Heuristic Evaluation of eGLU-Box:
A Semi-automatic Usability Evaluation Tool
for Public Administrations

Stefano Federici[1(✉)] , Maria Laura Mele[1] , Rosa Lanzilotti[2] ,
Giuseppe Desolda[2] , Marco Bracalenti[1] , Arianna Buttafuoco[1] ,
Giancarlo Gaudino[3], Antonello Cocco[3], Massimo Amendola[3],
and Emilio Simonetti[4]

[1] Department of Philosophy, Social and Human Sciences and Education,
University of Perugia, Perugia, Italy
stefano.federici@unipg.it, marialaura.mele@gmail.com,
marco.bracalenti90@gmail.com,
arianna.buttafuoco@gmail.com
[2] Department of Computer Science, University of Bari Aldo Moro, Bari, Italy
{rosa.lanzilotti,giuseppe.desolda}@uniba.it
[3] ISCOM – Superior Institute of Communication and Information Technologies,
Ministry of Economic Development, Rome, Italy
{giancarlo.gaudino,antonello.cocco,
massimo.amendola}@mise.gov.it
[4] Department of Public Service, Prime Minister's Office, Rome, Italy
e.simonetti@funzionepubblica.it

Abstract. This paper illustrates the heuristic evaluation of a web-based tool for usability testing for Public Administrations called eGLU-box. eGLU-box is an online platform aiming at supporting practitioners in the process of designing usability tests, analyzing data, and helping step-by-step participants to complete assessment tasks. Web users of Public Administrations can report their perceived quality of experience by completing a library of questionnaires shown to them by eGLU-box at the end of the test. This work is part of a multi-step user experience (UX) evaluation methodology to assess the platform. The UX evaluation methodology of eGLU-box uses standard and bio-behavioural evaluation methods. This work shows the results of the heuristic evaluation of eGLU-box involving five human factors experts and 20 practitioners working in Italian Public Administrations. Findings show that most of the problems are rated as minor problems and related to Nielsen's heuristic, "visibility of the system." Only 9% of problems are rated as major problems. These major problems are related to the "problematic match between system and the real world" heuristic. Evaluators provided indications for improvements that will be applied for the next version of the platform.

Keywords: User experience · Heuristic evaluation ·
Remote semi-automatic usability assessment · Usability assessment tools

1 Introduction

Remote usability evaluation tools are becoming more and more relevant in the field because they allow practitioners to reach wide and differentiated pools of test users at the same time, with lower costs and efforts than traditional laboratory usability approaches. Remote and automatic usability evaluation methodologies ask users to test web interfaces in their usual work or living environments while evaluators collect and analyze data by remotely controlled systems [1]. During the remote automatic assessment, evaluators do not directly monitor users during the interaction, but they remotely analyze their behavior collected by log files (e.g., [2, 3]). The limits of these tools are that (i) they often capture users' logs without analyzing them, (ii) they are not able to detect detailed information about users' actions such as facial expression recognition, (iii) they need to be pre-installed in the client's device (e.g., Morae, https://www.techsmith.com/morae.html), and (iv) they are often not platform-independent [4].

This work describes the heuristic evaluation of eGLU-box, a new remote semi-automatic usability assessment tool that overcomes each of the aforementioned limits. eGLU-box is a re-engineered version of a previous platform called UTAssistant [5–8], a web-based usability assessment tool developed to provide the Italian Public Administration with an online tool to conduct remote user studies. Both UTAssistant and its renewed version eGLU-box are designed according to usability guidelines provided by GLU, a group working on usability founded by the Department of Public Function, Ministry for Simplification and Public Administration in 2010. The latest version of the eGLU protocol (eGLU 2.1) was released in 2015 [9, 10].

The re-engineering process of UTAssistant was made possible by previous studies by Federici and colleagues, who evaluated user experience (UX) expert users of public administration (PA) websites [6]. In laboratory conditions, they used psychophysiological techniques [5] to measure the underlying reactions of participants through the recognition of facial expressions and electroencephalography (EEG). This work describes the usability evaluation of the renewed platform by a heuristic evaluation with both UX experts and PA practitioners. A heuristic evaluation is a usability assessment method in which an expert user performs a simulation of a typical user-system interaction with the aim of identifying critical points and weaknesses by means of heuristics. Heuristics in the UX context are simple and efficient rules that have been proposed since 1990 [11] to explain how people perceive, judge, and make decisions when facing interaction problems with a given system.

The rest of the paper is organized as follows. Section 2 describes the proposed usability assessment platform. Section 3 introduces the experimental methodology for the assessment of eGLU-box. Section 4 describes the results. Section 5 is for discussions, conclusions, and future directions.

2   From UTAssistant to eGLU-BOX: A Remote Usability Testing Tool for Public Administrations

Italian public administrations are the main public services that can benefit from simple and easy-to-use remote usability tools for assessing their websites. This is why in 2017 a web platform called UTAssistant was developed in line with the last Italian PA usability protocol, eGLU 2.1 [9]. Thanks to a UX evaluation of UTAssistant with expert users [6] and in laboratory conditions with two biobehavioral implicit measures [5], a re-engineering process of UTAssistant led to the development of the current version of the platform, eGLU-box. It is divided into two modules, one (the "tester module") for the practitioner who has to create, administer, and analyze the test, and another (the "end-user module") for end-users for whom the test is intended.

eGLU-box aims to facilitate evaluators in performing evaluation design activities such as creating a script, defining a set of tasks, or deciding which questionnaire to administer by means of three wizard procedures. Firstly, it guides the evaluators in specifying general information (e.g., a title, the script), data to gather during user task execution (e.g., mouse/keyboard data logs, webcam/microphone/desktop recordings), and post-test questionnaires to administer. The second procedure assists evaluators in creating the task lists and, for each task, specifying the starting/ending URLs, the goal, and the duration. The third procedure allows evaluators to decide which users to evaluate by selecting them from a list of users already registered to the platform or by typing their email addresses.

The second aim of eGLU-box is to help practitioners to manage all the tasks necessary for usability test execution such as emailing users with information, carrying out the evaluation test, capturing the session, and privacy policies regarding data from mouse/keyboard logs or webcam/microphone/desktop recordings. Each task is strongly guided by the platform, which shows the task description in a pop-up window and opens the web page from which users begin the test. To keep the platform as non-invasive as possible during the evaluation test, all the functions and indications (such as the current task goal and instructions, duration time, task number, and buttons to proceed to the next task or stop the evaluation) are grouped in a toolbar placed on top of the web page. Moreover, the button to proceed to the next task becomes "Complete Questionnaire" when the users finish the last task and must complete a questionnaire. During the task execution, the platform collects all data set by the evaluator in the study design in a transparent and non-invasive way.

eGLU-box automatizes all activities (such as collecting, storing, merging, and analyzing) related to data analysis, removing barriers in gathering usability test data. The evaluators access the data analysis results in their control panel by exploiting different tools that provide useful support in finding usability issues. The next sub-section provides an overview of the tools. Moreover, the platform calculates the task success rate (the percentage of tasks that users correctly complete during the test, which can also be calculated for each task estimating the percentage of users who complete that task) and visualizes them in a table in which the columns represent the tasks while the rows represent the users. The last row reports the success rate for each task, and the

last column depicts the success rate for each user. The global success rate is reported under the table.

When the test is concluded, summarizing the questionnaire results would improve the efficiency of evaluations. With eGLU-box, the evaluators can administer one or more questionnaires at the end of the usability evaluation. The platform automatically stores the user's answers and produces results by means of statistics and graphs.

The platform analyses audio-video information by collecting and storing the participants' voices with a microphone, their facial expressions with a webcam, and desktop activity with a browser plugin. The implemented player also provides an annotation tool so that when evaluators detect difficulties externalized with verbal comments or facial expressions, they can annotate the recorded audio/video tracks. If the evaluators decide to record both camera and desktop videos, their tracks are merged in a picture-in-picture fashion.

Finally, starting from the collected data, the platform shows performance statistics for each task as well as mouse and keyboard user logs.

Section 3 illustrates the heuristic evaluation for each of the above-described features of eGLU-box.

3 The Heuristic Evaluation of eGLU-Box

The study is a heuristic evaluation of the proposed web platform for remote and semi-automatic usability assessment of a PA website. The evaluation focuses on both the "tester module" and the "user module" of eGLU-box. Two groups of experts performed the evaluation, UX experts and PA practitioners involved in the design, development, and/or management of PA websites. As typically three to five UX experts are required to find 80% of usability issues [12], the group of experts is composed of four participants. A second group composed of 20 PA practitioners also performed the heuristic evaluation. The two groups followed the heuristic evaluation procedure as described in the following subsections.

3.1   Methods and Techniques

Heuristic evaluation is a usability evaluation method by which an expert evaluator performs a simulation of using a system to identify its critical points and weaknesses. The evaluation method uses heuristics, which are "simple and efficient rules that have been proposed to explain how people solve, make judgments, or make decisions about complex problems or incomplete information" (https://it.wikipedia.org/wiki/Euristica).

In a heuristic evaluation, the expert evaluator interacts with the system interface by simulating the actions and thoughts of a typical user following representative tasks. Tasks should guarantee that the system applies the main usability rules. For this study, we adopted the 10 Nielsen's heuristics (Table 1) [11], which consist of 10 principles derived from the factorial analysis of a list of 249 problems detected by numerous usability assessments. Nielsen's heuristics are used for the evaluation of the two modules of eGLU-box, the evaluator module and the end-user module.

Table 1. Nielsen's heuristics as described in the seminal work [12].

1.	**Visibility of system status.** The system should always keep users informed about what is going on through appropriate feedback within a reasonable timeframe
2.	**Match between system and the real world.** The system should speak the user's language with words, phrases, and concepts familiar to the user rather than system-oriented terms. It should follow real-world conventions, making information appear in a natural and logical order
3.	**User control and freedom.** Users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue. It should support undo and redo functions
4.	**Consistency and standards.** Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions
5.	**Error prevention.** Even better than helpful error messages would be a careful design that prevents a problem from occurring in the first place
6.	**Recognition rather than recall.** Make objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for the use of the system should be visible or easily retrievable whenever appropriate
7.	**Flexibility and efficiency of use.** Accelerators – unseen by the novice user – can often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions
8.	**Aesthetic and minimalist design.** Dialogues should not contain information that is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility
9.	**Help users recognize, diagnose, and recover from errors.** Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution
10.	**Help and documentation.** Although it is better if the system can be used without documentation, it might be necessary to provide help and documentation. Any such information should be easy to find, focused on the user's task, list concrete steps to be carried out, and not be too large

An impact rating is assigned to each problem. The problems are rated considering the frequency of the problem and the severity ratings assigned by the UX experts to each problem based on the scale shown in Table 2 [13].

Table 2. Nielsen's severity rating scale [13].

0	**Not a usability problem at all**
1	**Cosmetic problem only.** Need not be fixed unless extra time is available
2	**Minor usability problem.** Fixing this should be given low priority
3	**Major usability problem.** Important to fix. Should be given high priority
4	**Usability catastrophe.** Imperative to fix this before product can be released

The impact is calculated as the weighted average of the frequency of a problem and the related severity rating and defines how much a problem would affect the interaction with a system. Impacting problems with high frequencies and severity are then rated with high impact ratings.

3.2   Materials and Apparatus

During the assessment, both UX experts and PA practitioners are asked to report any violated heuristics in two grids, one for the assessment of the "tester module" and another for the assessment of the "end-user module." For each heuristic, the practitioner is asked to report any violated heuristics, the specific problem, the frequency of the problem, and suggestions on how to solve the issue. The grids were distributed in the spreadsheet file format. Participants are allowed to use their own laptop devices in indoor conditions. All tests were done in the Google Chrome web browser and the latest Windows operating system.

Instructions and tasks were created for the evaluation of the "tester module" and the "end-user module" and described as follows. The eGLU-box platform was tested on the PA website Ministry of Economic Development (MISE) (URL: https://www.sviluppoeconomico.gov.it).

3.3   Procedure

Both UX experts and PA practitioners are asked to evaluate the two main modules composing eGLU-BOX, the tester module and the end-user module. During each evaluation, participants are asked to compile the provided spreadsheet file format.

The expert is asked to evaluate the tester's experience for each of the following tasks.

Task 1 – Tester Module. Create a new test – Basic information.

- Login
- Create a new study, name it "Usability evaluation of the MISE website"
- Set website URL: to http://www.sviluppoeconomico.gov.it
- Set description, name it "Evaluation of UTAssistant"
- Ensure anonymity for the participants
- Set the following input peripheral devices to capture: microphone, webcam, and desktop
- Set duration to 5 min
- Select questionnaires to be administered to users at the end of the test: NPS, UMUX, and SUS

Task 2 – Tester Module. Define tasks for end-users.

- Create four tasks by setting a title, description, maximum duration, and start and end URLs (tasks for end-users are reported below in this section)

Task 3 – Tester Module. Invite participants to the test.

- The tester should invite participants to their test

Task 4 – Tester Module. Open the test to users.

- Save the test and make it public to participants
- Logout
- Once the test is created, the expert is asked to evaluate the end-users' experience for each of the following tasks. Before starting, end-users should log in to the eGLU-box platform and agree to participate in the test. The heuristic evaluation begins at the login phase.

Task 1 – End-User Module. *Scenario 1.* You inherited a farmhouse in Italy from your elderly grandmother. You go to the MISE website to search for information about the energy redevelopment of your type of building, and you come to know that the Ministry of Economic Development allows you to benefit from a bonus of tax deductions in the case of interventions on individual units. To find out more, you need to:

1. Identify the website area related to the incentives for the citizen;
2. Identify the deadline for requesting incentives in case of interventions on the single real estate units;
3. Look for the designated office to call for more information about it.

    - Start URL: https://www.sviluppoeconomico.gov.it/
    - End URL: https://www.sviluppoeconomico.gov.it/index.php/it/component/organigram/?view=structure&id=563.

Task 2 – End-User Module. *Scenario 2.* You are a non-EU citizen. In your country of residence abroad, you have obtained a professional hairdresser qualification. You would like to do your profession in Italy. You go the MISE website to search for information about it. To find out more, you need to:

1. Identify the list of recognized professional qualifications;
2. View the most up-to-date document regarding the recognition of your profession.

    - Start URL: https://www.sviluppoeconomico.gov.it/
    - End URL: https://www.sviluppoeconomico.gov.it/images/stories/impresa/mercato/MethnaniMounir.pdf.

Task 3 – End-User Module. *Scenario 3.* You are a journalist for a national information magazine for agricultural companies. For the article you are working on, you are looking for the 2018 statistics on rice export authorizations. Starting from the MISE homepage, you need to:

1. Identify the existence of a page containing the import/export statistics of agri-food products;

2. Look for statistics relevant to the export of rice issued in 2018.

- Start URL: https://www.sviluppoeconomico.gov.it/
- End URL: https://www.sviluppoeconomico.gov.it/images/stories/commercio_internazionale/agroalimentare/statistiche_2018/export_2018.pdf.

Task 4 – End-User Module. *Scenario 4.* A friend of yours recently told you that the Italian citizens are provided with an online information tool that monitors fuel prices in both Italy and Europe. You want to know more about it. Starting from the MISE homepage, you try to:

1. Identify the actual existence of an observatory for fuel prices;
2. Identify the existence of an observatory for fuel prices in Italy.

- Start URL: https://www.sviluppoeconomico.gov.it/
- End URL: https://carburanti.mise.gov.it/OssPrezziSearch/.

3.4   Subjects

Four UX experts (50% female, mean age 33 years old) and 20 practitioners working for the Public Administration in the field of website development and management (70% female, mean age 52 years old) participated in the study. The experimental sessions were conducted separately for UX experts, who accessed the eGLU-box platform through the servers of the University of Perugia and in a single session in a laboratory with multiple workstations through the PA servers.

4   Results

Severity ratings were computed for the usability problems identified for the tester module and the end-user module, respectively, as explained in Sect. 3.1.

The numbers of heuristic violations for the two modules assessed by the two groups of experts according to Nielsen's heuristics were calculated (Fig. 1). In the tester module of eGLU-box, heuristics were violated 19 times. The violated heuristics were "1. Visibility of system status" (68.4%), "5. Error prevention" (26.3%), and "3. User control and freedom" (5.3%). No other heuristics were violated in the tester module. In the end-user module, heuristics were violated 31 times. The most frequently violated heuristics were "1. Visibility of system status" (35.5%), "3. User control and freedom" (22.6%), and "4. Consistency and standards" (19.4%), followed by "2. Match between system and the real world" (9.7%), "5. Error prevention" (9.7%), and "7. Flexibility and efficiency of use" (33.2%). Across both modules, the heuristic "1. Visibility of system status" comprised 48% of all violations.
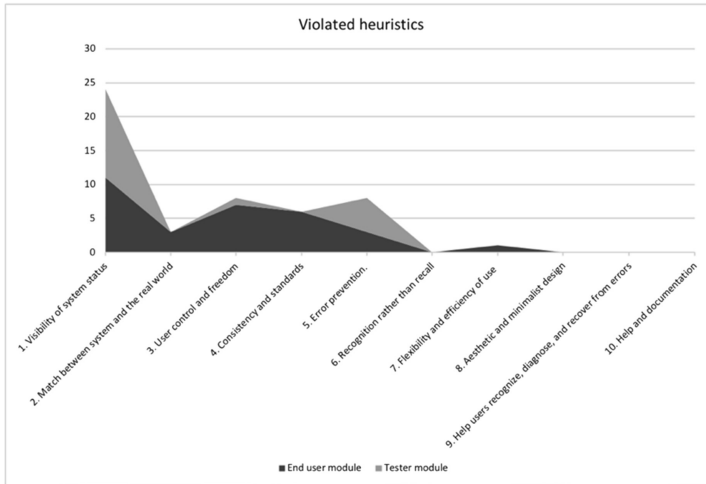
Fig. 1. Violated heuristics for both end-user and tester modules.

Comparisons of the two modules show no significant difference in the frequency of usability problems between the two modules (t(8) = 1.315, p > .05), meaning that the two modules have an equal chance of generating errors.

Figure 2 shows the level of severity of the usability problems found in both the tester and end-user modules. The severity ratings calculated for the tester module revealed 20% major severity problems, 60% minor severity problems, and 20% cosmetic usability problems. In the end-user module, 100% of problems were rated as minor usability problems. Overall, the problems belonging to the most violated heuristics in both modules (i.e., "1. Visibility of system status," 48% of all the violations) were rated as minor usability problems.
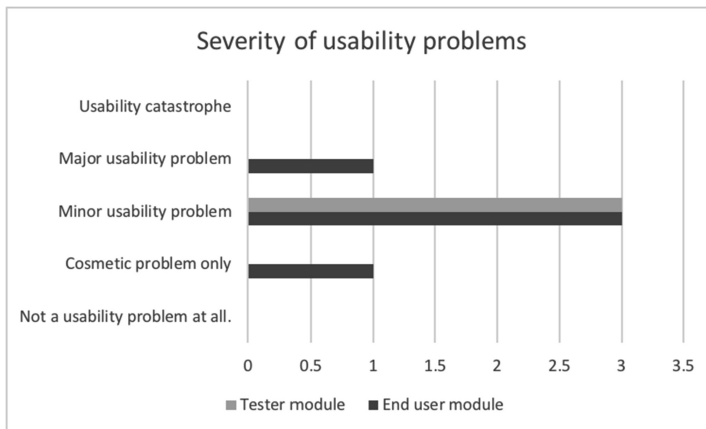

Fig. 2. Severity of the usability problems found for both modules as defined by the UX experts.

Figure 3 illustrates the impact of the problems found for both modules. Impact was calculated as described in Sect. 3.1. The results show that minor usability problems related to the first Nielsen's heuristic "1. Visibility of system status" have an impact of 24.5% on the interaction during the test. These problems are related to misleading/missing feedback that notifies end-users/testers about the effects of their actions – for example, "notify end-users with clear feedback as soon as each test ends." Major problems related to the second heuristic "2. Match between system and the real world" had 22.3% of the impact. This result shows that although the second Nielsen's heuristic was violated only 9.7% of all violated heuristics, it might have a significant impact on the end-user's experience. This problem is mostly related to the tester module and refers to a missing mark warning that all fields during the creation of a test are mandatory. Comparisons of the two modules show no significant difference in the impact of usability problems between the two modules (t(9) = 0.922, p > 0.05).
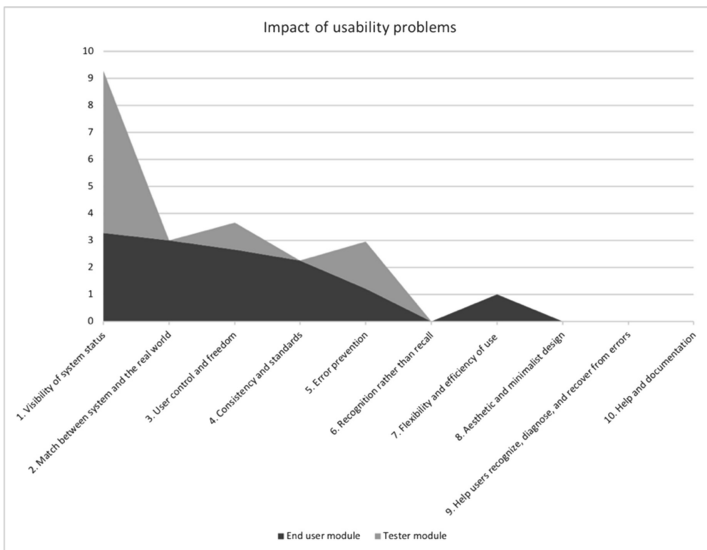

Fig. 3. Impact of the usability problems found for both modules calculated by considering severity and frequency of each problem.

The evaluators' suggestions for improvement were collected for each problem. The areas of the platform that received the most advice from evaluators were grouped into five parts: (1) preliminary actions to access the platform (e.g., access via link and login), (2) the homepage (i.e., exploration of the functions displayed in the user's dashboard), (3) areas for the creation of a test or agreement to participate, (4) invitation of users to conduct a test or participate in a test, and (5) the end of a test (i.e., feedback to users after completing the test or reports of users' results to testers). In the tester module, the areas with the most usability problems and the corresponding indications of improvements concern the end of part 3. The problems with the highest impact on

usability were related to the heuristic "5. Error prevention" and are related to missing/unclear feedback during the invitation of users to take part in the test. In the end-user module, the areas with the highest number of suggestions for improvement concern the end of part 5. These problems are linked to the heuristic "1. Visibility of system status" and are mostly linked to missing feedback to users after completing each task or the whole test.

5   Conclusions

This paper proposed the heuristic evaluation of eGLU-box, a web-based platform for the remote and semi-automatic evaluation of websites. The platform is specifically designed for Italian Public Administrations, and it follows the latest usability guidelines for PAs as recommended by the eGLU 2.1 protocol. eGLU-box is divided into two modules specifically designed to be used by two main categories of users: the tester who creates and administers usability tests and the end-user who performs the required tasks. eGLU-box overcomes the main issues related to remote automatic tools for usability assessment because it (i) is able to capture users' logs and provide pre-processed analysis reports; (ii) detects information about users' interactions through the main peripheral devices such as microphones, webcams, and desktops; (iii) does not need to be pre-installed; and (iv) is platform-independent. eGLU-box is the result of a re-engineering process on a previous version of the platform called UTAssistant subsequent to its UX evaluation with PA practitioners and in laboratory conditions with bio-behavioral measures. The heuristic assessment of eGLU-box involved two groups of evaluators, UX experts and PA practitioners. The evaluators highlighted the critical points and weaknesses of the system by using the 10 heuristics provided for the first time by Nielsen and Molich [11]. Results show that the modules of the system have an equal chance of generating usability errors. No differences in the extent to which errors impacted the interaction were found. Most of the problems found in both modules were minor usability problems (60%) related to visibility of the system status. Major problems (20%) were due to functions that are more system-oriented than user-oriented in the way they are presented to users. Updates and error fixes of the system will be performed by following the evaluators' suggestions for improvement provided in this study. Future work will focus on the assessment of the updated platform with end-users recruited via an online platform.

References

1. Andreasen, M.S., Nielsen, H.V., Schrøder, S.O., Stage, J.: What happened to remote usability testing? An empirical study of three methods. In: SIGCHI Conference on Human Factors in Computing Systems: CHI 2007, pp. 1405–1414 (2007). https://doi.org/10.1145/1240624.1240838

2. Obendorf, H., Weinreich, H., Hass, T.: Automatic support for web user studies with scone and tea. In: Extended Abstracts on Human Factors in Computing Systems, CHI 2004, pp. 1135–1138. ACM, Vienna, AT (2004). https://doi.org/10.1145/985921.986007

3. Atterer, R.: Logging usage of Ajax applications with the "Usaproxy" Http Proxy. In: Workshop on Logging Traces of Web Activity: The Mechanics of Data Collection, pp. 1–4 (2006)

4. Sauer, J., Sonderegger, A., Heyden, K., Biller, J., Klotz, J., Uebelbacher, A.: Extra-laboratorial usability tests: an empirical comparison of remote and classical field testing with lab testing. Appl. Ergon. **74**, 85–96 (2019). https://doi.org/10.1016/j.apergo.2018.08.011

5. Federici, S., Mele, M.L., Bracalenti, M., Buttafuoco, A., Lanzilotti, R., Desolda, G.: Bio-behavioral and self-report user experience evaluation of a usability assessment platform (UTAssistant). In: International Conference on Human Computer Interaction Theory and Applications, HUCAPP 2019, pp. 73518 (2019)

6. Federici, S., et al.: UX evaluation design of utassistant: a new usability testing support tool for italian public administrations. In: 20th International Conference on Human-Computer Interaction, pp. 55–67 (2018). https://doi.org/10.1007/978-3-319-91238-7_5

7. Catarci, T., et al.: Digital interaction: where are we going? In: Proceedings of the 2018 International Conference on Advanced Visual Interfaces, AVI 2018, pp. 1–5 (2018). https://doi.org/10.1145/3206505.3206606

8. Desolda, G., Gaudino, G., Lanzilotti, R., Federici, S., Cocco, A.: UTAssistant: a web platform supporting usability testing in italian public administrations. In: CHItaly, CHItaly 2017, 12th edn., pp. 138–142 (2017)

9. Dipartimento della Funzione Pubblica: Il Protocollo Eglu 2.1: Come Realizzare Test Di Usabilità Semplificati Per I Siti Web E I Servizi Online Delle Pa. Formez PA, Rome, IT (2015)

10. Borsci, S., Federici, S., Mele, M.L.: Eglu 1.0: un Protocollo Per Valutare La Comunicazione Web Delle Pa. Diritto e Pratica Amministrativa, vol. 1, pp. 9–10. Il Sole 24 Ore, Milan, IT (2014)

11. Nielsen, J., Molich, R.: Heuristic evaluation of user interfaces. In: SIGCHI Conference on Human Factors in Computing Systems: CHI 1990, pp. 249–256 (1990). https://doi.org/10.1145/97243.97281

12. Nielsen, J.: Heuristic evaluation. In: Nielsen, J., Mack, R.L. (eds.) Usability Inspection Methods. Wiley, New York (1994)

13. Nielsen, J., Mack, R.L. (eds.): Usability Inspection Methods. Wiley, New York (1994)