins, which still introduce a delay in opening the web pages of the tested application and require the installation of software on the participant's personal computer. A reverse proxy, or similar solutions, also has a cost: it introduces a delay that increases task execution time, even if, as discussed in Section 4.7, the very low effect of this delay can be ignored in tasks that do not require access to many pages. However, this aspect must be taken into account in the case of tasks that require the navigation of several web pages and when time is a critical dimension (e.g., in the case of comparative studies). One way to collect more precise task times that are not affected by this delay and thus reflect the time spent by the users without external delays is to estimate the delay introduced by the reverse proxy (or by the browser plugin or ad hoc browsers) and to subtract it from each task time to compute the actual task time.

Slow internet connections must be managed. It is widely known that usability tests must be no longer than 45-60 minutes to avoid boring and tiring the participants, who do not pay enough attention to their work if the test lasts too long [54]. Evaluators, when designing the test, must be sure that performing the tasks and completing the questionnaire does not take too long. However, the study has shown that the tool for asynchronous remote testing requires time to transfer large multimedia files to the server. In eGLU-Box PA, such files are transferred at the end of each task. This harms those participants who do not have a stable, fast internet connection, as the waiting time can be several minutes before the participants can start a new task. We were aware of this problem; thus, we set the audio and video quality as low as possible to minimize the amount of data to be sent, and in the invitation email, we asked the participants to make sure they had a fast internet connection (upload at least 20 Mb/s) and provided a link for measuring the network speed. Unfortunately, not all participants followed our indications. We can safely assume that this is not a limitation of eGLU-Box PA itself but of all the recent web-based tools that need to upload the recorded data to their webserver. Future tools should integrate the check of the participants' internet connection speed or other mitigation strategies.

Pay attention to possible interference with website usability. The use of a tool for asynchronous remote testing alters not only user performance (task execution time) but also other usability dimensions of the tested website. The results of the SUS questionnaire revealed an important interference in learnability. Indeed, in the real-time in-lab test, the participants interacted with only the tested website; however, in the asynchronous remote test, they also interacted with the tool. Thus, they had to learn how to interact not only with the tested website but also with the asynchronous remote testing tool. This requires extra effort, even if eGLU-Box PA was indicated to be a very usable tool by the study participants and by the participants of previous studies (see, e.g., [37]). Evaluators should take into account possible interference with usability and learnability of evaluated website, particularly when, as in the case of time (see previous paragraph), these dimensions are critical in the usability evaluation. Tools for asynchronous remote usability testing might consider the possibility of including a training session on the use of the tool before the execution of the test to minimize this effect.

The facilitator presence matters. In most cases, the participants were interacting with the tool for asynchronous remote testing for the first time; thus, participants might need help resolving doubts or solving problems with the tool or even with the website to be tested. The study showed that some negative emotions were significantly higher in the asynchronous remote test (fear, sadness, surprise), in particular fear with a medium/large effect. Moreover, the workload was significantly higher in the asynchronous remote test with a very large effect because, as revealed by two specific subscales of NASA-TLX, the participants felt less successful in accomplishing asynchronous remote testing and more discouraged, stressed or irritated. As remarked in the discussion, in the real-time in-lab test the participants might feel confident that the experimenter would advise them if they were not performing well. Thus, the lack of a facilitator in asynchronous remote testing could contribute to the participants feeling less successful, and more insecure, discouraged, irritated and stressed. This was also suggested by some

[5] https://www.w3.org/TR/mixed-content/

comments of the participants that they would have liked to be guided and assisted by a human facilitator, not only to ask for help but also to feel more confident in case of need. Implementing a chatbot in the tool seems promising to support participants in asynchronous remote testing and could be a substitute for a facilitator in answering participants' questions. The chatbot could also provide useful feedback to participants about the accomplished tasks, thus improving their self-confidence.

V. Conclusions

Since the late 1990s, several tools have been developed to support evaluators in performing remote usability testing of websites. Most of them appeared in the first decade of this millennium. Then, there was a gap until advanced web technology recently enabled the creation of more powerful tools that offer innovative features and claim to further facilitate the execution of asynchronous remote tests and improve their effectiveness. To the best of our knowledge, no study has been performed to understand how these tools influence asynchronous remote testing. This article presented a study that compared real-time in-lab usability tests with asynchronous remote tests performed with a recently developed tool. It investigated whether and how the possibilities offered by the tool affect the outcomes of the usability tests as well as the emotions, workload and overall experience of the participants during the test. The results showed that the innovative features of a modern tool are not a panacea for asynchronous remote testing; conversely, they highlighted some critical aspects that the designers of these tools should be aware of.

An important finding derives from the analysis of the participants' emotions; it provides empirical evidence that asynchronous remote test participants, acting without the support of a facilitator, are more prone to feel negative emotions, also due to missing human support. However, echoing the literature (see, e.g., [39]), our study confirms that in-lab test participants sometimes feel pressured by the presence of the facilitator. These contrasting results deserve more attention. Future work includes the execution of an experimental study comparing real-time in-lab tests versus synchronous remote tests (where the facilitator is at a distance) and asynchronous tests, where the facilitator is not present.

As a further research direction focusing on the role of facilitators, we are currently working on integrating a chatbot into a tool supporting asynchronous remote tests [55]; the chatbot assists participants as the facilitator does in real-time in-lab tests. Then, we plan to use this new tool for a new comparison study, whose underlying hypothesis is that participants of asynchronous remote tests will act more comfortably even without the physical presence of a facilitator.

References

[1] M. M. Sebrechts and J. B. Black (2008). Software psychology: Human factors in computer and information systems. Ben Shneiderman. Cambridge, Mass.: Winthrop, 1980. Pp. xiv + 320. *Applied Psycholinguistics*, 3(4), 373-381.

[2] R. Lanzilotti, C. Ardito, M. F. Costabile, and A. De Angeli (2011). Do patterns help novice evaluators? A comparative study. *International journal of human-computer studies*, 69(1-2), 52-69.

[3] J. Rubin and D. Chisnell 2008. *Handbook of usability testing: how to plan, design and conduct effective tests*. John Wiley & Sons.

[4] J. S. Dumas and J. C. Redish 1999. *A Practical Guide to Usability Testing*. Intellect Books.

[5] C. Ardito, P. Buono, D. Caivano, M. F. Costabile, and R. Lanzilotti (2014). Investigating and promoting UX practice in industry: An experimental study. *International Journal of Human-Computer Studies*, 72(6), 542-551.

[6] A. Fernandez, E. Insfran, and S. Abrahão (2011). Usability evaluation methods for the web: A systematic mapping study. *Information and software Technology*, 53(8), 789-817.

[7] G. Salvendy 2012. *Handbook of human factors and ergonomics*. John Wiley & Sons.

[8] J. Sauer, K. Seibel, and B. Rüttinger (2010). The influence of user expertise and prototype fidelity in usability tests. *Applied Ergonomics*, 41(1), 130-140.

[9] H. R. Hartson, J. C. Castillo, J. Kelso, and W. C. Neale (1996). Remote evaluation: the network as an extension of the usability laboratory. *Proc. of the International Conference on Human Factors in Computing Systems (CHI '96)*. ACM, 228–235.

[10] M. Hammontree, P. Weiler, and N. Nayak (1994). Remote usability testing. *interactions*, 1(3), 21–25.

[11] P. V. Selvaraj, "Comparative study of synchronous remote and traditional in-lab usability evaluation methods," Virginia Tech, 2004.

[12] J. Sauer, A. Sonderegger, K. Heyden, J. Biller, J. Klotz, and A. Uebelbacher (2019). Extra-laboratorial usability tests: An empirical comparison of remote and classical field testing with lab testing. *Applied Ergonomics*, 74, 85-96.

[13] A. J. B. Brush, M. Ames, and J. Davis (2004). A comparison of synchronous remote and local usability studies for an expert interface. *Proc. of the International Conference on Human Factors in Computing Systems -Extended Abstracts (CHI '04)*. ACM, 1179–1182.

[14] K. E. Thompson, E. P. Rozanski, and A. R. Haake (2004). Here, there, anywhere: remote usability testing that works. *Proc. of the Conference on Information Technology Education (ITE '04)*. ACM, 132–137.

[15] T. Tullis, S. Fleischman, M. McNulty, C. Cianchette, and M. Bergel (2002). An empirical comparison of lab and remote usability testing of web sites. *Proc. of the Usability Professionals Association Conference (UPA '02)*.

[16] S. Waterson, J. A. Landay, and T. Matthews (2002). In the lab and out in the wild: remote web usability testing for mobile devices. *Proc. of the International Conference on Human Factors in Computing Systems - Extended Abstracts (CHI '02)*. ACM, 796–797.

[17] A. Edmonds (2003). Uzilla: A new tool for Web usability testing. *Behavior Research Methods, Instruments, & Computers*, 35(2), 194-201.

[18] L. Paganelli and F. Paternò (2003). Tools for remote usability evaluation of Web applications through browser logs and task models. *Behavior Research Methods, Instruments, & Computers*, 35(3), 369-378.

[19] T. Carta, F. Paternò, and V. F. de Santana (2011). Web Usability Probe: A Tool for Supporting Remote Usability Evaluation of Web Sites. *Proc. of the IFIP Conference on Human-Computer Interaction (INTERACT '11)*. Springer Berlin Heidelberg, 349-357.

[20] A. Baravalle and V. Lanfranchi (2003). Remote Web usability testing. *Behavior Research Methods, Instruments, & Computers*, 35(3), 364-368.

[21] A. S. Alghamdi, A. Al-Badi, R. Alroobaea, and P. Mayhew (2013). A comparative study of synchronous and asynchronous remote usability testing methods. *International Review of Basic and Applied Sciences*, 1(3), 61-97.

[22] M. S. Andreasen, H. V. Nielsen, S. O. Schrøder, and J. Stage (2007). What happened to remote usability testing? an empirical study of three methods. *Proc. of the International Conference on Human Factors in Computing Systems (CHI '07)*. ACM, 1405–1414.

[23] C. Andrzejczak and D. Liu (2010). The effect of testing location on usability testing performance, participant stress levels, and subjective testing experience. *Journal of Systems and Software*, 83(7), 1258-1266.

[24] R. West and K. Lehman (2006). Automated summative usability studies: an empirical evaluation. *Proc. of the International Conference on Human Factors in Computing Systems (CHI '06)*. ACM, 631–639.

[25] Loop11. *Loop11 User Testing*. Retrieved from https://www.loop11.com/ Last Access July 14th, 2022.

[26] Lookback_2020. *Lookback*. Retrieved from https://lookback.io/ Last Access July 14th, 2022.

[27] Userlytics. *Userlytics: Remote User Testing Platform*. Retrieved from https://www.userlytics.com/ Last Access July 14th, 2022.

[28] UserTesting. *UserTesting: The Human Insight Platform*. Retrieved from https://www.usertesting.com/ Last Access July 14th, 2022.

[29] S. Federici, M. L. Mele, M. Bracalenti, A. Buttafuoco, R. Lanzilotti, and G. Desolda (2019). Bio-behavioral and self-report user experience evaluation of a usability assessment platform (UTAssistant). *Proc. of the International Conference on Human Computer Interaction Theory and Applications (HUCAPP '19)*, 73518.

[30] G. M. Burghardt *et al.* (2012). Perspectives–minimizing observer bias in behavioral studies: a review and recommendations. *Ethology*, 118(6), 511-517.

[31] J. Scholtz (2004). Usability evaluation. *National Institute of Standards and Technology*, 1.

[32] J. I. Hong, J. Heer, S. Waterson, and J. A. Landay (2001). WebQuilt: A proxy-based approach to remote web usability testing. *ACM Transaction on Information Systems*, 19(3), 263–285.

[33] M. A. A. Winckler, C. M. D. S. Freitas, and J. V. d. Lima (2000). Usability remote evaluation for WWW. *Proc. of the International Conference on Human Factors in Computing Systems - Extended Abstracts (CHI '00)*. ACM, 131–132.

[34] T. Corporation. *Morae*. Retrieved from https://www.techsmith.com/morae.html Last Access July 10th, 2022.

[35] F. Paternò, L. Paganelli, and C. Santoro (2001). Models, tools and transformations for design and evaluation of interactive applications. *Proceedings of PC-HCI*, 23-28.

[36] UserTesting. *MISE receives the ISO/IEC 25000 certificate for their software producteGLU-box PA*. Retrieved from https://iso25000.com/index.php/en/news/190-mise-receives-the-iso-iec-25000-certificate-for-their-software-producteglu-box-pa Last Access July 13th, 2022.

[37] S. Federici *et al.* (2019). Heuristic Evaluation of eGLU-Box: A Semi-automatic Usability Evaluation Tool for Public Administrations. *Proc. of the International Conference on Human-Computer Interaction (HCII '19)*. Springer International Publishing, 75-86.

[38] J. R. Lewis (2006). Usability testing. *Handbook of human factors and ergonomics*, 12, e30.

[39] K. C. Madathil and J. S. Greenstein (2011). Synchronous remote usability testing: a new approach facilitated by virtual worlds. *Proc. of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, 2225–2234.

[40] S. Federici, M. L. Mele, M. Bracalenti, A. Buttafuoco, R. Lanzilotti, and G. Desolda (2019). Bio-behavioral and Self-Report User Experience Evaluation of a Usability Assessment Platform (UTAssistant). *Proc. of the International Conference on Human Computer Interaction Theory and Applications (HUCAPP '19)*.

[41] S. Federici *et al.* (2018). UX Evaluation Design of UTAssistant: A New Usability Testing Support Tool for Italian Public Administrations. *Proc. of the International Conference on Human-Computer Interaction (HCII '18)*. Springer International Publishing, Lecture Notes in Computer Science, 55-67.

[42] I. Etikan, S. A. Musa, and R. S. Alkassim (2016). Comparison of convenience sampling and purposive sampling. *American journal of theoretical and applied statistics*, 5(1), 1-4.

[43] J. Brooke (1996). SUS-A quick and dirty usability scale. *Usability evaluation in industry*, 189(194), 4-7.

[44] D. B. Grisaffe (2007). Questions about the ultimate question: conceptual considerations in evaluating Reichheld's net promoter score (NPS). *Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior*, 20, 36.

[45] K. Finstad (2010). The usability metric for user experience. *Interacting with Computers*, 22(5), 323-327.

[46] R. Ekman 1997. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA.

[47] S. Stöckli, M. Schulte-Mecklenbeck, S. Borer, and A. C. Samson (2018). Facial expression analysis with AFFDEX and FACET: A validation study. *Behavior Research Methods*, 50(4), 1446-1460.

[48] R. W. Taggart, M. Dressler, P. Kumar, S. Khan, and J. F. Coppola (2016). Determining emotions via facial expression analysis software. *Proc. of the Student-Faculty Research Day (CSIS)*. Pace University.

[49] L. Kulke, D. Feyerabend, and A. Schacht (2020). A Comparison of the Affectiva iMotions Facial Expression Analysis Software With EMG for Identifying Facial Expressions of Emotion. *Frontiers in Psychology*, 11(329).

[50] J. Cohen (1988). The effect size. *Statistical power analysis for the behavioral sciences*, 77-83.

[51] V. Braun and V. Clarke (2006). Using Thematic Analysis *Psychology Qualitative Research in Psychology*, 3(2), 77-101.

[52] J. Lewis and J. Sauro (2009). The Factor Structure of the System Usability Scale. In: M. Kurosu Ed. *Human Centered Design - HCD 2009.* LNCS, Vol. 5619, Springer Berlin Heidelberg, 94-103.

[53] S. G. Hart (2006). Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Human Factors and Ergonomics Society Annual Meeting*, 50(9), 904-908.

[54] J. Preece, H. Sharp, and Y. Rogers 2019. *Interaction Design-beyond human-computer interaction*. John Wiley & Sons.

[55] S. Federici *et al.* (2021). A Chatbot Solution for eGLU-Box Pro: The Usability Evaluation Platform for Italian Public Administrations. *Proc. of the International Conference on Human-Computer Interaction (HCII '21)*. Springer International Publishing, 268-279.