International Journal of Human–Computer Interaction

ISSN: 1044-7318 (Print) 1532-7590 (Online) Journal homepage: https://www.tandfonline.com/loi/hihc20

Inclusive Web Empirical Studies in Remote and In-Situ Settings: A User Evaluation of the RemoTest Platform

Myriam Arrue, Xabier Valencia, J. Eduardo Pérez, Lourdes Moreno & Julio Abascal

To cite this article: Myriam Arrue, Xabier Valencia, J. Eduardo Pérez, Lourdes Moreno & Julio Abascal (2019) Inclusive Web Empirical Studies in Remote and In-Situ Settings: A User Evaluation of the RemoTest Platform, International Journal of Human–Computer Interaction, 35:7, 568-583, DOI: 10.1080/10447318.2018.1473941

To link to this article: https://doi.org/10.1080/10447318.2018.1473941

Published online: 23 May 2018.

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
2019, VOL. 35, NO. 7, 568–583
https://doi.org/10.1080/10447318.2018.1473941

Taylor & Francis
Taylor & Francis Group

Inclusive Web Empirical Studies in Remote and In-Situ Settings: A User Evaluation of the RemoTest Platform

Myriam Arrue[a], Xabier Valencia[a], J. Eduardo Pérez[a], Lourdes Moreno[b], and Julio Abascal[a]

[a]EGOKITUZ, University of the Basque Country (UPV/EHU), Donostia (Gipuzkoa), Spain; [b]Computer Science Department, Universidad Carlos III, Madrid, Spain

ABSTRACT

Web accessibility evaluation requires tests to be carried out with real users with disabilities performing real tasks or activities. To recruit an appropriate group of users and to observe their performance in the real world is difficult. For this reason we have developed RemoTest, a platform that assists researchers designing experiments, conducting remote and in-situ experimental sessions and analyzing the data gathered while the users are accessing the Web. Although this tool is oriented to experimenters, it is necessary to check whether the evaluation environments created by RemoTest are accessible or not to the users that participate in the tests. To this end, we conducted formal in-situ evaluations with 36 users with diverse characteristics. For this assessment, the participants were asked to install the platform, to fill in some automatically created questionnaires and to carry out several web navigation tasks. From the data gathered we analyzed the ease of the installation process, the accessibility of the automatically generated questionnaires, and user satisfaction. The results revealed the suitability of the platform for conducting inclusive experiments both in remote and in-situ contexts and provided guidelines on how the experiments should be set out.

1. Introduction

The involvement of users in the evaluation of web services is fundamental in order to achieve universal access to the information society. Other methods exist to assess the usability and accessibility of web sites, such as the use of automatic checkers of sets of accessibility guidelines or standards or the hiring of experts to perform manual evaluations (Petrie & Bevan, 2009). However, evaluation by real users is the most valuable technique because it enables the detection of real problems and barriers that users experience while using the web pages. The expertise of each user, the configuration of the system, the assistive technology utilized by the user, are just a few of the variables that can determine whether the user manages to overcome a potential barrier or not.

User evaluation requires conducting experimental sessions with large and diverse groups of users. Researchers need to clearly define the tasks to be performed as well as the specific questionnaires which are required in order to explicitly obtain certain data from the participants such as satisfaction level, socio-demographic data and emotional aspect. Therefore, the experiment design process is demanding and requires experience from different areas: human factors, hypertext, web technology, etc.

In addition, involving an appropriate number of participants for a specific experiment is also challenging. Frequently, this is due to the location and the rigorous timing of sessions. Nowadays, interest in the use of software tools to conduct experiments remotely is increasing because they allow participants to be observed while they perform the tasks in their habitual daily environment. To work with their own resources and devices (which are already adapted to their needs) is particularly important when working with people with disabilities, as it facilitates the conduction of experiments "on the wild." Moreover, this type of experiment gathers real interaction data without any obtrusive observation mechanism (Apaolaza et al., 2013). It also makes it possible to involve a larger number of participants, as they do not have to physically get to a specific location.

However, remote usability testing also has drawbacks. For example, it may not provide a thorough understanding of the users and their behavior and it is also necessary for all participants to have access to a reliable Internet connection (Albert, Tullis, & Tedesco, 2009). In order to carry out remote experimental sessions the remote tool should meet accessibility requirements to ensure that the tool can be used by a wider range of users. Finally, the set-up, installation and configuration processes must be accessible and user-friendly.

This article presents an evaluation of the accessibility and suitability of the platform RemoTest (Valencia, Pérez, Muñoz, Arrue, & Abascal, 2015) to carry out sessions with people with disabilities. The RemoTest platform objective is to assist researchers to design experiments, conduct

CONTACT Lourdes Moreno lmoreno@inf.uc3m.es Computer Science Department, Universidad Carlos III de Madrid, Avda. de la Universidad Nº 30, 28911 Leganés, Spain.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hihc.

experimental sessions and analyze data gathered in the evaluation sessions.

In order to verify whether the evaluation environments created by RemoTest are accessible or not to the users that participate in the test, the accessibility and usability of the test environment created for participants was evaluated by 36 users with different characteristics: 13 people with physical disabilities, 10 blind people, 8 people with low vision and 5 able-bodied people.

An in-situ experimental session was conducted and participants were asked to install the testing tool which is a Firefox add-on and perform different types of tasks such as filling in questionnaires automatically generated by the tool and web navigation tasks. Results revealed the suitability of the platform for conducting inclusive experiments both in remote and in-situ contexts.

2. Systems for web testing

Several remote web usability-testing tools have been developed in the last decade. They can be classified as server-side, proxy-based or client-side tools depending upon their architecture. Server-side tools are the most transparent for users, since no installation or configuration is needed (Etgen & Cantor, 1999; Google analytics, 2018; Leiva & Vivó, 2013; Optimizely, 2018; Paganelli & Paternò, 2002; Santana & Baranauskas, 2010; Scholtz, Laskowski, & Downey, 1998). Even though only HTTP requests can be gathered by the tool developed by Scholtz et al. (1998), adding some additional code, usually some JavaScript, to the web pages enables significant user interaction data to be gathered (Claypool, Le, Wased, & Brown, 2001; Etgen & Cantor, 1999; Leiva & Vivó, 2013, 2013; Paganelli & Paternò, 2002). This approach can be considered only when the web pages being evaluated are located in servers to which the researchers have access. On the other hand, proxy-based tools allow the evaluation of un-owned web pages but they require some configuration parameters to be fixed by the users (users have to configure their browser to access via the proxy) (Atterer, Wnuk, & Schmidt, 2006; Hong, Heer, Waterson, & Landay, 2001). Client-side tools (Claypool et al., 2001; Edmonds, 2003; Gajos, Reinecke, & Herrmann, 2012) are the most appropriate for usability testing since researchers can have access to any local interaction data generated in the experimental sessions (browser back, forward, bookmark, print options, mouse contextual menus, vertical/horizontal scrolling actions, etc.). Moreover, this type of architecture facilitates the inclusion of a questionnaire during the experimental session so that explicit data can be gathered from the participants. Other systems, such as USERZOOM (2018) can act as a server-side tool or client-side tool depending on the data to be gathered or the type of web site.

Table 1 presents information regarding the architecture of the most used remote usability-testing tools as well as the implicit interaction data gathered during the experimental sessions (mouse events, keyboard events, window events, browser actions and information in HTTP requests). In addition, other events collected by tools are also specified in Table 1. For instance, the NIST WebMetrics Suite (Scholtz et al., 1998) allows the injection of code to links in order to track the path followed by the user. Almost all the tools gather mouse, keyboard and window events, but there are more differences between them when it comes to browser actions and information in HTTP requests. Only UZILLA (Edmonds, 2003) and MORAE (2018) collect browser actions such as back/forward buttons. WEBQUILT (Hong et al., 2001) only obtains information from the HTTP requests. The last five tools indicated in Table 1 are commercial whereas the others were developed in an academic environment.

The system developed by Gajos et al. (2012) is devoted to gathering interaction data for data mining purposes. The USAPROXY tool (Atterer et al., 2006) injects tracking code automatically via proxy. Neither tool includes features for analyzing or visualizing the gathered interaction data. The rest of the tools have some functionality in order to facilitate the visualization and analysis of collected interaction data in remote experimental sessions.

Regarding the different types of experiments that can be executed by the analyzed tools, two main kinds of tasks are found: target searching and free navigation tasks. Target searching tasks require some features to be included in the testing tool: defining the target of the tasks, determining their duration, giving instructions to the participants and informing them when the target has been reached or when they are out of time. UZILLA, USERZOOM, Loop11 (LOOP11, 2018) and Morae are tools that include all these features. The other tools are devoted to conducting experiments based on free navigation tasks.

Table 1. Web usability-testing tools classification.

		Implicit interaction data					
Tool name	Architecture	Mouse	Keyboard	Window	Browser	HTTP requests	Other
NIST WebMetrics Suite (Scholtz et al., 1998)	Server-side	No	No	No	No	No	Path
WET (Etgen & Cantor, 1999)	Server-side	Yes	Yes	Yes	No	No	No
SMT2 (Leiva & Vivó, 2013)	Server-side	Yes	Yes	Yes	No	No	No
WELFIT	Server-side	Yes	Yes	Yes	No	No	Customized events
WebRemUSINE (Paganelli & Paternò, 2002)	Server-side	Yes	Yes	Yes	No	No	No
Curious Browser (Claypool et al., 2001)	Client-side	Yes	Yes	No	No	No	No
USAPROXY (Atterer et al., 2006)	Proxy-based	Yes	Yes	Yes	No	No	No
WEBQUILT (Hong et al., 2001)	Proxy-based	No	No	No	No	Yes	No
UZILLA (Edmonds, 2003)	Client-side	Yes	Yes	Yes	Yes	No	No
Gajos et al. (2012)	Client-side	Yes	Yes	Yes	No	No	No
Optimizely (2018)	Server-side	Yes	Yes	Yes	No	No	Customized events
MORAE (2018)	Client-side	Yes	Yes	Yes	Yes	No	Customized events
Google analytics (2018)	Server-side	Yes	Yes	Yes	No	No	Customized events
LOOP11 (2018)	Proxy-based	Yes	Yes	Yes	No	No	No
USERZOOM (2018)	Client/Proxy	Yes	Yes	Yes	No	No	No

When it comes to collecting explicit data from participants by means of questionnaires (for measuring satisfaction, emotions, etc.), Morae, USERZOOM, UZILLA and LOOP11 include features for presenting and getting information through questionnaires before and after completing the tasks. Curious Browser (Claypool et al., 2001) presents questionnaires after every visited new page in order to study the relation between the events gathered during the session with users' interest in the page being evaluated.

Many of those systems, such as Loop11, Morae, Usaproxy, WELFIT (Santana & Baranauskas, 2010) or the one presented by Gajos et al. have been or are being used by people with disabilities. But only Morae and Loop11 can be used to perform guided user testing. The other tools are more focused on free navigation tasks. No accessibility evaluations could be found about the use or installation of Morae or Loop11. Morae is a powerful tool with which to perform user behavior studies but it is quite difficult to use or to be installed by people with disabilities. On the contrary, Loop11 can be easily used due to its proxy-based architecture. It does not require any installation as user testing starts when accessing a predetermined URL. One drawback of this tool is that it does not gather any information about the browser events occurring during the experimental sessions since the system acts as a proxy within the user and the evaluated web pages.

The RemoTest platform (Valencia et al., 2015) is a web testing tool which gathers most of the implicit interaction data presented in Table 1: mouse, keyboard, window and browser events and HTTP Requests. In addition, it includes features for defining different types of experiments and tasks as well as questionnaires for gathering explicit data from participants. It was developed as an inclusive testing platform which takes accessibility into account throughout the process. The following section describes the general architecture of the platform.

3. Remotest, platform for inclusive web experiments

The RemoTest platform provides evaluators with functionalities to facilitate the definition of experiments, manage experimental remote/in-situ sessions, describe questionnaires/surveys to be displayed to participants and to gather interaction data produced during the sessions and analyze this interaction data. This platform admits a wide range of experiments with a variety of objectives, for instance, to study user behavior when performing a task on different websites, to analyze and compare the navigational strategies of different types of participants when interacting with the same website, to evaluate the accessibility-in-use of several websites, to gather significant information through surveys, to measure user satisfaction when using certain web services, to analyze user performance improvement when interacting with adapted versions of original web pages and so on.

The architecture of the platform has been designed taking all these different types of experiments into consideration. In this case, we opted for a hybrid architecture model that includes some functionalities from a client-side module and other ones from some server-side modules. The platform is split into four modules: experimenter module (EXm), participant module (PAm), coordinator module (COm) and results viewer module (RVm). Figure 1 shows the architecture and interactions between these modules.

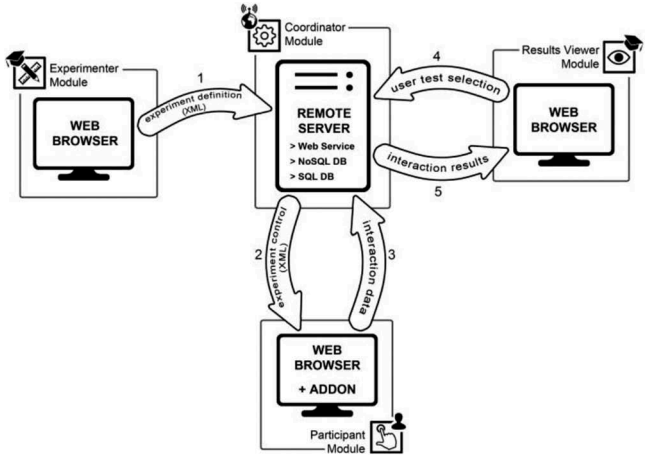

Figure 1. The RemoTest platform general architecture and interactions between modules.

Each module has specific functions and uses different technologies. The EXm is responsible for assisting researchers during the experiment definition process. The experiment definition is stored in an XML file based on specific vocabulary created for specifying experiments. This vocabulary is comprehensive enough to define the tasks, objectives, stimuli to be presented, task time limits, questionnaires to be filled in by participants and so on. The COm exploits the information in this XML file (Step 1 in Figure 1) in order to create personalized experimental sessions for each participant. These personalized sessions are transferred to the corresponding PAm (Step 2 in Figure 1). The PAm guides participants during the experimental sessions, presents the stimuli to participants and gathers the interaction data created during the experimental sessions. The interface of the PAm has been designed with accessibility aspects taken into account so the initial login screen, task description screens and the presented questionnaires conform to WCAG 2.0 accessibility guidelines (W3C, 2008). This module is developed as an add-on for Firefox and has to be locally installed in the participants' computer. The interaction data are centrally stored in a remote server for future analysis (Step 3 in Figure 1). The RVm organizes and presents the abundant interaction data gathered in the experiments (Step 4 and 5 in Figure 1).

4. Evaluation methods for assessing web-based tools

The evaluation of web-based tools entails significant challenges, as different aspects have to be considered. This work focuses on the evaluation of the user-testing tool installation process (PAm of the RemoTest platform), accessibility and usability of the interfaces automatically created and displayed by the tool and users' overall satisfaction and acceptance. This system has been developed to be used by people with different skills and ways of access, including people with and without

disabilities as well as users employing different system configurations.

Several methods have been considered for carrying out the evaluation of the RemoTest platform: rating pragmatic quality (PQ) attributes of the installation process and emotional aspects during the installation, user testing and expert-based evaluations for detecting accessibility barriers, observational methods and inquiry methods for assessing the overall satisfaction and acceptance of the tool.

4.1. User experience (UX) evaluation

UX can be defined as

> the entire set of affects that is elicited by the interaction between a user and a product, including the degree to which all his or her senses are gratified (aesthetic experience), the meanings we attach to the product (experience of meanings) and the feelings and emotions that are induced (emotional experience). Law, Roto, Hassenzahl, Vermeeren, & Kort, 2009)

UX covers different aspects such as aesthetics, emotions, usability/pragmatic attributes, hedonic attributes, cognitive load, interactivity, social responses, persuasion and acceptability (Brajnik & Giachin, 2014). Each of these aspects is evaluated with different metrics or established methods. Pragmatic attributes and emotions were especially considered for evaluating the installation process of the RemoTest platform.

Pragmatic attributes are connected to the users' need to achieve behavioral goals (Hassenzahl, 2004). "A product may be perceived as pragmatic because it provides effective and efficient means to manipulate the environment" (Hassenzahl, 2005). Thus, PQ can be understood as perceived usability. Hassenzahl's Attrakdiff questionnaire[1] was selected as the method to gather the perceived PQ of the installation process. The set of seven word pairs reflecting opposite adjectives that can be rated on a 7-point scale to measure users' perceptions of the PQ attributes was translated to Spanish and introduced to the RemoTest platform so that a questionnaire was automatically generated and displayed to the participants after installing the tool.

Gathering information about emotions of participants during the installation process enables the appraisal of situations from an affective point of view (i.e., assigning arousal and valence value). In this work, emotional feedback was collected from participants in terms of the self-assessment manikin (SAM) scale to measure dimensions of valence (pleasantness of the emotion) and arousal (strength of the emotion) (Bradley & Lang, 1994).

4.2. Accessibility evaluation

Effective evaluation of websites for accessibility remains problematic. Automated evaluation tools still require significant manual testing and human judgment. Furthermore, the evaluation methodologies such as the one proposed by the web accessibility initiative presupposes that the evaluator has considerable knowledge about accessibility and assistive technology. There are other methods such as the Barrier Walkthrough Method[2] that can be performed more easily and is reliable and efficient in terms of the time required for carrying out the evaluation. This evaluation method in combination with conformance to WCAG 2.0 guidelines was applied during the RemoTest platform development and experiment preparation.

However, accessibility barriers that can only be discovered by user testing may be overlooked. In this work, user testing has been conducted in order to analyze the interaction of participants with the interfaces which are automatically created and displayed by the RemoTest platform. Participants were observed during the experiment and their interaction was video recorded so any accessibility barrier was detected during the analysis.

4.3. Interaction data gathering and inquiry methods

Participants' interaction data, such as the time required to complete the given navigational tasks or the time required for filling in the questionnaires, were automatically collected by the RemoTest platform and were analyzed in order to detect any barrier. In addition, participants were interviewed to record their accessibility/usability perceptions. Some questions were directly related to rating the ease of filling in the automatically generated questionnaires and to comment upon any difficulty presented by this process.

Inquiry methods were also used to measure participants' overall satisfaction and their acceptance of the RemoTest platform.

5. Experimental study

5.1. Research goals

The aim of the experimental study was to evaluate the RemoTest tool from the participants' perspective. In this case, an in-situ setting was chosen for the experimental sessions in order to obtain first-hand direct feedback from users and to be able to help them in any problem occurring during the interaction. The main objectives were evaluating the suitability and accessibility of the tool for carrying out experimental sessions with different groups of users. The study included the evaluation of the installation process as well as performing different tasks managed by the tool and obtaining participants' information through automatically generated questionnaires. Any accessibility barrier encountered during the process was immediately communicated to and annotated by experimenters. Participants' perceptions and opinions about the accessibility, usability and usefulness of the tool were gathered through semi-structured interviews.

The experimental study was designed to explore the following research questions:

- Q1: The RemoTest installation process is accessible and usable regardless of the participants' characteristics and the assistive technology used.

- Q2: The questionnaires automatically generated by RemoTest are accessible regardless of the participants' characteristics and the assistive technology used.
- Q3: Participants are satisfied with the tool performance, consider that RemoTest is easy to use and would use it in future experimental sessions even in remote settings.

5.2. Participants

The evaluation required participants of different groups of users. A call for participation was disseminated through several organizations of people with disabilities, social networks and email distribution lists. A total of 36 users were recruited. As required by the study all of them had some experience in using computers and Internet browsing.

Table 2 shows the description of the 36 participants who took part in the study. These participants have been grouped together into four user groups: physical disability (13 participants, 36.1%), blind (10 participants, 27.8%), low vision (8 participants, 22.2%) and participants without disabilities (5 participants, 13.9%). Figure 2 shows the frequency distribution bar chart of each user group.

The assistive technology used by participants is also included in Table 2. With regard to the participants with physical disabilities, four users employed joysticks, four adapted mouse, two head pointers, two users did not use any specific assistive technology but some specific configuration (such as switching right and left button functions in the mouse) and one interacted using the touchpad. Regarding the blind users, nine users employed a screen reader (one of them used it jointly with a braille display), all of them used JAWS screen reader. In the low vision user group, four users employed the ZoomText screen magnifier software, three users applied browser zoom functionalities and one user configured system settings to obtain high contrast interfaces.

Of the 36 participants in the study 17 were female and 19 were male. Mean age was 44.06 (SD = 9.9), see Figure 3 for the frequency distribution bar chart.

The Internet usage experience (1–3 years, 4–6 years, more than 7 years), the Internet expertise level (beginner, intermediate or advanced) and the Internet use frequency (daily, weekly, monthly) varied among participants. Generally, most of them claimed to have a usage experience of more than 7 years (80.56% of participants), have an intermediate expertise level (58.33% of participants) and to use the Internet daily (80.33% of participants). Figure 4 shows the frequency distribution bar chart of Internet usage experience, Internet expertise level and Internet use frequency. No significant variations are noticed in those values among user groups.

Participants were also asked about their experience with Mozilla Firefox browser as the RemoTest has been developed as an extension of this browser. The data obtained indicated that it is not the favorite browser among participants, eight participants (P2, P7, P9, P14, P23, P25, P26 and P30) said they have never used it. Only seven participants (19.4% of participants) stated that it is

Table 2. Description of the participants in the study.

Id	Sex	Age	Disability | nondisabled	Assistive technology (AT)	Expertise level	Setting
P1	Male	45	Physical disability	Joystick	Intermediate	Elkartu
P2	Female	39	Physical disability	Joystick	Intermediate	Elkartu
P3	Male	53	Physical disability	Adapted mouse	Beginner	Elkartu
P4	Male	40	Physical disability	Joystick	Advanced	Home
P5	Female	59	Physical disability	Nothing	Beginner	Elkartu
P6	Male	50	Physical disability	Nothing	Beginner	Elkartu
P7	Female	54	Physical disability	Touchpad	Beginner	Elkartu
P8	Female	42	Physical disability	Head pointer	Intermediate	Elkartu
P9	Male	41	Physical disability	Head pointer	Advanced	Elkartu
P10	Female	76	Physical disability	Joystick	Intermediate	Elkartu
P11	Male	50	Blind	Screen reader	Intermediate	LabUC3M
P12	Male	44	Low vision	Screen magnifier	Intermediate	LabUC3M
P13	Female	54	Low vision	Browser zoom	Intermediate	LabUC3M
P14	Female	52	Low vision	Screen magnifier	Intermediate	LabUC3M
P15	Female	44	Low vision	High Contrast	Intermediate	Servimedia
P16	Female	39	Low vision	Screen magnifier	Intermediate	Servimedia
P17	Male	47	Blind	Screen reader	Intermediate	LabUC3M
P18	Male	32	Blind	Screen reader	Intermediate	LabUC3M
P19	Female	54	Low vision	Screen magnifier	Intermediate	LabUC3M
P20	Female	34	Blind	Screen reader	Intermediate	LabUC3M
P21	Male	45	Low vision	Screen magnifier	Intermediate	Servimedia
P22	Male	45	Blind	Braille display screen reader	Advanced	Servimedia
P23	Female	41	Low vision	Browser zoom	Intermediate	Servimedia
P24	Male	36	Blind	Screen reader	Intermediate	Servimedia
P25	Male	23	Blind	Screen reader	Intermediate	Servimedia
P26	Female	30	Blind	Screen reader	Beginner	Servimedia
P27	Female	31	Blind	Screen reader	Advanced	LabEHU
P28	Male	40	Blind	Screen reader	Advanced	LabEHU
P29	Male	43	Physical disability	Adapted mouse	Advanced	LabEHU
P30	Female	57	Physical disability	Adapted mouse	Advanced	Home
P31	Female	52	Physical disability	Adapted mouse	Intermediate	Home
P32	Female	32	Nondisabled	Nothing	Advanced	LabEHU
P33	Male	41	Nondisabled	Nothing	Intermediate	LabEHU
P34	Male	34	Nondisabled	Nothing	Advanced	LabEHU
P35	Male	45	Nondisabled	Nothing	Advanced	LabEHU
P36	Male	42	Nondisabled	Nothing	Intermediate	LabEHU

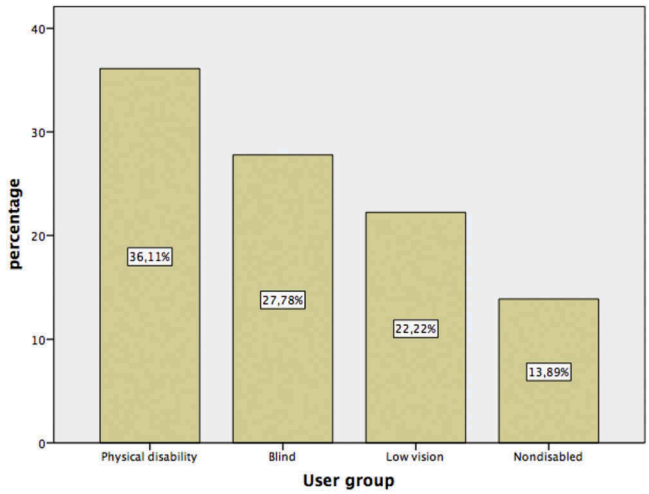

Figure 2. Frequency distribution bar chart of user group.

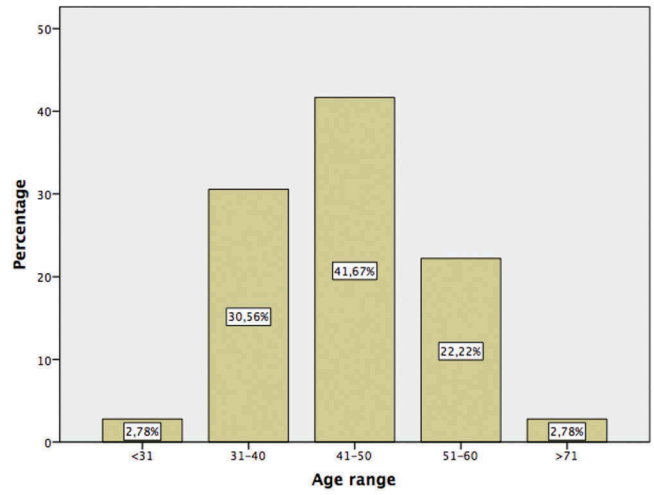

Figure 3. Frequency distribution bar chart of age.

their habitual browser and that they always use it for connecting to the Internet, 38.9% of participants used it sporadically and 19.4% of participants usually used Firefox. These frequencies do not have many variations segmented by user group.

5.3. Setting and equipment

The experimental sessions were carried out in different settings. There were four main different settings: the computer room at Elkartu (an association of people with physical disabilities), a laboratory located in the Carlos III University of Madrid (LabUC3M), a laboratory of the Computer Science School at the University of the Basque Country (LabEHU) and Servimedia (a news agency that employs people with visual disabilities). Nine experimental sessions were conducted on the Elkartu premises and eight sessions were conducted in each location: LabUC3M, LabEHU and Servimedia. In addition, the experimental sessions of three participants (P4, P30, P31) were conducted at the participant's home.

Participants were encouraged to use their own laptop and assistive technology whenever possible. The objective was to evaluate the RemoTest on different platforms and settings adapted to the participants. However, a desktop PC or laptop was configured for the sessions in Elkartu, LabUC3M and LabEHU. The desktop PC in Elkartu and the laptop in LabUC3M run Microsoft Windows 7 and Mozilla Firefox 25.0. The desktop PC in Elkartu was utilized by seven participants with their own assistive technology (joystick, head pointer, etc.), the other two participants (P3, P7) used their own laptop with the same configuration (Microsoft Windows 7 and Mozilla Firefox 25.0). The laptop in LabUC3M was utilized in 5 sessions in which a ZoomText magnifier and a JAWS 15 screen reader were also installed. The other three participants (P12, P13 and P17) used their own laptop with different configurations: Microsoft Windows XP and Mozilla Firefox 25 (P12), Windows Vista and Mozilla Firefox 25 (P13) and Microsoft Windows 7 and Mozilla Firefox 9.0.1 (P17). The desktop PC in LabEHU runs Microsoft Windows XP and Mozilla Firefox 22.0. All participants except for P27 and P28 used it. The laptops of these participants did not differ on the Mozilla Firefox version from the one installed in the PC but the laptop of P28 ran Microsoft Windows 7. Participant P29 used his own trackball to interact with the PC. Different configurations were found in Servimedia as the experimental sessions were conducted in the participant workplace. All computers run Microsoft Windows XP but differ in the version of Mozilla Firefox (we found 8.0.1, 19.0, 21.0, 23.0, 25.0 and 26.0 versions). Two of the participants conducting the session at home (P30, P31) had the same configuration: Microsoft Windows 7 and Mozilla Firefox 22.0. Finally, participant P4 used his own desktop PC running Microsoft Windows 7 and Mozilla Firefox 25.0.

5.4. Tasks and stimuli

Users were asked to perform different types of tasks. The first task (Task 1) was to install the RemoTest tool based on the instructions provided on a web page. Then, participants were asked to log in and fill in a set of questionnaires and complete some web navigation tasks. The tasks proposed to participants by the RemoTest were the following: filling in a questionnaire to provide some feedback about their perception about the tool installation process (Task 2), a free navigation task on a website (Task 3), a target searching task on a website (Task 4) and filling in a questionnaire to provide their socio-demographic data (Task 5).

Some of the stimuli presented by the RemoTest in the experimental sessions were manually designed web pages. For instance, a web page was developed for giving instructions for the installation process of the RemoTest and was displayed in Task 1. Other stimuli were automatically generated by the RemoTest such as the questionnaires and task description pages displayed in the rest of the tasks. Table 3 shows the description of the stimuli presented in the experimental sessions. It describes each stimuli by indicating the task to which it is related, the type of stimuli (informational web page, task description web page, questionnaire to be fulfilled, task completion indication), the description

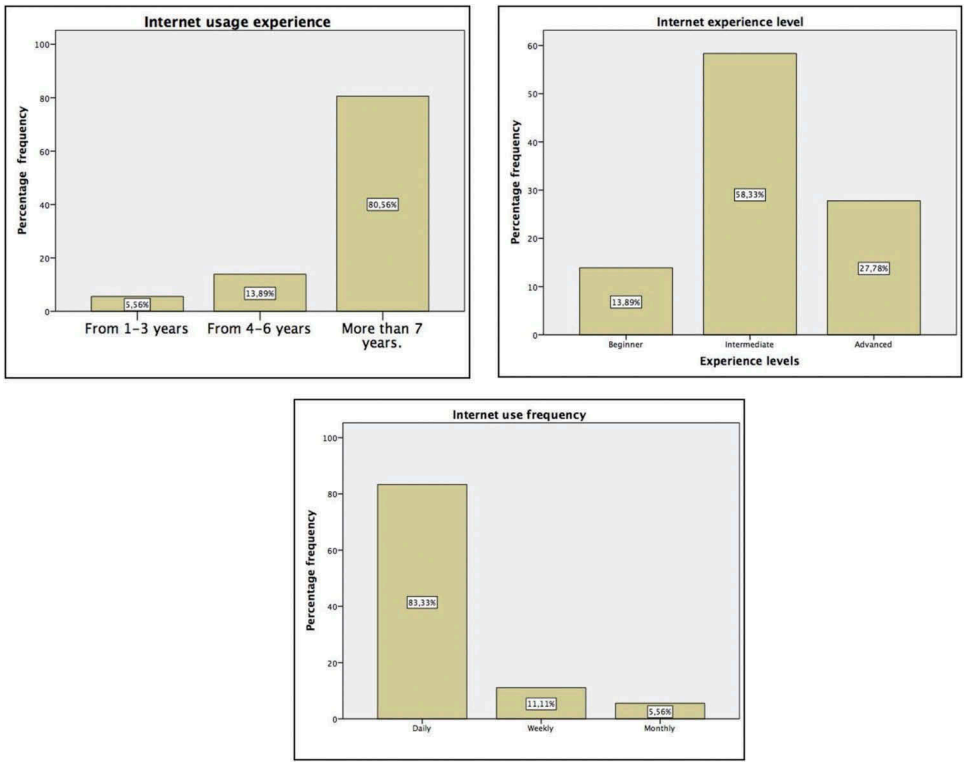

Figure 4. Frequency distribution bar chart of internet usage experience, internet expertise level and internet use frequency.

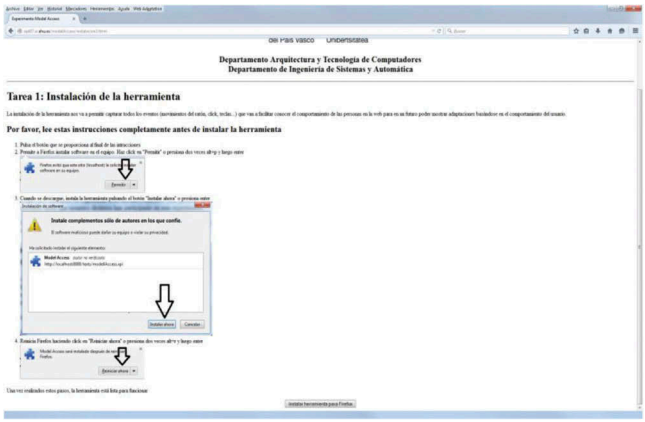

Figure 5. Manually developed web page containing the description of the RemoTest installation process (S1).

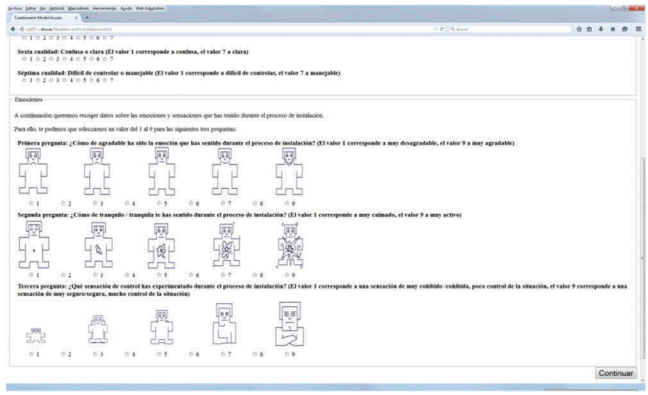

Figure 6. Automatically created questionnaire for gathering participants' perceived pragmatic aspects of the installation process and their emotional state during Task 1 (S2).

of the stimuli and whether it had been manually developed for the experiment or automatically generated by the RemoTest tool.

The Figures 5, 6, 7 and 8 show some screenshots of the stimuli displayed during the experimental session (S1, S2, S3 and S7). This set of screenshots includes all the different types of web pages displayed by RemoTest. The remaining stimuli (S4, S5, S6 and S8) are the ones created automatically by the tool for describing a task or for indicating the completion of the task and they all are similar to S3.

As stated above, S1 was manually generated with all the information needed by participants about the installation process. The different popup windows that would be presented by Mozilla Firefox were displayed and explained in order to install the add-on. Once they started the process (clicking on the button at the bottom of the web page) all the installation process was guided by Mozilla Firefox as in the installation of any other add-on.

The stimuli automatically generated by the RemoTest (S2–S8) were created first in XUL but an expert evaluation carried out by two blind people showed some critical accessibility issues (Valencia, Arrue, Rojas-Valduciel, & Moreno, 2014). Therefore, the stimuli generation process was updated to create the stimuli in HTML to avoid accessibility barriers. A Barrier Walkthrough Method was carried out by two of the authors in order to detect any accessibility barrier in these stimuli in HTML for blind people, people with physical

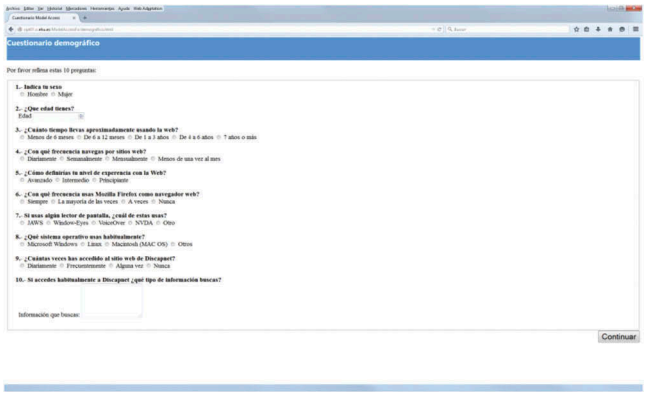

Figure 7. Automatically created questionnaire for gathering participants' socio-demographic data (S7).

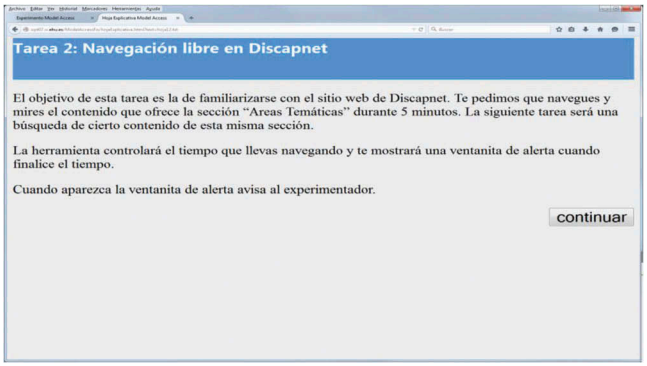

Figure 8. Automatically created web page with the description of the Task 3 (S3).

disabilities or low vision users. There were no significant accessibility barriers detected though there were some minor issues which will be improved upon in future versions of the tool such as including shortcuts for activating the button in the web pages (the button with the text "Continuar") and skipping links to directly access specific questions in the questionnaires (S2 and S7).

The free navigation and target searching tasks (Tasks 3 and 4, respectively) were carried out on the Discapnet website [www.discapnet.com]. This website focuses on providing information to people with disabilities. It officially conforms to the AA level defined in WCAG 1.0 accessibility guidelines.

5.5. Procedure

The sessions with participants were conducted one at a time. The whole test was conducted in the participants' mother tongue, Spanish. Each session started by providing information about the objectives of the study. Participants were told that their contribution to the scientific experiment was voluntarily, and that they could withdraw from the study at any point. All participants followed the same sequence of tasks in the experimental session. Then, all of them started installing the RemoTest tool and carried out the questionnaire completion tasks and navigation tasks. Finally, they were briefly interviewed. All the interactions with RemoTest platform were video recorded and the interviews were audio recorded.

5.6. Data Collection

The following methods were used for data collection:

- Interaction data: Every user interaction with the Discapnet website and the web pages automatically generated by RemoTest was monitored and stored in XML files. These files contain information such as the time at which each task was started, web pages visited, cursor movements and browser events.
- Video recordings: User interactions were recorded with a video camera. These recordings provided us with information about the users' interaction with the interfaces displayed by RemoTest.
- Observations: Interaction-specific aspects that drew the attention of the experimenters were noted (for instance, problems that occurred during the interaction or installation of the tool).
- Semi-structured interview: Two short post-interaction interviews were carried out and were audio recorded. Both interviews focused on getting information about users' satisfaction levels and opinions on the RemoTest tool, displayed interfaces, difficulties encountered when accomplishing tasks, etc. The objective was to gain direct feedback from participants.

Table 4 presents the data collected during each task in the experimental sessions.

This section is devoted to the analysis of the data gathered in the experimental session regarding the problems detected in the RemoTest tool installation process and in the automatically generated questionnaire completion tasks. Thus, data collected in Task 1, Task 2, Task 5 and the interviews are analyzed in this section. The analysis of the user interaction data automatically gathered by RemoTest (data collected in Task 3 and Task 4) is beyond the scope of this article and was carried out in other previously published research papers (Pérez, Arrue, Valencia, & Moreno, 2014; Valencia et al., 2015).

5.6.1. Remotest installation process

The installation process was evaluated based on the data collected in Task 1, Task 2 and the first short interview. Task 1 was completed by all of the participants even though some of them, by means of the responses given in the questionnaire of Task 2 and the comments in the short interview, reported several issues which could be improved upon and minor accessibility barriers they were faced with.

Analysis of the data gathered through the questionnaire in Task 2:

Task 2 consisted in filling in a questionnaire about the installation which was used to measure the perceived usability of the installation process and the emotions felt by participants during the installation. This questionnaire consisted of two parts: the first one was devoted to gathering users'

Table 3. Information about the stimuli presented by RemoTest during the experimental sessions.

Id	Task	Type	Description	Manually/ Automatically
S1	Task 1	Task description	Web page containing the description of the RemoTest installation process.	Manually
S2	Task 2	Questionnaire	Participants were asked to fulfill a questionnaire about perceived pragmatic aspects of the installation process and their emotional state during the installation task (Task 1).	Automatically
S3	Task 3	Task description	Web page containing the description of Task 3 (free navigation on a website).	Automatically
S4	Task 3	Task completion	Web page indicating the completion of Task 3.	Automatically
S5	Task 4	Task description	Web page containing the description of Task 4 (searching a target link on a website).	Automatically
S6	Task 4	Task completion	Web page indicating the completion of Task 4.	Automatically
S7	Task 5	Questionnaire	Participants were asked to fulfill a questionnaire about their socio-demographic data.	Automatically
S8	Task 5	Task completion	Web page indicating that the provided data has been correctly stored and informing about the end of the experimental session.	Automatically

Table 4. Information about the data collected in each task of the experimental session.

Task	Description	Data collected
Task 1	Installing the RemoTest tool based on the instructions provided on a web page (stimuli S1)	• -Annotations of problems occurred during the installation process based on direct observation and video recordings.
Task 2	Filling in a questionnaire automatically generated by the RemoTest tool (stimuli S2) for gathering participants' perception about the installation task (Task 1).	• -Time for completing the task. • -Perceived UX usability and pragmatic attributes based on the Hassenzahl's model. • -Participants' emotions during installation task based on the SAM scale.
Short/brief Interview	Interview with questions about the previous tasks (Task 1, Task 2) and user satisfaction.	Semi-structured interview focused on gathering data about: • -Accessibility barriers • -User satisfaction
Task 3	Free navigation on the Discapnet website for 5 minutes.	Interaction data collected by the RemoTest tool: • -Visited web pages • -Time in each web pages • -Cursor movements, etc.
Task 4	Target searching task on the Discapnet website for a maximum of 10 minutes.	• -Completing the task • -Task completion time • -Interaction data collected by the RemoTest tool
Task 5	Filling in a questionnaire automatically generated by the RemoTest (stimuli S7) for collecting socio-demographic data.	• -Time for completing the task. • -Participants' socio-demographic data.
Short/brief interview	Semi-structured interview with questions about the overall user satisfaction with the tool.	Semi-structured interview focused on gathering data about: • -Accessibility barriers • -User satisfaction

perceptions about the PQ of the installation process and the second one for collecting participants' emotions during the installation process.

The first part of the questionnaire was based on the Attrakdiff questionnaire and consisted of a set of seven word pairs reflecting opposite adjectives to be rated on a 7-point Likert scale: technical-human, complicated-simple, impractical-practical, cumbersome-straightforward, unpredictable-predictable, confusing-clearly structured and unruly-manageable.

The second part of the questionnaire consisted of three questions to be rated on a 9-point scale based on the SAM: pleasure, arousal and dominance. The original SAM scale is a non-verbal pictorial assessment technique and so alternative texts (in Spanish) were added in order to make it accessible for the participants with visual disabilities.

The reliability of the used scales in this questionnaire was analyzed based on the Cronbach coefficient. For the PQ attributes all the Cronbach coefficients were greater than 0.70 indicating moderate-to-good reliability. On the contrary,

Table 5. Cronbach for the different scales that were used.

	PQ attributes	Emotions (SAM)
N items	7	3
Cronbach *α*	0.864	0.233

no reliable data could be gathered from the emotions scale as can be appreciated in Table 5. We found that some users did not understand the semantics of the images of the SAM scale or the alternative text provided. No further quantitative analysis was made with the values gathered with this scale. However, more information about the feelings during the installation was gathered in the interviews.

Each PQ attribute was analyzed separately and the results are presented in Table 6. It can be observed that the Corrected Item-Total Correlation for each item is favorable or positive as all the values are above 0.4 except one: the technical-human item (0.387). This result confirms that this item had less consistency compared with the other attributes. However, the general Cronbach's Alpha value increment is low when removing this item (from 0.864 to 0.883). These results suggest that the item has moderate reliability. During the interview, which was carried out after the installation task, some users commented that they had difficulties understanding the Technical-Human adjectives pair.

Participants tended to give high values to the PQ attributes. The median, mean confidence interval and standard deviation for each item are presented in Table 7. The mean of six of the attributes was higher than 5 (5.22–5.92). The technical-human question has the lowest value (4.25). This could be due to its moderate reliability, previously calculated with the Cronbach coefficient. The same tendency can be observed for the median and the standard deviation values.

The obtained values reveal a moderately good perception of the PQ of the installation process of the RemoTest tool.

5.6.1.1. Insights from the short interview. Once the participants installed the RemoTest tool and filled in the questionnaire a short interview was carried out to evaluate the process. They were asked to quantify the ease of the installation process based on a Likert 7-point scale. Only one participant said that the installation process was very complex. The 58.33% of the participants rated the installation process as simple or very simple (values 6 and 7). The mean value of the rates given by participants was 5.48. Figure 9 shows the values given by participants.

Table 6. Cronbach for the items of perceived usability/ PQ attributes.

Pragmatic quality (PQ)	Corrected item-total correlation	Cronbach's alpha if item deleted
Technical – human	.387	.883
Complicated – simple	.649	.844
Impractical – practical	.725	.831
Cumbersome – direct	.681	.838
Unpredictable – predictable	.774	.827
Confusing – clear	.641	.843
Unruly – manageable	.652	.842

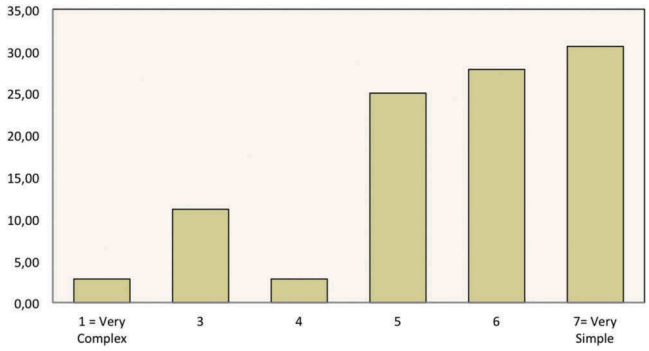

Figure 9. Graph showing the frequency of the answers to the question regarding the ease of the installation process.

In order to detect any cognitive barriers with the provided installation process instructions, participants were asked if they were easy to follow and understand. The results were positive since 24 of 36 of the users told us that the instructions were simple. Nevertheless, they did make some suggestions for improvement. The same suggestions from participants within the same user group were not obtained in all cases:

- Some suggestions were about the aesthetics and were not regarding the comprehension of the content. Two participants (P4 and P2) with physical disability commented on the text style. P4 stated that the text was too close together and could be quite confusing and P2 said that the font size was small. Two participants with low vision (P12 and P24) said that the font used was not very accessible due to the use of a font with serif. P12 suggested not using the whole screen for the text since going across from left to right tires people with low vision. Another participant also asked for more colorful instructions.
- One suggestion made by some participants was related to the difficulty they had to read all the instructions at once, since this required them to remember all the steps needed to install the tool. The blind participant P11 and the participant with low vision P12 asked for a step-by-step installation. P18, a blind participant, instead, asked for a shorter installation with fewer steps and simpler instructions. Participant P10, a participant with physical disability, had difficulties to understand the instructions and recommended using clearer and simpler language.

Participants were asked about accessibility barriers they found in the installation process and the responses given by those participants who rated the process as very complex or complex were thoroughly analyzed. Some of the barriers reported by participants are related to the Mozilla Firefox browser. In the case of accessibility barriers, coincidences in the answers of the participants corresponding to the same user group were obtained. These are the main barriers classified by user groups detected through the interview:

- Most of the screen reader users (P11, P18, P22, P25 and P26) did not notice the popup alert window opened by the browser agent in order to initiate the installation process.

Table 7. Mean, median and standard deviation for each of the PQ attributes.

	Technical – human	Complicated – simple	Impractical – practical	Cumbersome – direct	Unpredictable– predictable	Confusing – clear	Unruly– manageable
Mean (μ)	4.25	5.92	5.22	5.56	5.39	5.58	5.28
Lower limit	3.60	5.43	4.65	4.99	4.87	5.03	4.65
Upper limit	4.90	6.41	5.80	6.12	5.91	6.13	5.90
Median	4.00	7.00	6.00	6.00	6.00	6.00	6.00
Standard deviation (σ)	1.991	1.500	1.758	1.731	1.591	1.680	1.907

- Problems with popup alerts were also detected by low vision users using magnification software. The installation alert was positioned by Firefox in the upper left corner, which most of the time was out of their field of vision (P8 and P16). On the other hand, participants P12 and P14 saw the window but they did need more time to find it. Moreover, P12 had to change his strategy using the magnifier to decrease the zoom in order to access the alert window more easily. P23 said that the popup window size was too small and that she would prefer a bigger alert window in a centered position.
- One participant with physical disability (P29) proposed an improvement for stimuli S1 (installation process description). He suggested breaking the web page into smaller ones to avoid the use of scrolling.

Despite the accessibility barriers encountered by participants during the installation process, all of them were able to overcome them and install the tool without significant difficulties. The comments gathered in the interview are valuable for improving future versions of the RemoTest tool.

Chi-square and Fisher's exact tests were carried out in order to determine whether there were significant associations between variables of characteristics of users (user group, with or without disability, assistive technology used, age, Internet usage experience, Internet expertise level, Internet use frequency, Mozilla Firefox browser experience, etc.) and the opinions of the participants about the complexity-simplicity of installation process (Task 1). No statistical evidence of associations in the results was found.

5.6.2. Questionnaire completion tasks

Task 2 and Task 5 consisted of filling in some questionnaires. The questionnaires (stimuli S2 and S7) were automatically generated by the RemoTest tool based on the parameters specified by researchers when defining the experiment. The objective was to generate accessible questionnaires. All the participants were able to complete these tasks despite using different assistive technologies. This section explores the data gathered during these tasks such as the time required for filling in the questionnaires and any barrier detected by participants.

5.6.2.1. Time required for filling in the questionnaires.

Table 8 shows the time required by participants to fill in each of the questionnaires presented to users by Remotest during the experimental session.

Due to some technical problems the demographic questionnaire could not be presented to participant P32. These data were gathered through specific questions in the short interview.

Table 8. Time required to complete the questionnaires.

Participant	Task 2	Task 5	Total
P1	582	321	903
P2	638	269	907
P3	381	257	638
P4	230	208	438
P5	342	142	484
P6	280	92	372
P7	797	371	1168
P8	1784	543	2327
P9	993	377	1370
P10	756	485	1241
P11	262	183	445
P12	375	313	688
P13	207	143	350
P14	430	279	709
P15	648	235	883
P16	456	170	626
P17	299	212	511
P18	243	289	532
P19	182	90	272
P20	455	374	829
P21	329	161	490
P22	411	222	633
P23	270	80	350
P24	175	164	339
P25	316	192	508
P26	224	328	552
P27	229	158	387
P28	881	828	1709
P29	328	176	504
P30	184	263	447
P31	645	314	959
P32	143	–	–
P33	275	76	351
P34	148	39	187
P35	103	46	149
P36	272	58	330

Regarding the time required to fill in the questionnaires, P8, P28 and P9 needed appreciably more total time to complete both questionnaires: 2327, 1709, and 1370 seconds respectively. Participants P8 and P9 used a head pointer which takes considerably more time to point and click on the answers. In addition, video recordings of P8 showed that she needed longer to read the questions and response options than other participants. She also had some difficulties when answering the questions related to the PQ attributes in Task 2. Moreover, she clicked on the radio buttons in order to select her answers even though the clickable area was wider (in fact it encompassed all of the text of the answer as well as the radio button). Participant P28 was a blind user. He experienced some problems with the JAWS screen reader. He did not get any advice on the option selected in a question and sometimes the screen reader cursor returned to the beginning of the form and he had to navigate through all the questions he had already answered.

Table 9. Mean and standard deviation for the time required to complete the questionnaires by user group.

User Group	Task 2		Task 5	
	Mean (μ)	Standard deviation (σ)	Mean (μ)	Standard deviation (σ)
Blind	349.50	205.69	295.00	200.76
Low vision	362.13	151.84	183.88	84.79
Physical disability	610.77	430.36	293.69	129.29
Nondisabled	188.20	79.80	43.80	28.21

Video recordings of the participants using a screen reader were observed in order to detect any general barrier experienced by them. Results showed that participant P20 also experienced some inconvenience with the use of the screen reader, as it did not read out all the questions as apparently it skipped some of them. The participant became aware of this problem when an alert advising that all questions had to be filled in appeared when the "Continue" button was clicked. Even though these issues did not prevent participants from completing Task 2 and Task 5 they have to be analyzed and fixed for the next version of the tool.

Analyzing all of the questionnaires and all the participants it was seen that users required a mean time of 334 seconds to fill in a questionnaire. Table 9 presents the mean and standard deviation for the time required for each questionnaire by the different user groups.

As can be appreciated in the table, the time needed to complete the Task 2 questionnaire is higher than the Task 5 questionnaire for all the user groups. The user group of participants with physical disabilities obtained the slower mean value of 610.77 seconds for filling in the questionnaire in Task 2, but the standard deviation is also a very high value (430.36). Thus, the high time required by participants with physical disability is not data which gives us a central trend. The data scatter is due to certain specific participants such as P8 and P9 who obtained very high values due to using a head pointer and some problems they had to understand the PQ attributes.

The difference of the time distribution between user groups was able to be confirmed by running the Kruskal–Wallis one-way test. A rejection in the null hypothesis of independence (Task 2 ($\chi^2(3) = 8.6, p = .035$) and Task 5 ($\chi^2(3) = 13.9, p = .03$)) was obtained. The ranking average by user group was 7.5, 15.3, 17.8 and 23.5 for nondisabled people, people with blindness, with low vision and with motor impairments respectively in Task 2. While in Task 5 2.5, 14.3, 21.0, 22.8 ranking averages were obtained by nondisabled people, people with low vision, with blindness and with motor impairments respectively.

5.6.2.2. Insights from the short interviews. An interview about the participants' feelings about the questionnaire completion task was carried out just after filling in each questionnaire. This semi-structured interview consisted of some questions relating to the complexity of the completion process, previous experience with such kind of questionnaires and about barriers detected when filling them in. One participant (P15) was not interviewed as she performed the experimental session at her workplace and she was interrupted by a phone call. Nevertheless, she was able to complete both questionnaires without any noticeable problem. In total, 35 participants were interviewed about the questionnaire completion tasks.

We asked participants to rate the ease of filling in the questionnaire on a 7-point Likert scale. Figures 10 and 11 show the result of this question for the questionnaire in Task 2 and Task 5 respectively. The mean value for the questionnaire in Task 2 was 5.2 CI [4.61, 5.78]. It was rated as very simple (value 6–7) by 19 participants with adjusted-Wald 95% binomial confidence range (BC) of [38.18, 69.54%]. Only 6 from 35 participants gave a value lower than 4 BC [7.72, 33.06%]. Regarding the questionnaire in Task 5, the mean

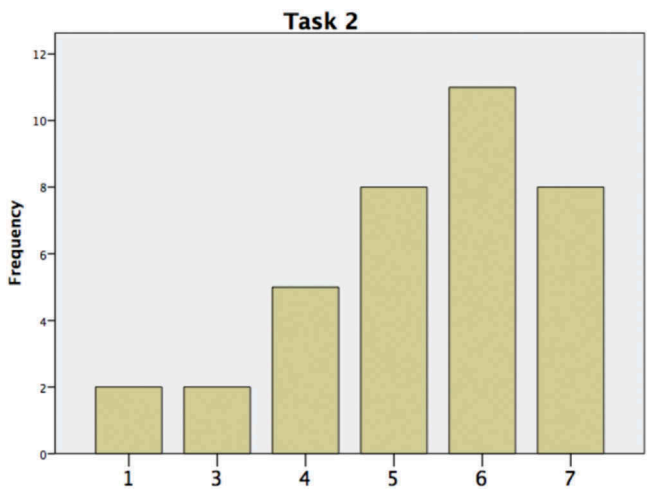

Figure 10. Ease score frequency for completing the questionnaire in Task 2.

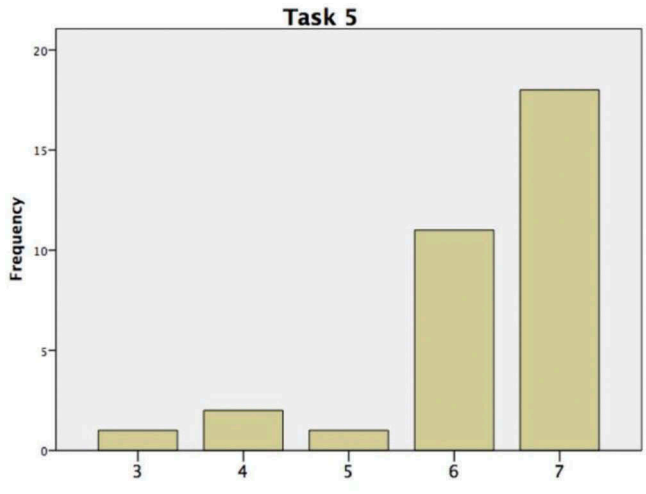

Figure 11. Ease score frequency for completing the questionnaire in Task 5.

value was 6 CI [5.61, 6.39]. Most of the participants, 29 out of 35, rated it as very easy (value 6–7) BC [66.94, 92.28%] . Only two participants rated it with a value lower than 4 BC [6.2, 19.57%].

All of them were able to access to all the content of the questionnaire in Task 2 but participant P14 with low vision had some problems with the figures due to her using a screen magnifier and losing the context of the images. 10 BC [16.19, 45.20%] participants had problems understanding some of the pairs of words of the PQ attributes, mainly the technical-human word pair. 6 BC [7.72, 33.06%] participants, on the other hand, stated that it was very simple and clear. It is worth mentioning that near the half of them (15 participants out of 35 BC [27.97, 59.16%]) had never filled in this kind of questionnaires before.

Regarding the questionnaire in Task 5, one participant commented on the screen reader cursor problem. Another participant reported some doubts about a question. In contrast to the previous questionnaire, most of the users had previously filled in similar questionnaires (29 participants from 35 BC [66.94, 92.28%]).

As in Task 1, Chi-square and Fisher's exact tests were carried out in order to determine whether there were significant associations between variables of characteristics of users and the opinions of the participants about the complexity-simplicity of Tasks 2 and 5. As a result, statistical evidence of associations were not found for either.

5.6.3. User satisfaction

The final interview of the experimental session was also intended to obtain participants' satisfaction and acceptance of the RemoTest tool. It was a semi-structured interview and participants were asked for their opinion about the tool and whether they would use it in remote settings. Participants were asked three questions to in order to obtain their opinion:

- Question 1: "What do you think about having a system for conducting inclusive experiments remotely?
- Question 2: "Would you participate in remote experiments? Would you encourage friends to participate?"
- Question 3: "Would you feel more comfortable doing the experiment remotely (e.g. from home, office, etc.)?"

All participants except P35 thought it would be interesting to have such a system to conduct experiments remotely. P35 expressed concerns about security. He did not intend to install the add-on on his computer as he thought such a system could get personal data from the system. However, he would try if the experimenters were people whom he trusted.

Most participants responded that they would participate in remote experiments and would encourage friends to do so:

- P28: "Sure, I have done some tests in remote before and everything was OK"
- P10: "It would be very interesting to carry out experiments from home and I would participate"
- P11: "I think it is amazing to have such a system to conduct experiments remotely. There is a lot of work to do and people often have a great sense of helplessness"

However, some participants indicated they would participate only if it did not take a long time:

- P21: "Yes, I would participate if I had enough time. I think that this is to help others and it is easier if you can do it from anywhere".
- P22: "Yes I would, depending on the time it would take"
- P5: "Yes I would if it is not difficult and it is not everyday"
- P6: "I don't know if I would be available to do tests at home"

Finally, other participants revealed some concerns about doing experiments in remote settings:

- P9: "I think that I could have some problems when installing and it would be necessary to provide good instructions"
- P34: "Yes I would participate if it is not too difficult and the experiments are helpful"
- P31: "Yes I would participate. If I had any problem I would email you"

Regarding Question 3, 16 out of 35 participants replied that they would be more comfortable performing the experiment at home or in the office:

- P2: "Yes I would be more comfortable without cameras and a tape recorder and I think I would be more efficient completing the tasks at home"
- P13: "I would prefer to do it at home because I have a huge screen and I see everything much better there"
- P14: "I would be more comfortable at home because I have all my tools there"
- P23: "The best place for me is the office because I have everything adapted"

Three participants (P11, P18 and P28) would be more comfortable in remote settings as long as they were provided with a chat system to resolve any problem encountered during the experimental sessions.

Another three participants (P7, P10 and P30) would prefer to carry out experiments in local settings. They claimed that they were more comfortable when the experimenters were on hand so they could ask about any doubt concerning the tasks.

The remaining participants responded that they would be comfortable in both a local and a remote setting.

5.7. Discussion

The results gathered served to explore the research questions defined for the experimental study. In relation to research question Q1, the results obtained revealed that the installation process of the RemoTest tool proved to be successful irrespective of the disability and the assistive technology used by

participants. All participants were able to install the tool even if some screen reader users and low vision users had to cope with problems when managing the popup windows which appeared during the installation process by the Mozilla Firefox browser. It is worth mentioning that Mozilla Firefox was not their usual web browser. This issue ought to be thoroughly tested for future experiments so that the screen reader used by each participant is proven to be compatible with the Mozilla Firefox version used.

The usability of the installation process was tested according to the PQ attributes. All participants gave positive scores to the seven pairs of adjectives displayed in the questionnaire of Task 2. Their feelings about the installation process were also positive. 83.33% BC [66.73, 92.51%] the participants rated the ease of the installation process with a value greater than 4 in a 7-point Likert scale. Statistical associations between variables of characteristics of users and their perception of the complexity of Task 1 were not found. This supports the assumption that the tool is accessible, simple and usable irrespective of the participants' characteristics and the assistive technology used.

However, some potential improvements on the presentation of the instructions for the installation were discerned during the interviews. There were some similar comments from participants within the same user group such as those from two participants with low vision (P12 and P24) who questioned the choice of text font. We also found opposing comments from participants from the same user group such as two blind participants (P11 and P18) one of whom requested a step-by-step installation process whereas the other asked for a shorter installation with fewer steps and simpler instructions. There were also some coincidences between participants within different user groups. These comments showed us the importance of including some mechanism of personalization so the installation and the presented stimuli can be adapted according to participants' preferences.

As regards research question Q2, the stimuli automatically generated by RemoTest tool proved to be accessible to all participants. All of them were able to fill in the questionnaires displayed by the tool and they rated them positively. According to the time required for completing the questionnaires, results revealed variations between user groups and among users within the same group. Participants in the group of physical disability required more time on average than others to fill in questionnaires but they showed high deviations in their results. These results led us to consider defining parameters in the configuration of tasks in future versions of the tool so that enough time is provided to each participant independent of their user group (this is related to Guideline 2.2 of WCAG 2.0). There were some participants using screen readers who reported some minor problems. They were able to fill in all the questions but the compatibility of the questionnaires with different versions of screen readers should be thoroughly analyzed and improved.

Participants gave a mean value of 5.2 out of 7 to the ease of filling in the questionnaire of Task 2 and a mean value of 6 to the questionnaire in Task 5. Moreover, there was no significant evidence found to contradict that these automatically generated questionnaires were accessible and simple to fill in regardless of the participants' characteristics.

The main concerns commented on by participants in the interviews were related to the semantics of the questions. This highlights the importance of using a clear and easy language, even more so when the studies are specifically focused on people with disabilities. In addition, some aspects of the design could be improved according to comments gathered in the interviews and these will be considered in future versions of the tool.

The lack of reliability of the SAM questionnaire about the feelings of participants during the installation might be due to the fact that some participants had problems understanding the semantic of the pictures or the alternative texts provided. The SAM questionnaire was selected because of its simplicity, however due to the results obtained the suitability of alternatives should be explored, such as, for example, the hedonic quality measure of Hassenzahl (2001).

Regarding research question Q3, the last interview revealed interesting data about user satisfaction with and acceptance of the RemoTest tool. All but one participant found this kind of tool interesting for conducting experimental sessions in local or remote settings. However, some of them showed some reluctance to perform remote experiments. Their main concerns were the length of time experiments would take and the problem solving mechanism during the sessions. They would appreciate some kind of support such as chat systems to ask for help if they are locked in any step or need some clarification about tasks. Nevertheless, most of them showed a positive attitude toward participating in other experiments even if they were to be remotely performed.

6. Conclusions

The need for remote web usability-testing tools in order to test web services in real contexts is growing. Several tools have been developed in the last decade and the most used ones are discussed in this paper. One of the common drawbacks of such tools is their lack of inclusiveness for people with disabilities. They have not been formally evaluated with users from different user groups.

We presented the RemoTest platform for designing, conducting and analyzing the data gathered in experimental sessions. It has been evaluated from the perspective of the experiment participants in a formal empirical study including 36 participants with different characteristics. Results revealed that all the participants, irrespective of their characteristics and the assistive technology used, were able to install the tool when provided with specific instructions. However, it was seen that the installation process could be improved by applying some of the suggestions made by participants. Future work will be focused on improving this process so it will be more personalized to tailor for the specific characteristics and the assistive technology used by participants.

Regarding the stimuli automatically generated by the RemoTest platform, results from the empirical study showed that they were accessible for a wide range of users. All participants were able to complete the questionnaires presented by the tool in a reasonable amount of time and most of the problems or barriers

detected by participants and pointed out in the interviews were more related to the complexity of the questions rather than difficulties encountered operating with the form controls. However, there were some compatibility issues between the assistive technology employed by users and the web browser version that should be considered in future versions of the platform. Moreover, based on feedback provided by participants during the interviews, some design aspects could also be improved such as providing shortcuts, bigger text and controls, numbering the questions, use of clear and simple language, etc. Future versions of the platform will incorporate these suggestions.

All in all, participants expressed their satisfaction with the platform and in general they were confident enough to take part in remote experimental sessions. One aspect to be taken into consideration in designing future versions of the platform is the possibility of dividing the experimental session into shorter sessions to encourage more participants to take part. However, the implications of such a feature at experiment design time would need to be thoroughly analyzed. Another potential feature suggested by participants was a mechanism, such as a chat system, for problem solving during experiment time. This will be also considered for future work.

Summarizing, in order to carry out an inclusive remote usability study with RemoTest or other remote usability tools, the following recommendations should be followed.

- Provide clear instructions about the installation or configuration of the user-testing tool including explanatory images when required
- Task descriptions, questionnaires, alarms etc. should be set up based on standards and be accessible in order to ensure their compatibility with the assistive technologies
- Texts must be short and clear and technical language should not be used
- Long experimental sessions should be divided into shorter sessions to avoid tiring the user and to encourage a greater number of participants to get involved.
- Due to the diversity of users, personalization features should be included, for instance allowing the preferred contrast, color, text size etc. to be set or allowing images, shortcuts etc. to be removed.

Notes

1. http://attrakdiff.de/index-en.html.
2. https://users.dimi.uniud.it/~giorgio.brajnik/projects/bw/bw.html.

Funding

This research work was developed within the project eGovernability, funded by the Spanish Government, Ministry of Economy, Industry and Competitiveness, and the European Regional Development Fund, under grant (TIN2014-52665-C2-1-R MINECO/FEDER). J.E.P. holds a PhD Scholarship from the University of the Basque Country (UPV/EHU). Some of the authors are members of the EGOKITUZ/ADIAN research team, supported by the Basque Government, Department of Education, Universities and Research under grant (IT980-16).

References

Albert, W., Tullis, T., & Tedesco, D. (2009). *Beyond the usability lab: Conducting large-scale online user experience studies*. Morgan Kaufmann, 1–314. ISBN: 9780080953854,

Apaolaza, A., Harper, S., & Jay, C. (2013). Understanding users in the wild. In *Proceedings of the 10th international cross-disciplinary conference on web accessibility*, 1–13.

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*, 203–212.

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*(1), 49–59. doi:10.1016/0005-7916(94)90063-9

Brajnik, G., & Giachin, C. (2014). Using sketches and storyboards to assess impact of age difference in user experience. *International Journal of Human-Computer Studies*, *72*(6), 552–566. doi:10.1016/j.ijhcs.2013.12.005

Claypool, M., Le, P., Wased, M., & Brown, D. (2001). Implicit interest indicators. In *Proceedings of the 6th international conference on intelligent user interfaces*, 33–40.

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

Etgen, M., & Cantor, J. (1999). What does getting WET (web event-logging tool) mean for web usability. In *Proceedings of fifth human factors and the web conference*.

Gajos, K. Z., Reinecke, K., & Herrmann, C. (2012). Accurate measurements of pointing performance from in situ observations. *Proceedings of the SIGCHI conference on human factors in computing systems*, 3157–3166.

Google analytics. 2018. Retrieved from http://www.google.com/analytics/

Hassenzahl, M. (2001). The effect of perceived hedonic quality on product appealingness. *International Journal of Human-Computer Interaction*, *13*(4), 481–499. doi:10.1207/S15327590IJHC1304_07

Hassenzahl, M. (2004). The interplay of beauty, goodness, and usability in interactive products. *Human-Computer Interaction*, *19*(4), 319–349. doi:10.1207/s15327051hci1904_2

Hassenzahl, M. (2005). The thing and I: Understanding the relationship between user and product. In Blythe, M. A., Overbeeke, K., Monk, A. F., & Wright, P. C., (Eds.), *Book funology: From usability to enjoyment* (pp. 31–42). ISSN 1571-5035, ISBN 978-1-4020-2966-0.

Hong, J. I., Heer, J., Waterson, S., & Landay, J. A. (2001). WebQuilt: A proxy-based approach to remote web usability testing. *ACM Transactions on Information Systems*, *19*(3), 263–285. doi:10.1145/502115.502118

Law, E. L. C., Roto, V., Hassenzahl, M., Vermeeren, A. P., & Kort, J. (2009). Understanding, scoping and defining user experience: A survey approach. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 719–728.

Leiva, L. A., & Vivó, R. (2013). Web browsing behavior analysis and interactive hypervideo. *ACM Transactions on the Web (TWEB)*, *7*(4), 1–28. doi:10.1145/2540635

LOOP11. (2018). Retrieved from http://www.loop11.com/

MORAE. (2018). Retrieved from http://www.techsmith.com/morae.html

Optimizely. (2018). Retrieved from https://www.optimizely.com/

Paganelli, L., & Paternò, F. (2002). Intelligent analysis of user interactions with web applications. *Proceedings of the 7th international conference on intelligent user interfaces*, 111–118.

Pérez, J. E., Arrue, M., Valencia, X., & Moreno, L. (2014). Exploratory study of web navigation strategies for users with physical disabilities. In *Proceedings of the 11th web for all conference*, 1–4.

Petrie, H., & Bevan, N. (2009). The evaluation of accessibility, usability and user experience. In Stephanidis, C. (Ed.), *The universal access handbook* (pp. 10–20). Boca Raton, FL: CRC Press Reference. ISBN 9780805862805

Santana, V. F., & Baranauskas, M. C. C. (2010). Bringing users of a digital divide context to website evaluation using WELFIT. In *Proceedings of the IX symposium on human factors in computing systems*, 31–40.

Scholtz, J., Laskowski, S., & Downey, L. (1998). Developing usability tools and techniques for designing and testing web sites. *Proceedings HFWeb*, *98*, 1–10.

USERZOOM. (2018). Retrieved from https://www.userzoom.com

Valencia, X., Arrue, M., Rojas-Valduciel, H., & Moreno, L. (2014). Interdependent components for the development of accessible XUL applications for screen reader users. *Webist*, *2*, 65–73.

Valencia, X., Pérez, J. E., Muñoz, U., Arrue, M., & Abascal, J. (2015, July). Assisted interaction data analysis of web-based user studies. *Human-Computer Interaction – INTERACT 2015* (Vol. 9296; pp. 1–19), Vancouver, BC. Springer International Publishing.

W3C, WAI. (2008, December 11). *Web content accessibility guidelines (WCAG) 2.0*. W3C Recommendation. Retrieved from http://www.w3.org/TR/WCAG20/

About the Authors

Myriam Arrue is an Assistant Professor at the University of the Basque Country/Euskal Herriko Unibertsitatea (UPV/EHU) and a member of the Egokituz Laboratory of HCI for Special Needs at UPV/EHU. Her research focuses on web accessibility evaluation, web navigation behavior analysis, and the application of transcoding techniques for improving UX in web interactions.

Xabier Valencia is a researcher at the Egokituz Laboratory of HCI for Special Needs (UPV/EHU). His main research interest is focused on providing higher accessibility to the Web to people with diverse capabilities, principally applying transcoding techniques.

J. Eduardo Pérez is a PhD student at the University of the Basque Country/Euskal Herriko Unibertsitatea (UPV/EHU) and a member of the Egokituz Laboratory of HCI for Special Needs at UPV/EHU. His research focuses on studying web navigation behaviors within diverse groups of users with special needs and testing different GUI enhancements in order to improve their web browsing ability.

Lourdes Moreno is Visiting Associate Professors in the Computer Science Department at Carlos III University of Madrid, Spain (UC3M), and researcher at the LaBDA Research Group in UC3M. Her research is focused on the design, development, and validation of accessible technology targeted to users with disabilities in different application domains.

Julio Abascal, BSD in Physics (U. de Navarra, 1978) and PhD in Informatics (University of the Basque Country-Euskal Herriko Unibertsitatea, 1987), is a Professor of the Computer Architecture and Technology Department of the University of the Basque Country, where he has worked since 1981. In 1985 he co-founded Egokituz Laboratory of Human–Computer Interaction for Special Needs which has participated in several R&D projects at national and international level.