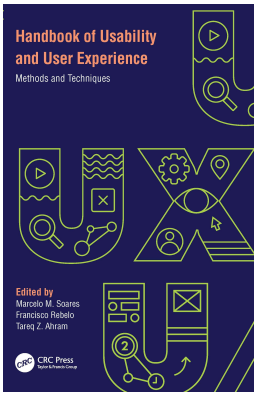

Handbook of Usability and User Experience Methods and Techniques

Marcelo M. Soares, Francisco Rebelo, Tareq Z. Ahram

Remote Usability Testing

Publication details
https://test.routledgehandbooks.com/doi/10.1201/9780429343490-5
J. M. Christian Bastien, Kevin Falzone
Published online on: 13 May 2022

4 Remote Usability Testing

J. M. Christian Bastien and Kevin Falzone

CONTENTS

4.1 INTRODUCTION

Usability evaluation is an essential step in the user-centered design cycle (International Organization for Standardization, 2019). For usability evaluation, different approaches and methods are available: model-based evaluations (Kieras, 2012), inspection-based evaluations (Cockton, Woolrych, Hornbæk, & Frøkjær, 2012) and user testing (Dumas & Fox, 2012; Lewis, 2012). This latter method is probably the most documented one. There are countless articles and books on it (Barnum, 2020; Dumas & Redish, 1999; Rubin & Chisnell, 2008; Tullis & Albert, 2013).

With the Internet, remote usability testing has gained popularity, especially for testing Web sites. In remote user tests, researchers and participants are in different locations and participants use their own hardware and software. This is made possible by different technologies.

The aim of this chapter is to present a state of the art in remote usability testing. The differences between the two approaches in terms of methodology and tools, advantages and drawbacks of each will be addressed. Before presenting the state of the art in remote usability testing and allowing the comparison between in situ user tests and remote testing, we will describe how the user tests are prepared and conducted in the traditional lab.

DOI: 10.1201/9780429343490-5

4.2 PREPARING AND CONDUCTING A USER TEST

To assess the usability of interactive systems, experts collect behavioral, physiological and self-reported data (Bergstrom & Schall, 2014; Sauro & Lewis, 2016; Tullis & Albert, 2013). But before capturing these data, experts have to go through the following steps for preparing the test (Bastien, 2010):

- The definition of the test objectives,
- The selection and recruitment of test participants,
- The selection of tasks that participants will be asked to perform,
- The creation and description of task scenarios,
- The choice of the measures and the way the data will be recorded,
- The preparation of the test material and test environment (the usability laboratory),
- The choice of the tester and the design of the test protocol per se (instructions, design protocol, etc.),
- The selection of satisfaction questionnaires,
- The analysis of the data,
- The presentation and communication of the test results.

Some of these steps, as they take place in a usability lab, are detailed in the following sections.

4.2.1 The Definition of Test Objectives

The design flaws may be identified during the development cycle of an interactive system or when the system is released. In the first case, we talk about formative evaluations which are conducted at the beginning of the design process and continue until the final system is released. In other words, evaluations are usually performed with each new version of the underdevelopment system in order to identify and fix usability problems. These evaluations end when predefined criteria are met (e.g., a planned number of iterations, a percentage of successful tasks, etc.).

In the second case, summative evaluations are intended to assess the final system in order to measure its performance, to validate that the system meets a set of requirement criteria and to benchmark the system to previous versions or to competing products.

4.2.2 The Selection and Recruitment of Test Participants

Tests participants should be representative of the end users in terms of characteristics, knowledge and skills. The number of users that need to be mobilized is an issue that has been addressed by several authors. Early studies concluded that five users were sufficient to identify 80–85% of usability problems (Lewis, 1994; Nielsen & Landauer, 1993; Virzi, 1990, 1992).

However, more recent studies have indicated that five users could not be sufficient. For example, Spool and Schroeder (2001) report that 35% of usability problems were

found with the first five users, but that critical usability problems were found from the 13th and 15th test participants. Faulkner (2003) also conducted a study involving 60 users who were asked to complete a computerized timesheet. The author found that the risk of relying on a group of five users could cause half of the usability problems to be missed.

4.2.3 The Definition of Task Scenarios

During user testing, participants are usually asked to perform predefined tasks. These tasks are selected according to several criteria such as the objectives of the test or research hypotheses, the end user's goals, the frequency with which they are performed by the end users, the areas of the system where there may be potential usability problems, the system's business objectives, the results of a previous test or inspection methods or the new functionalities which have to be tested.

Following the selection and definition of the tasks, the experts have to write instructions and test their comprehensibility. When writing the instructions, experts will keep in mind that users are invited to achieve specific objectives and not to follow a succession of actions.

4.2.4 The Choice of Measures, Their Analyses and Their Representation

During a user test, experts collect specific information related to effectiveness and efficiency of the interaction and also satisfaction with the system, which are the characteristics of usability (International Organization for Standardization, 2018). The data can be classified into three categories:

Behavioral data. The most commonly used behavioral data collected during a user test are the task status (i.e., success or failure), task duration, error rate and physical or/and cognitive efforts. In the context of Web sites, clickstreams and the lostness metric (Smith, 1996) can also be collected and computed.

The task status is a way of reporting whether or not the user has completed the task and to what extent. From this task status, several analyses and representations can be produced: the ratio of successes (or failures), the average calculation of the tasks (i.e., the number of participants who succeeded or failed in completing the task) and the average calculation for each participant (i.e., the number of tasks that the participant was able to complete or fail to complete).

The time on task is the time elapsed between the start of a task and its completion. Duration of the task can be calculated and presented not only for each task (i.e., the average time the user takes to complete the task) but also for each user (i.e., the average time the user takes to complete all tasks).

Errors are the actions that can cause the task to fail. Errors can be calculated and presented as error rate per task per user or for each task per user. It is also possible to assign a severity score (low, medium, high) and then calculate the frequency for each category of errors. These errors are generally due to usability problems.

Physical efforts refer to the physical activity required to perform the task, whereas cognitive efforts are the mental resources involved in responding to a particular task.

They can be analyzed by counting the number of actions, such as clicks, performed to complete the task. Usually, experts calculate an average number of actions for each task (per participant).

Clickstreams illustrate the paths taken (by users) on a Web site. It can be used to identify the pages that participants go through in their search for information and to calculate the percentage of participants taking each route (e.g., Figure 4.1). It allows assessing the heading and links and their relation to their content.

The lostness metric (Smith, 1996) indicates whether or not users are lost on a Web site. The coefficient is calculated from three elements: (1) *N*: the number of unique Web pages that were visited during the task (pages that are visited several times only count once); (2) *S*: the total number of Web pages visited during the task, including page revisits; (3) *R*: the (optimal) number of pages that must be visited to complete the task. These three elements are then used in the following formula to calculate the lostness metric noted *L*:

$$L = \text{sqrt}((N / S - 1)^2 + (R / N - 1)^2).$$

Smith (1996) found that the score *L* of less than 0.4 shows no sign of being lost. In contrast, participants with a score above 0.5 appear disoriented. When the *L* value was calculated for each participant, it is possible to obtain an average score for each task.

The use of eye-tracking techniques makes it possible to know precisely where the participant's gaze lands throughout the test session (Bergstrom & Schall, 2014). This technique can be used, for example, to analyze cognitive processing, stimulation and interest using the user's pupillary response, or to determine whether the user, while browsing, correctly saw the Web link to successfully complete the task or whether the user took it into account but did not click on it. From the eye-tracking data, it is possible to obtain the following: (a) The scanpath which allows the visual representation of the participant's eye path on the interface. This technique takes into account two types of data: fixations and saccades. The fixations, which are a pause in the eye's movement over an area, are usually expressed in numbered circles, while saccades, which are brief, rapid eye movements, are represented by lines joining two fixations. (b) The heat map which is used to represent the eye movements of several participants on the Web page. The colors of the heat map indicate the density of eye fixations. Usually, the warmer the color, the higher the fixation density and, conversely, the colder the color, the lower the fixation density. The heat map is an excellent way to know which area(s) of the page is attracting more (or less) attention from participants. (c) The focus map which makes areas that have received the most visual attention transparent, while it darkens areas that have received little or no visual attention.

It is also possible to make analyses of eye-tracking data according to specific areas (called areas of interest [AOI]) that have been delineated by the experts. From these areas of interest, experts can have dwell times, number of fixations within an AOI, the sequence, time to first fixation, revisits, hit ratio, etc. From these analyses, two types of visualizations are possible: (1) binning charts which show the percentage of time spent

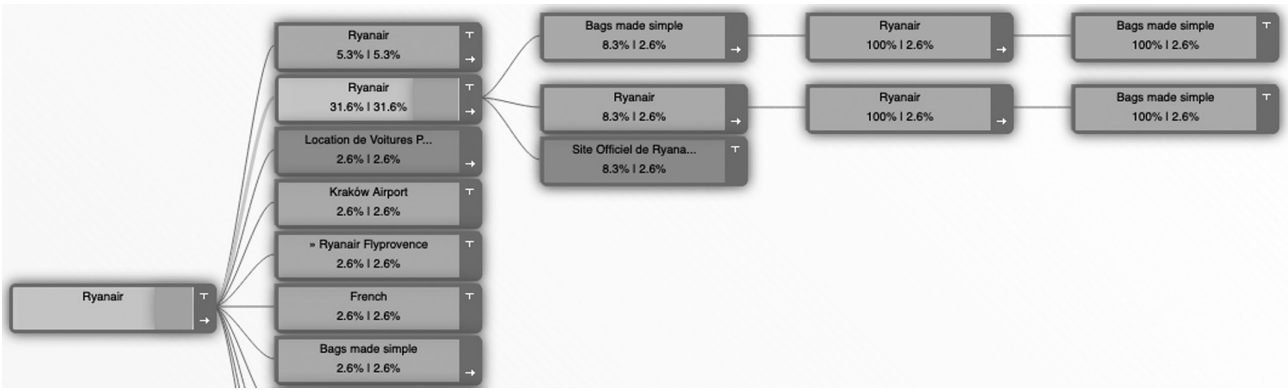

FIGURE 4.1 Schematic example of a clickstream taken from a test conducted with Evalyzer (www.evalyzer.com).

looking at each area of interest according to a time interval; (2) areas of interest grids which show the amount of visual attention given to cut out cells of equal size.

Physiological data. The most common physiological measures used in user tests are facial expressions, skin conductance and electroencephalography (EEG). Software such as FaceReader[1] can determine the emotional state of the participant based on his or her facial expressions (based on the taxonomy of Ekman and Friesen (1975)). The conductance (or electrodermal activity) of the skin is measured using sensors to detect emotional activation or stimulation. Three types of activation exist: an increase in mental load, an increase in emotion/emotional state and/or an increase in physical activity. Emotional states associated with increased electrodermal activity include fear, anger and joy. Knowing the emotional states of participants during the user testing session can be useful not only in evaluating the user experience, but also in detecting usability problems.

Brain waves measured by electroencephalogram are associated with cognitive and emotional states. For example, they can detect states of activation or excitement or calm in users (Alfimtsev, Basarab, Devyatkov, & Levanov, 2015).

Self-reported data. Most of the time, self-reported data used in user testing are collected in three different ways: with the think-aloud protocol, with questionnaires and with standardized satisfaction questionnaires.

User verbalizations are collected by the think-aloud protocol method. The think-aloud protocol involves asking participants to think aloud while interacting with the system. Users are invited to express anything that comes to their mind, i.e., their ways of doing things, their opinions, their reactions and so on. Experts can have users perform verbalizations during the test session (which is called concurrent think-aloud protocol) or after the test session accompanied with a video recording of their performance (which is called retrospective think-aloud protocol).

Written and oral comments are collected with open-ended, closed-ended, single choice, multiple-choice questions, scales and ranking questions (e.g., Likert scale [Likert, 1932] and semantic differentiators (Osgood, Suci, & Tannenbaum, 1957)). In addition, from the 1980s onward, authors have developed satisfaction questionnaires. These questionnaires can be categorized globally by their number of items, the usability dimensions they evaluate, the scale format and the type of systems they are designed for. Some examples of satisfaction questionnaires are the ASQ (*After Scenario Questionnaire*) (Lewis, 1995), the AttrakDiff (Hassenzahl, Burmester, & Koller, 2003), the CSUQ (*Computer Usability Questionnaire*) (Lewis, 1995), the meCUE (Minge & Riedel, 2013), the PSSUQ (Post-Study System Usability Questionnaire) (Lewis, 2002), the PUTQ (Purdue Usability Testing Questionnaire) (Lin, Choong, & Salvendy, 1997), the QUIS (*Questionnaire for User Interface Satisfaction*) (Chin, Diehl, & Norman, 1988), the SUMI (Software Usability Measurement Inventory) (Kirakowski & Corbett, 1993), the SUS (System Usability Scale) (Brooke, 1996), the UEQ (User Experience Questionnaire) (Laugwitz, Held, & Schrepp, 2008; Laugwitz, Schrepp, & Held, 2006), the UMUX (Usability Metric for User Experience) (Finstad, 2010), the UMUX-LITE (Usability Metric for User Experience) (Lewis, Utesch, & Maher, 2013) and the USE (Usefulness, Satisfaction and Ease of Use) (Lund, 2001).

Some satisfaction questionnaires are more dedicated to Web sites: e.g., DEEP (Design-oriented Evaluation of Perceived Usability) (Yang, Linder, & Bolchini, 2012), EUCS (End User Computing Satisfaction) (Doll & Torkzadeh, 1988), the perceived Web site usability measurement scale (Wang & Senecal, 2007), SUPR-Q (Sauro, 2015), the user-perceived Web quality instrument (Aladwani & Palvia, 2002) and WAMMI (Web Analysis and Measurement Inventory, www.wammi.com).

4.2.5 The Preparation of the Test Material and the Test Environment (Test Laboratory)

The structure of a test laboratory is usually composed of several rooms. Although the number of rooms differs from laboratory to laboratory, there is a minimum of two rooms (Nielsen, 1994): a testing room for the participant(s) and an observation room for usability professionals. In this first room, you will find materials for conducting the test, such as a computer, tablet or smartphone, for presenting the interface (Web site, application) to be evaluated. The instructions, usage scenarios and questionnaires are generally provided in paper or online version or eventually verbally given by a facilitator.

The test room also contains recording devices allowing the user's actions on the interface to be collected through software snapshots, cameras for observing the users, user verbalizations with the help of microphones and physiological measurements using specific devices (e.g., cardio-frequency meters and electrodermal sensors).

This room is usually separated by a one-way glass. The observation room contains equipment for the observation of the users and instruments to interact with the user (i.e., microphones).

Sometimes additional rooms are used in some laboratories (Figure 4.2): an observation room where additional people can observe the test without interfering with the users or assessors, a reception room and audiovisual control room.

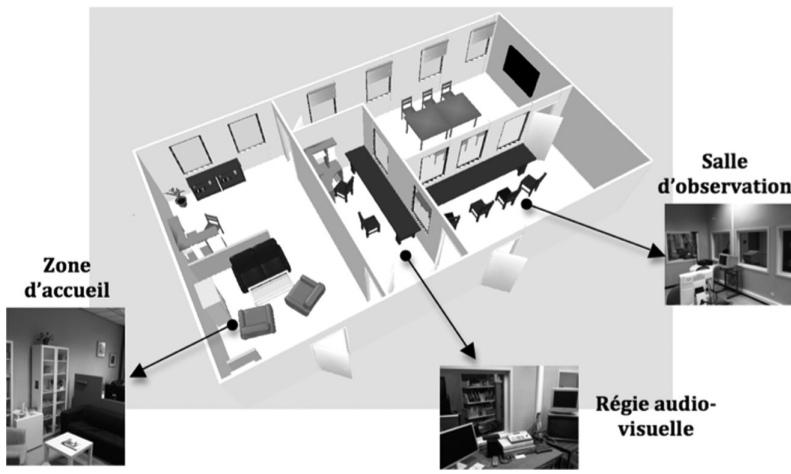

FIGURE 4.2 Pergolab user test laboratory (Metz, France).

4.2.6 The Presentation and Communication of Test Results

Although the presentation and communication of test results can be left to the choice of the usability professionals, a template (the CIF[2]) has been proposed (American National Standard, 2001; International Organization for Standardization, 2010). The purpose of this template is to standardize the writing of test reports.

4.3 REMOTE USER TESTING

In the previous sections, the preparation of the test session, as well as the tools used during the laboratory sessions, has been presented. In this and the following sections, we will address the remote user testing and the tools used. We will also ask ourselves if this kind of test allows gathering the same kind of data that are gathered in a laboratory and if the results of these tests compare to those obtained in a laboratory.

Over the past two decades or so, the Internet has made it possible to conduct user tests remotely. Thus, "the evaluator, performing observation and analysis, is separated in space and/or time from the user" (Hartson, Castillo, Kelso, & Neale, 1996). Remote user testing addresses several issues encountered in laboratory settings. In fact, remote testing does not require testers to travel to the testing labs, saving time and money. It also makes it possible to involve testers who are far away, and to reduce the costs of traditional lab. Moreover, users are in their natural environment (i.e., they use their own hardware and software). However, the remote user test may face some new issues such as installation and configuration of the software to perform the test as well as confidentiality issues (De Bleecker & Okoroji, 2018).

Two ways of conducting remote user testing have been used by experts: (1) User tests that involve the supervision of an evaluator (synchronous and moderated testing). (2) User tests that do not require the presence of a moderator (asynchronous and unmoderated).

4.3.1 Moderated User Testing

Moderated remote user testing was born in the 1990s, thanks to the development of information sharing and collaborative tools (Hammontree, Weiler, & Nayak, 1994). Like user testing labs, evaluators conduct the tests and collect information while users perform different actions on the system, but in this context evaluators are just geographically separated from users. Evaluators interact directly with test participants.

To perform user tests, three key elements are required (Dumas & Loring, 2008): the system being evaluated, a sharing application and a recording application.

Nowadays, evaluators usually use videoconferencing tools (e.g., WebEx,[3] Zoom,[4] etc.) in order to collect user screen actions, user verbalizations and user facial expressions. Commercial tools such as Lookback[5] and Loop11[6] can be used in moderated remote user testing.

4.3.2 Unmoderated User Testing

Unmoderated remote testing appeared in the late 1990s (Scholtz, Laskowski, & Downey, 1998). In unmoderated testing, the evaluators are physically and temporally separated from the users. In other words, evaluators are replaced by a platform which is in charge to conduct user tests and to collect data from users (and in some circumstances analyzes them). The advantages are that many users may participate at the same time, thus more participants can be recruited in a given period of time, reducing the duration of the test campaign. The test is also independent of time zones. In this situation, users are not influenced by the expert's comments or behaviors and the test situation is less impressive because the user is in a familiar environment at home or at work.

But there are some drawbacks. In fact, users may not be able to get assistance if needed. Experts cannot observe test participants while they are running the test and cannot interact with them. But some of these drawbacks can be mitigated depending on the technology used.

Over the last 20 years, unmoderated remote user testing tools have evolved significantly in terms of the architecture used to collect data, the type of data collected, the skills and amount of effort required by usability experts.

Three approaches have been adopted to conduct remote testing: server-based, proxy-based or client-based approach.

Server-based approaches are normally able to collect navigation data and even interaction data by adding some JavaScript code on Web pages which require access to the Web server. With this approach, it may be difficult to interpret users' actions, paths and goals. Nevertheless, a possible solution has been proposed to address these drawbacks. This solution combines users' actions data and subjective data collected through questionnaires in the same tool (Winckler, Freitas, & Valdeni de Lima, 2000).

Proxy-based approaches consist in adding an "intermediary" between the client which sends requests to obtain Web pages and the server which provides specific Web pages according to the requests. By being located between the two, the proxy can retrieve some data. Like server-based tools, this approach gathers navigation data (Hong & Landay, 2001), interaction data (Atterer & Schmidt, 2007; Atterer, Wnuk, & Schmidt, 2006; Baravalle & Lanfranchi, 2003) as well as subjective data (Baravalle & Lanfranchi, 2003). This approach solves the main issue related to the server-based approach, i.e., the access to the server.

Client-based approaches consist in using either an instrumented browser (e.g., Uzilla) (Edmonds, 2003) or browser plugins (e.g., Evalyzer[7] and Loop11[8]). Thanks to this approach, not only the above-mentioned data but also the user's actions on the Web browser (e.g., backward and forward buttons) can be recorded. With this approach, however, the user must have a compatible operating system or Web browser, the required privileges and the aptitude or the inclination to install a browser or plugins on his machine.

Recent commercial tools mainly use the plugin approach and (try to) integrate all the steps required to conduct user tests for reducing the amount of effort needed by usability experts (e.g., Evalyzer,[9] Lookback,[10] Loop11[11] and UserTesting[12]).

To a certain extent, these tools can manage the different steps that were described in the previous sections for preparing and conducting a user test. For instance, user selection and description can be done by a screening questionnaire. Evalyzer and UserTesting, for example, offer this functionality. All the platforms mentioned above allow defining tasks. However, given that the test is conducted without supervision, experts or evaluators cannot know when participants have succeeded a task or not. Thus, conditions of success must be defined in order for the tool to be able to calculate success rates and failures. Some of the tools provide this functionality (Evalyzer and Loop11). To our knowledge, only one of them allows the evaluator to randomize the task order (Evalyzer). At the end of the test, satisfaction questionnaires are provided as well as the possibility to develop different types of questions both after each task and at the end of the test (e.g., Evalyzer, Loop11 and UserTesting).

The measures, their analyses and their representation may vary depending on the platform. Behavioral data, self-reported data and user software and hardware environment information can also be collected. However, no platforms allow capturing physiological data. The use of Webcams has been attempted to record the position of the gaze on Web pages on the desktop of the remote participant, but the data collected is not very reliable (Chynał & Szymański, 2011) and the software used may compete with the remote testing application for the Webcam resources.

After collecting user data, some platforms provide automatic analyses and representations of the results. Individual results are provided as well as group statistics on each task and questionnaire responses. Thus, it is possible to know for each user the number of tasks on which he failed, the task duration, the efforts (e.g., number of clicks, scrolls, pages consulted), the pages consulted (i.e., clickstream) and the lostness coefficients. In addition, the results of standardized satisfaction questionnaire (e.g., the SUS) can be computed automatically. At the level of the tasks, we get the number of users which have failed on the task, the average duration of the task, the average effort on the task, the clickstream on the task and the average of the lostness metric.

These platforms allow not only exploring data and visualizing the analyses but also exporting data in spreadsheet format, the figures, and the video recordings of the test session which may contain the comments made by the participants and the participants' faces in picture-in-picture if the participants are allowed the use of the Webcam and the microphone. Finally, some platforms allow generating a PowerPoint report.

4.3.3 Comparing In Situ and Remote User Testing

Several studies have examined the effects of user test situations (i.e., laboratory testing, remote and supervised testing, remote and unsupervised testing) on dependent variables such as the number of usability issues identified and their severity, task duration, task completion, number of errors, satisfaction, etc.

As for the number of usability problems, five studies found that the lab situation and the remote testing situation were comparable (Andreasen, Nielsen, Schrøder, & Stage, 2007; Brush, Ames, & Davis, 2004; Chalil Madathil & Greenstein, 2017;

Hartson et al., 1996; Thompson, Rozanski, & Haake, 2004). In the same way, the severity of the problems found was similar in both situations (Andreasen et al., 2007; Brush et al., 2004; Chalil Madathil & Greenstein, 2017).

Among the four studies which have measured the task duration, three of them showed that there were no significant statistical differences (Andreasen et al., 2007; Brush et al., 2004; Chalil Madathil & Greenstein, 2017) and only one study showed that local participants took less time (Thompson et al., 2004). One of the studies reported that the setup and wrap-up of the test took significantly more time for remote moderated testing and the discussion was slightly longer in the laboratory condition (Brush et al., 2004).

One study reported that the task completion time was not different between the two test contexts (Andreasen et al., 2007) and another one reported that the number of errors was less for the local participants (Thompson et al., 2004).

Two studies measured satisfaction through questionnaires (Hartson et al., 1996; Thompson et al., 2004). They concluded that the results were comparable.

A point to note is that "the majority of participants felt that the remote condition was more convenient" and "half would prefer to be involved in remote studies over local studies in the future, while none preferred local over remote condition" (Brush et al., 2004).

Only one study compared the laboratory and the unmoderated remote testing (Tullis, Fleischman, McNulty, Cianchette, & Bergel, 2002), and in this study both configurations allowed identification of the most critical usability issues. Task completion rates and task duration were found comparable.

However, the results of these comparisons should be considered with caution as the technologies used are different.

4.4 CONCLUSION

The aim of this chapter was to give the reader an overview of the state of the art in remote usability testing. To do so, we first presented the different steps an evaluator has to go through in order to prepare and conduct the usability test and to analyze the data captured. These steps were described as they are usually conducted in a local usability laboratory. They were thus used as a reference. We then gave a brief overview of the evolution of the technologies used to conduct remote usability tests and presented some commercial platforms. Studies comparing the local laboratory test session with the remote testing situation were then presented. We conclude that with recent technologies, the data that can be obtained from remote testing is no longer different from the data captured in a local usability laboratory except for physiological data and eye-tracking recordings. But remote usability platforms which allow conducting usability tests remotely can also be used in a local laboratory so as to complete the test campaign with other kinds of measures. What emerges from these comparisons and comparative analyses of the platforms is that the most recent ones use plugin technologies which allow the evaluators to capture performance data (quantitative) as well as subjective data. The platforms differ from one another in the way they allow, for example, managing the tasks presented to the users as well as

the way they analyze the data. It can be expected that these will continue to evolve and will make it possible to integrate other measurement tools whether in a remote situation or as a complement to it.

NOTES

1. https://www.noldus.com/facereader
2. Common Industry Format (for usability test reports).
3. https://www.webex.com/
4. https://zoom.us/
5. https://lookback.io/
6. https://www.loop11.com/
7. https://www.evalyzer.com/
8. https://www.loop11.com
9. https://www.evalyzer.com/
10. https://www.lookback.com/
11. https://www.loop11.com/
12. https://www.usertesting.com/

REFERENCES

Aladwani, A. M., & Palvia, P. C. (2002). Developing and validating an instrument for measuring user-perceived web quality. *Information & Management*, *39*(6), 467–476. doi:10.1016/S0378-7206(01)00113-6

Alfimtsev, A. N., Basarab, M. A., Devyatkov, V. V., & Levanov, A. A. (2015). A new methodology of usability testing on the base of the analysis of user's electroencephalogram. *Journal of Computer Sciences and Applications*, *3*(5), 105–111. doi:10.12691/jcsa-3-5-1

American National Standards Institute. (2001). *Common Industry Format for Usability Test Reports (Formerly ANSI INCITS 354-2001).* New-York, NY: American National Standards Institute, Inc.

Andreasen, M. S., Nielsen, H. V., Schrøder, S. O., & Stage, J. (2007). What happened to remote usability testing? An empirical study of three methods. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, San Jose, California, USA. https://doi.org/10.1145/1240624.1240838.

Atterer, R., & Schmidt, A. (2007). Tracking the interaction of users with AJAX applications for usability testing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, San Jose, California, USA. https://doi.org/10.1145/1240624.1240828.

Atterer, R., Wnuk, M., & Schmidt, A. (2006). Knowing the user's every move: user activity tracking for website usability evaluation and implicit interaction. In: *Proceedings of the 15th International Conference on World Wide Web*, Edinburgh, Scotland. https://doi.org/10.1145/1135777.1135811.

Baravalle, A., & Lanfranchi, V. (2003). Remote web usability testing. *Behavior Research Methods, Instruments, & Computers*, *35*(3), 364–368. doi:10.3758/BF03195512

Barnum, C. M. (2020). *Usability testing essentials: ready, set... test!*. (2nd ed.). Amsterdam: Morgan Kaufmann.

Bastien, J. M. C. (2010). Usability testing: A review of some methodological and technical aspects of the method. *International Journal of Medical Informatics*, *79*(4), e18–e23. doi:10.1016/j.ijmedinf.2008.12.004

Bergstrom, J. R., & Schall, A. J. (2014). *Eye tracking in user experience design* (J. R. Bergstrom & A. J. Schall, eds.). Boston, MA: Elsevier.

Brooke, J. (1996). SUS: A quick and dirty usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Eds.), *Usability evaluation in industry* (pp. 189–194). London: Taylor & Francis.

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

Chalil Madathil, K., & Greenstein, J. S. (2017). An investigation of the efficacy of collaborative virtual reality systems for moderated remote usability testing. *Applied Ergonomics*, *65*, 501–514. doi:10.1016/j.apergo.2017.02.011

Chin, J. P., Diehl, V. A., & Norman, K. L. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Washington, D.C., USA. https://doi.org/10.1145/57167.57203.

Chynał, P., & Szymański, J. M. (2011). Remote usability testing using eyetracking. In P. Campos, N. Graham, J. Jorge, N. Nunes, P. Palanque, & M. Winckler (Eds.), *Human-computer interaction – INTERACT 2011* (Vol. 6946, pp. 356–361). Berlin: Springer.

Cockton, G., Woolrych, A., Hornbæk, K., & Frøkjær, E. (2012). Inspection-Based Evaluations. In J. A. Jacko (Ed.), *The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications* (3rd ed.). Boca Raton, FL: CRC Press.

De Bleecker, I., & Okoroji, R. (2018). *Remote Usability Testing: Actionable insights in user behavior across geographies and time zones.* Birmingham, UK: Packt Publishing..

Doll, W. J., & Torkzadeh, G. (1988). The measurement of end-user computing satisfaction. *MIS Quarterly*, *12*(2), 259–274. doi:10.2307/248851

Dumas, J. S., & Fox, J. E. (2012). Usability testing. In J. A. Jacko (Ed.), *The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications* (3rd ed.). Boca Raton, FL: CRC Press.

Dumas, J. S., & Redish, J. (1999). *A practical guide to usability testing* (Rev. ed.). Portland, OR: Intellect Books.

Dumas, J. S., & Loring, B. A. (2008). *Moderating usability tests. principles & practice for interacting.* Morgan Kaufmann.

Edmonds, A. (2003). Uzilla: A new tool for web usability testing. *Behavior Research Methods, Instruments, & Computers*, *35*(2), 194–201. doi:10.3758/BF03202542

Ekman, P., & Friesen, W. V. (1975). *Unmasking the face: A guide to recognizing emotions from facial clues*. Englewood Cliffs, NJ: Prentice-Hall.

Faulkner, L. (2003). Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments, & Computers*, *35*(3), 379–383. doi:10.3758/BF03195514

Finstad, K. (2010). The usability metric for user experience. *Interacting with Computers*, *22*(5), 323–327. doi:10.1016/j.intcom.2010.04.004

Hammontree, M., Weiler, P., & Nayak, N. (1994). Remote usability testing. *Interactions*, *1*(3), 21–25. doi:10.1145/182966.182969

Hartson, H. R., Castillo, J. C., Kelso, J., & Neale, W. C. (1996). Remote evaluation: The network as an extension of the usability laboratory. In: *Paper presented at the CHI96: CHI '96 ACM conference on human factors.*

Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In G. Szwillus & J. Ziegler (Eds.), *Mensch & computer 2003* (vol. 57, pp. 187–196). Wiesbaden: Springer.

Hong, J. I., & Landay, J. A. (2001). WebQuilt: A framework for capturing and visualizing the web experience. In: *Paper presented at the WWW '01: Proceedings of the 10th international Conference on World Wide Web*.

International Organization for Standardization. (2010). ISO/IEC TR 25060:2010—Systems and software engineering—Systems and software product quality requirements and evaluation (SQuaRE)—Common industry format (CIF) for usability: General framework for usability-related information. Retrieved from https://www.iso.org/standard/35786.html

International Organization for Standardization. (2018). ISO 9241-11:2018: Ergonomics of human-system interaction - Part 11: Usability, definitions and concepts. Retrieved from https://www.iso.org/standard/63500.html

International Organization for Standardization. (2019). ISO 9241-210:2019: Ergonomics of human-system interaction - Part 210: Human-centred design for interactive systems. Retrieved from https://www.iso.org/standard/77520.html

Kieras, D. (2012). Model-based evaluation. In J. A. Jacko (Ed.), *The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications* (3rd ed.). Boca Raton, FL: CRC Press.

Kirakowski, J., & Corbett, M. (1993). SUMI: The software usability measurement inventory. *British Journal of Educational Technology*, *24*(3), 210–212. doi:10.1111/j.1467-8535.1993.tb00076.x

Laugwitz, B., Held, T., & Schrepp, M. (2008). Construction and evaluation of a user experience questionnaire. In A. Holzinger (Ed.), *HCI and usability for education and work* (vol. 5298, pp. 63–76). Berlin, Heidelberg: Springer Berlin Heidelberg.

Laugwitz, B., Schrepp, M., & Held, T. (2006). Konstruktion eines Fragebogens zur Messung der User Experience von Softwareprodukten. In H. M. Heinecke & H. Paul (Eds.), *Mensch und computer 2006*. München: OLDENBOURG WISSENSCHAFTSVERLAG.

Lewis, J. R. (1994). Sample sizes for usability studies: Additional considerations. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *36*(2), 368–378. doi:10.1177/001872089403600215

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, *7*(1), 57–78. doi:10.1080/10447319509526110

Lewis, J. R. (2002). Psychometric evaluation of the PSSUQ using data from five years of usability studies. *International Journal of Human-Computer Interaction*, *14*(3–4), 463–488. doi:10.1080/10447318.2002.9669130

Lewis, J. R. (2012). Usability testing. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (4th ed., pp. 1267–1305). Hoboken, NJ: Wiley.

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013). UMUX-LITE: when there's no time for the SUS. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Paris, France. https://doi.org/10.1145/2470654.2481287.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22*(140), 55–55.

Lin, H. X., Choong, Y.-Y., & Salvendy, G. (1997). A proposed index of usability: A method for comparing the relative usability of different software systems. *Behaviour & Information Technology*, *16*(4–5), 267–278. doi:10.1080/014492997119833

Lund, A. M. (2001). Measuring usability with the USE questionnaire. *Usability Interface*, *8*(2), 3–6. Retrieved from https://www.researchgate.net/profile/Arnold_Lund/publication/230786746_Measuring_Usability_with_the_USE_Questionnaire/links/56e5a90e08ae98445c21561c/Measuring-Usability-with-the-USE-Questionnaire.pdf

Minge, M., & Riedel, L. (2013). meCUE-Ein modularer fragebogen zur erfassung des nutzungserlebens. In: S. Boll, S. Maaß & R. Malaka (Hrsg.), *Mensch und Computer 2013: Inter-aktive Vielfalt* (S. 89–98). München: Oldenbourg Verlag.

Nielsen, J. (1994). Usability laboratories. *Behaviour & Information Technology*, *13*(1–2), 3–8. doi:10.1080/01449299408914577

Nielsen, J., & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. In: *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, Amsterdam, The Netherlands. https://doi.org/10.1145/169059.169166.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Champaign, IL: University of Illinois Press.

Rubin, J., & Chisnell, D. (2008). *Handbook of usability testing: How to plan, design and conduct effective tests*. New York, NY: John Wiley & Sons.

Sauro, J. (2015). SUPR-Q: A comprehensive measure of the quality of the website user experience. *Journal of Usability Studies 10*(2), 19.

Sauro, J., & Lewis, J. R. (2016). *Quantifying the user experience: Practical statistics for user research* (2nd ed.). Cambridge: Morgan Kaufmann.

Scholtz, J., Laskowski, S., & Downey, L. (1998, June 5). Developing usability tools and techniques for designing and testing Web sites. In : *Paper presented at the 4th Conference on Human Factors & the Web*, Basking Ridge, NJ.

Smith, P. A. (1996). Towards a practical measure of hypertext usability. *Interacting with Computers*, *8*(4), 365–381. doi:10.1016/S0953-5438(97)83779-4

Spool, J., & Schroeder, W. (2001, 31 March–5 April). Testing Web sites: Five users is nowhere near enough. In: *CHI '01 Extended Abstracts on Human Factors in Computing Systems*, Seattle, Washington, New York..

Thompson, K. E., Rozanski, E. P., & Haake, A. R. (2004). Here, there, anywhere: Remote usability testing that works. In: *Proceedings of the 5th Conference on Information Technology Education*, Salt Lake City, UT, USA. https://doi.org/10.1145/1029533.1029567.

Tullis, T., & Albert, B. (2013). *Measuring the user experience: Collecting, analyzing, and presenting usability metrics*. Amsterdam: Elsevier/Morgan Kaufmann.

Tullis, T., Fleischman, S., McNulty, M., Cianchette, C., & Bergel, M. (2002). An empirical comparison of lab and remote usability testing of Web sites. In: *Paper presented at the Usability Professionnal Association Conference*, https://www.researchgate.net/publication/228540469_An_empirical_comparison_of_lab_and_remote_usability_testing_of_Web_sites.

Virzi, R. A. (1990). Streamlining the design process: Running fewer subjects. *Proceedings of the Human Factors Society Annual Meeting*, *34*(4), 291–294. doi:10.1177/154193129003400411

Virzi, R. A. (1992). Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *34*(4), 457–468. doi:10.1177/001872089203400407

Wang, J., & Senecal, S. (2007). Measuring perceived website usability. *Journal of Internet Commerce*, *6*(4), 97–112. doi:10.1080/15332860802086318

Winckler, M. A. A., Freitas, C. M. D. S., & Valdeni de Lima, J. (2000). *Usability remote evaluation for WWW*. In: *CHI '00 Extended Abstracts on Human Factors in Computing Systems*, The Hague, The Netherlands. https://doi.org/10.1145/633292.633367.

Yang, T., Linder, J., & Bolchini, D. (2012). DEEP: Design-oriented evaluation of perceived usability. *International Journal of Human-Computer Interaction*, *28*(5), 308–346. doi:10.1080/10447318.2011.586320