













eGLU-Box Mobile: A Smartphone App for Usability Testing by Italian Public Administration Webmasters

Stefano Federici^{1,4} , Giovanni Bifulchi¹ , Marco Bracalenti¹ ,
Alessandro Ansani^{1,2} , Agnese Napoletti¹ , Rosa Lanzillotti³ ,
Giuseppe Desolda³ , Maria Laura Mele⁴ , Simone Borsci⁵ ,
Maria Laura de Filippis¹ , Giancarlo Gaudino⁶, Massimo Amendola⁶,
Antonello Cocco⁶, Aldo Doria⁶, and Emilio Simonetti⁷

¹ Department of Philosophy, Social and Human Sciences and Education, University of Perugia,
Perugia, Italy

`stefano.federici@unipg.it, alessandro.ansani@uniroma3.it`

² Department of Philosophy, Communication, and Performing Arts, Roma Tre University,
Rome, Italy

³ Department of Computer Science, University of Bari Aldo Moro, Bari, Italy
{`rosa.lanzillotti, giuseppe.desolda`}@uniba.it

⁴ Myèsis, Research and Development Company, Rome, Italy

⁵ Department of Learning, Data Analysis, and Technology – Cognition, Data and
Education – CODE group, Faculty of BMS, University of Twente, Enschede, The Netherlands

⁶ DGTCISI-ISCTI – Directorate General for Management and Information and Communications
Technology, Superior Institute of Communication and Information Technologies, Ministry of
Enterprises and Made in Italy, Rome, Italy

`{giancarlo.gaudino, massimo.amendola, antonello.cocco,
aldo.doria}@mise.gov.it`

⁷ Rome, Italy

Abstract. Smartphones and tablets now offer consumers unique advantages such as portability and accessibility. Developers are also working with a mobile-first approach, and are prioritizing mobile applications over desktop versions. This study introduces eGLU-box Mobile, an application for performing a drive usability test directly from a smartphone. An experimental study was conducted in which the participants were divided into two groups: an experimental group, which used the new mobile application from a smartphone, and a control group, which used the desktop application from a computer. The participants' behavior was assessed using explicit (self-report questionnaires) and implicit measures (eye movement data). The results were encouraging, and showed that both the mobile and desktop versions of eGLU-box enabled participants to test the usability with a similar level of UX, despite some minimal (although significant) differences in terms of satisfaction of use.

Keywords: Mobile · Application · App · Computer · Desktop · eGLU-box · Eye-Tracking · Eye-Tracker · Usability

E. Simonetti—Independent Researcher.

1 Introduction

Mobile devices such as smartphones and tablets are now commonly used. Oulasvirta and colleagues [1] have reported that smartphones are more widely available and are used more often throughout the day than laptops. In addition, smartphones provide faster access to content than laptops, due to their portability. In view of this, the present study compares a mobile version of eGLU-box developed for the Italian public administration (PA), a national government, with the original version for desktop use. eGLU-box PA (in the following eGLU-box) is a web platform that allows PA webmasters in Italy to evaluate the usability of their websites and digital services [2]. It was developed based on the eGLU LG 2018.1 protocol, and is designed to run on a personal computer or Mac [3]. The eGLU-box Mobile application, which is available for both Android and iOS systems, was developed to guide these webmasters to carry out semiautomatic evaluations directly from their smartphones. In this study, we aim to evaluate the user experience of eGLU-box Mobile compared to the desktop version (eGLU-box) by observing the users' implicit (eye movement) and explicit (satisfaction, cognitive workload, and promotability) behaviors. This study forms part of the eGLU-box Mobile project supported by the former Ministry of Economic Development, now known as the Ministry of Enterprises and Made in Italy (MIMIT), together with the universities of Perugia and Bari. The aim of the project was to create a new mobile application that would allow users to use eGLU-box on devices such as smartphones and tablets. Before its launch, the application needs to be tested under laboratory conditions with participants randomly selected from the population.

1.1 eGLU-box: From Its Inception to Today

Introduction to eGLU-box. eGLU-box is software developed by the PA in collaboration with the universities of Perugia and Bari (Italy), which allows for the evaluation of the usability of websites. It is an online tool that allows PA webmasters to create usability tests for a particular website and to invite participants to carry them out. eGLU-box is a re-engineered version of a previous platform called UTAssistant [4–6], a web-based usability assessment tool that was developed to provide the PA with an online tool to conduct remote user studies. Both UTAssistant and its newer version, eGLU-box, were designed according to usability guidelines provided by GLU, a group working on usability that was founded by the Department of Public Function, Ministry for Simplification and Public Administration in 2010. Currently, eGLU-box is based on eGLU LG version 2018.1, the latest version of the protocol [3].

The eGLU-box LG Protocol: Usability Guidelines for Everyone. Version 2018.1 of the eGLU LG protocol specifies the procedure that a tool should implement to investigate the usability of a product. This is a generic tool, as the protocol is defined independently of the technology. This means that it can be applied with minimal adjustment to a variety of products and services on different distribution channels and with different technologies, such as information websites, online services, paper documents, desktop applications (for computers), and mobile applications (for smartphones and tablets). This protocol was created with the aims of: (i) describing a procedure to promote the direct

involvement and observation of users in the evaluation of online sites and services; and (ii) encouraging public operators to pay greater attention to the issue of usability. The user observation procedure consists of five steps: (i) definition of the tasks (e.g., searching for specific information, filling in online forms, downloading documents) to be carried out by participants, which is done by the observer; (ii) selection of users; (iii) performance of the assigned tasks by users (during the observation, direct questions from the observer are not admitted); (iv) administration of user experience questionnaires when the tasks have been executed; (v) data analysis based on the quantitative or qualitative data collected. If carried out correctly, the entire procedure can be considered a minimum usability test, albeit simplified, and can be performed by non-experts. The protocol also provides the PA webmaster with an idea of the possible problems with interaction with the PA website and online services.

eGLU-box: From Theory to Practice. When the detailed instructions provided by the eGLU LG protocol version 2018.1 are followed, the eGLU-box platform allows even a non-expert in usability testing to conduct a usability test. eGLU-box reports the time taken for each task, its outcome, the results of the questionnaire, any registration or tracing performed by the participant as part of the task, etc. Via a single web platform, eGLU-box allows for the merging of data with different natures, which an observer would otherwise have to collect using different technologies and software, such as screen recordings, task durations, task outcomes, etc. To perform a usability test with eGLU-box, an observer (i.e., a webmaster or experimenter) accesses the platform as a “user-creator” in order to define the tasks to be performed by the users on the specific website under evaluation (Fig. 1). Furthermore, the user-creator can select one or more self-report questionnaires to be completed by the user at the end of the test, in order to measure: (i) usability, based on the System Usability Scale (SUS) [7] and UMUX-LITE [8]; and (ii) promotability, based on the Net Promoter Score (NPS) [9]. The user-creator also has the ability to add new questionnaires via the interface.

When the user-creator has set up the website that will be evaluated, the tasks and the questionnaires to be carried out, each user taking part in the usability evaluation is invited to access eGLU-box as a “user-tester”. This interface provides users with a step-by-step guide to navigating the website under evaluation, and displays the tasks and questionnaires set up by the user-creator. The user’s actions are recorded by eGLU-box through a webcam, screen recordings, and a microphone. eGLU-box was designed for the conduction of usability tests both remotely and in a laboratory. In this case, eGLU-box data can be combined with software that captures bio-behavioral data (such as electroencephalographic, skin conductance, heart rate, eye movement, and facial expression data) [5].

In 2022, a version of eGLU-box was developed for mobile testing (designed for both Android and iOS systems) to guide the webmasters of the PA to carry out a semiautomatic evaluation directly from smartphones and tablets.

The screenshot displays the 'eGLU-box PA' application interface. At the top, there is a navigation bar with 'Home', 'Create Test', 'Settings', and 'Logout (Mario)'. Below this is a header for 'Edit test' with two tabs: 'Main information' and 'Task definition'. The 'Task definition' tab is active, showing a task titled 'Task 1' with a 'Delete' button. The task details are as follows:

Field	Value
Title *	Send an email
Estimated duration (in minutes) *	5 minutes
Initial URL *	https://mail.google.com/mail/u/0/#inbox
URL to reach to complete the task correctly *	https://mail.google.com/mail/u/0/#inbox?compose=new
Instructions *	Connect to your google account and send an email.

At the bottom of the task definition form, there is a green '+ Add task' button. Below the form, there are two buttons: a grey 'Back' button and a green 'Save Test' button.

Fig. 1. A screenshot of eGLU-box (desktop version) showing the task creation screen. This figure shows an example created previously, in which the user-tester is asked to send an email via Gmail.

This study aims to evaluate the user experience with eGLU-box Mobile compared to the desktop version (eGLU-box), by observing users' implicit behaviors (eye movements) and explicit reactions (satisfaction, cognitive workload, and promotability) to the elements of this evaluation tool. The main goal was not to assess the usability of the specific interface through which participants carried out their tasks (i.e., the MIMIT website: <https://www.mise.gov.it/>), but to compare the experience of participants using the web and mobile versions of eGLU-box. As a semiautomatic assessment system, eGLU-Box was also designed to be used remotely by participants, with the application providing all the necessary instructions on the screen to allow participants to start and complete a usability test on their own.

2 Methods

2.1 Design

The study was designed as a between-subject experiment. The participants were divided into two groups: an experimental group (EG), which used the new eGLU-box Mobile application on a smartphone (Samsung Galaxy S8) to assess the MIMIT website, guided by the eGLU-box indications, and a control group (CG), which used the desktop version of the eGLU-Box on a computer (Lenovo ThinkPad T540p) to perform a usability test of the same website.

2.2 Material

A sociodemographic questionnaire was administered that asked participants about their age, the most frequently used device (i.e., computer or smartphone), device assigned by the experimenter. Furthermore, a series of generic questions were presented with the aim of investigating their use of smartphones and computers, such as the average period of use of the device in a typical day, the purposes of usage, etc.

The participants in the two groups were asked to interact with the MIMIT websites to achieve four tasks, which were structured as follows: (i) imagine you are the founder of a female start-up and you need to identify possible concessions for the “female business creation” category; (ii) imagine you want to buy a new vehicle and you want to look for the section that deals with the bonus that allows you to receive a refund for the purchase of low-emission vehicles; (iii) imagine you are the owner of a small farm that needs a supply of fuel, and you want to know its price; and (iv) imagine that you are the owner of a radio station that has been granted access to local broadcaster contributions, and you want to find a document containing a list of contribution amounts. All of the tasks had a maximum duration of 5 min.

After the interaction with the interface, three self-report questionnaires were administered in order to measure: (i) usability, through the SUS [7]; (ii) cognitive workload, through the Nasa Task Load Index (NASA TLX) [10]; and (iii) promotability, based on the NPS [9] for the eGLU-box application. The Partial Concurrent Thinking Aloud (PCTA) technique [11, 12] was used, in which participants were required to silently interact with the interface and ring a bell whenever they detected a problem; this bell represented a reminder signal, with the aim of aiding memorization of the moment at which the participant encountered the problem. The experimenter then needed to record what the participant was doing when the bell rang. As soon as the participant finished the test, they were invited to discuss and verbalize the problems encountered during the interaction [13]. Hence, at the end of the experimental procedure, a short interview was conducted in which the participant was asked how the interaction with the interface had gone in general, the reasons that had prompted them to ring the bell, and the observations they made regarding the application.

2.3 Participants

Forty-nine students attending the University of Perugia (eight males, 41 females) with an average age of 21.67 years (min = 18; max = 50; SD = 5.90) were recruited through social networks and academic mailing lists. A description of the sample can be seen in Table 1.

Table 1. Description of the sample, divided by sex and mean age

Group	Sex	N (%)	Mean age
<i>Experimental</i>	Male	4 (16.7%)	22.00
	Female	20 (83.3%)	21.60
	Total	24 (100%)	21.67
<i>Control</i>	Male	4 (16%)	21.50
	Female	21 (84%)	22.86
	Total	25 (100%)	23.20

In their answers to the sociodemographic questionnaire, 42 participants (85.7%) stated that they spent more time on a smartphone in a typical day, while seven (14.3%) declared that they spent more time on the computer. A total of 44 participants (89.9%) reported that they used a smartphone more often as a device for surfing the Internet, while five (10.2%) stated that they mainly used a computer for this. Twenty-two participants (44.9%) stated that they would feel more motivated to participate in an online questionnaire using a smartphone, while 27 (55.1%) said that they would prefer to use a computer. The answers to these questions can be viewed in Table 2.

Table 2. Questions about smartphone and computer use (responses from 49 participants)

Question	Answer	N (%)
<i>In a typical day, which device do you spend more time on?</i>	Smartphone	42 (85.7%)
	Computer	7 (14.3%)
<i>In a typical day, which device do you use most often to search the internet?</i>	Smartphone	44 (89.4%)
	Computer	5 (10.2%)
<i>Which device would you be more motivated to use to participate in an experimental study (online questionnaire)?</i>	Smartphone	22 (44.9%)
	Computer	27 (55.1%)

Participants were asked how much time they thought they spent on a smartphone and a computer in a typical day (Table 3). They reported using a smartphone between 2.5 and 3 h, while a computer was used between 1.5 and 2 h. When they were asked about their likelihood of agreeing to participate in an online study using a smartphone (Table 4), the data suggest that none of them would refuse to participate and all 44 would participate, whereas if they were asked to participate using a computer, one person would not participate and 43 would.

Table 3. Answers to the question: “How much time do you currently spend on your computer/smartphone in a typical day?” (responses from 49 participants)

Answer	Device	N (%)
None	Smartphone	0
	Computer	3 (6.1%)
Between 1 and 30 min a day	Smartphone	0
	Computer	5 (10.2%)
Between 60 (1 h) and 90 min (1 h 30 min)	Smartphone	6 (12.2%)
	Computer	8 (16.3%)
Between 90 min (1 h 30 min) and 120 min (2 h)	Smartphone	7 (14.3%)
	Computer	6 (12.2%)
Between 120 min (2 h) and 150 min (2 h 30 min)	Smartphone	6 (12.2%)
	Computer	7 (14.3%)
Between 150 min (2 h 30 min) and 180 min (3 h)	Smartphone	9 (18.4%)
	Computer	5 (10.2%)
Between 180 min (3 h) and 210 min (3 h 30 min)	Smartphone	5 (10.2%)
	Computer	5 (10.2%)
Between 210 min (3 h 30 min) and 240 min (4 h)	Smartphone	9 (18.4%)
	Computer	2 (4.1%)
More than 240 min (4 h)	Smartphone	7 (14.3%)
	Computer	5 (10.2%)

Table 4. Answers to the question: “How likely is it that you would accept an invitation to an experimental study (online questionnaire) using a computer/smartphone?” (responses from 49 participants)

Answer	Device	N (%)
Very unlikely	Smartphone	0
	Computer	0
Unlikely	Smartphone	0
	Computer	1 (2%)
Neutral	Smartphone	5 (10.2%)
	Computer	5 (10.2)
Likely	Smartphone	17 (34.7%)
	Computer	15 (30.6%)
Very likely	Smartphone	27 (55.1%)
	Computer	28 (57.1%)

2.4 Procedure

The experiment was conducted in a laboratory at the University of Perugia. The experimenter provided participants with a brief introduction to the study, randomly assigned each of them to a group (EG or CG), and assigned them an email for use in connecting to the eGLU-box. Eye-tracker calibration was performed only on participants in the CG at this time. The participants then opened the Qualtrics.xm (Provo, UT, USA) platform, via which the informed consent and the privacy policy were displayed and the sociodemographic questionnaire and the questionnaire on the frequency of smartphone and computer usage were administered. Next, the participants started to carry out the tasks through the eGLU-box platform. Before accessing the eGLU-box platform, eye-tracker calibration was performed for participants in the EG group. The tasks were the same for both groups, and eGLU-box guided participants to carry out these tasks on the MIMIT website. When these were complete, participants from both groups returned to Qualtrics to complete three self-report questionnaires (SUS, NASA TLX, and NPS) about the eGLU-box interface they had used to receive instructions, tasks and scenarios and to perform the overall assessment in a guided modality. Finally, participants were interviewed by an experimenter based on the PTCA procedure. Implicit data from eye movements were collected using a Tobii Pro Nano eye-tracker.

2.5 Data Analysis

The duration of the experiment with eGLU-box was calculated for all participants, and data from 11 participants were discarded as they were considered to be outliers. The time spent on navigation in eGLU-box was calculated by Qualtrics, using a widget that counted how much time the participant required before going to the next page. Although each participant was instructed to move to the next page only after completing the tasks in the eGLU-box, some proceeded before the task was concluded, and these data were not considered in the analysis. For this reason, the average navigation data may differ from those reported in the section on eye movement analysis. For the tasks in the eGLU-box, three possible outcomes were defined: (i) completed, i.e., the participant achieved the required goal; (ii) not completed, i.e., the participant did not reach the required goal; and (iii) missing, i.e., code problems with the eGLU-box application were encountered during the procedure (e.g., forced closure) that prevented execution, which allowed us to find possible bugs. The SUS results were transformed into grades [14, 15] ranging from F (absolutely unsatisfactory) to A+ (absolutely satisfactory). Mann-Whitney U test was conducted to find a possible significant difference in the SUS results for the two groups (EG and CG), and the total score was used as the test variable. For the NPS questionnaire, the participants' scores were transformed to classify them as promoters (scores of nine or 10), passives (scores of seven or eight) and detractors (scores of between zero and six). The NPS is calculated as the percentage of promoters minus the percentage of detractors. Mann-Whitney U test was used to find a possible significant difference in the results obtained in the NPS questionnaire for the EG and CG, and the participants' responses were used as the test variable. For the NASA TLX, the Mann-Whitney U test was applied to find a possible significant difference in the final results of the scales for the two groups (EG and CG), and the total score for each scale was used as the test variable. The eye-tracking data were analyzed using Tobii Pro Lab software.

For the EG, which used a smartphone, a mobile testing accessory (MTA) was used to ensure the tracking and collection of eye movement data. The implicit measures analyzed were the area of interest (AOI), mean webpage observation time, and ocular fixation (with a threshold set to ≥ 50 ms). To perform the analyses, data from participants with a gaze sample percentage of greater than or equal to 75% were used. Eye movement data were analyzed for only 12 participants from the EG, as only these met the requirements. The mean values for the number of fixations and time of visualization of the AOIs were analyzed. AOIs were inserted by the Tobii Pro Lab software. Eye movement data were analyzed only during the interaction with the elements of eGLU-box. In the analysis phase, five AOIs were identified for the mobile application, and eight for the desktop application (see Fig. 2).

IBM SPSS version 27 software was used for analysis of the questionnaire data, and Tobii Pro Lab software version 1.207 was used for the eye movement data.

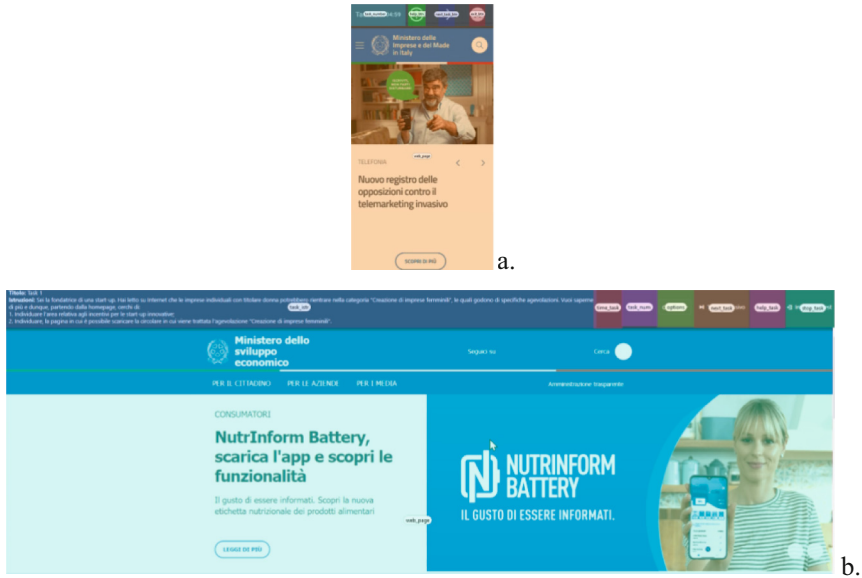


Fig. 2. Visualization of the eGLU-box platform: (a) smartphone mode; (b) computer mode. Both screens show the AOIs as colored squares or rectangles.

3 Results

3.1 Participants' Performance

On average, each participant took 6 min to complete the eGLU-box tasks, with a minimum time of 5 min and a maximum of 11 min. Of the 38 participants, 20 were allocated to the EG that used the smartphone. On average, the participants in the EG spent 6 min, with a minimum time of 5 min and a maximum of 11 min. The CG consisted of 18 participants, who used a computer to perform the tasks. On average, the participants in the CG spent 6 min, with a minimum time of 5 min and a maximum of 9 min. Mann-Whitney U test was conducted to determine whether there was a difference in the total time spent to perform tasks in eGLU-box between the two groups (EG and CG). The results indicate non-significant difference ($U = 123, p = .096$). A regression analysis of task achievement suggested that there were no significant differences between the EG and CG in terms of success or failure in achieving the tasks. Table 5 summarizes the performance of the participants in the two groups for each task, including missing data. On average, 40.8% of participants succeeded in carrying out the four tasks, 48.8% failed, and in about 10.7% of the cases there was a technical issue.

Table 5. Percentages of tasks completed by participants in the experimental (EG) and control (CG) groups

Task	Group	Completed	Not completed	Missing
Task 1	EG	0	18 (75%)	6 (25%)
	CG	2 (8%)	23 (92%)	0
	Total	2 (4.1%)	41 (83.7%)	6 (12.2%)
Task 2	EG	17 (70.8%)	3 (12.5%)	4 (16.7%)
	CG	23 (92%)	2 (8%)	0
	Total	40 (81.6%)	5 (10.2%)	4 (8.2%)
Task 3	EG	4 (16.7%)	15 (62.5%)	5 (20.8%)
	CG	10 (40%)	15 (60%)	0
	Total	14 (28.6%)	30 (61.2%)	5 (10.2%)
Task 4	EG	9 (37.5%)	9 (37.5%)	6 (25%)
	CG	15 (60%)	10 (40%)	0
	Total	24 (49%)	19 (38.8%)	6 (12.2%)

3.2 PCTA Interview

In the CG, 17 participants rang the bell, with a total of 22 events. However, the difficulties reported were related less to the functioning of the platform than to the difficulty of completing the tasks requested within the tested website. In fact, only two of 22 events were related to critical issues regarding the procedure, and these were particularly associated with the difficulty of understanding the requests made in one of the questionnaires (NASA-TLX). In the EG, 17 participants rang the bell, with a total of 38 events. In this case, most of the events concerned a malfunction or critical aspects of the application. Specifically, of these 38 events, three were related to the non-functioning of the hyperlinks in the tasks set by the application. In one event, the application crashed at the start of one of the tasks, requiring the user to close and reopen it. In 15 events, the participants complained that they could not read the task instructions again when they were doing it. Of these subjects, none had clicked on the button marked with the current task number, which would have displayed the instructions. Instead, the button these subjects had clicked was the one marked “?”, as they expected to be able to see the instructions from there. In one event, it was reported that the participant expected to be shown the on-screen instructions again each time he performed one of the requested tasks. Another event occurred when a participant exited the application by mistake. By clicking on the icon again, the participant could restart the task they were carrying out; however, the timer did not restart from the point where it left off, and it was necessary to click on the task number to make the instructions reappear and then resume the task, with the timer restarting from 5 min. Nine of the events referred to difficulties in completing the tasks requested for the website. In one event, the instructions did not disappear from the screen even though the 5-min timer had started, making it impossible to start the

task. In two events, the exact opposite occurred, i.e., the task began without the 5-min timer starting. In one event, it was not possible to move from one task to the next after communicating to the application whether or not the previous task had been completed successfully. Another participant complained that in the top bar, the application showed an excessive number of stimuli that were confusing while carrying out the tasks. In a further event, it was recommended to change the way the application works. Instead of having the participant press the button to finish the task, the application should itself show a message on the screen when the participant has reached the goal and thus finished the task. Finally, two events involved a sudden crash of the application, which required the participant to log in again and start the test from the beginning.

3.3 Usability Questionnaire

System Usability Scale (SUS). The results for the SUS questionnaire showed an average score of 63.78, with a minimum of 25.00, a maximum of 87.50 and a standard deviation of 14.43. The mean value for the EG was 58.02, with a minimum of 25.00, a maximum of 87.50 and a standard deviation of 14.28, while the mean value for the CG was 69.30, with a minimum of 40.00, a maximum of 87.50 and a standard deviation of 12.49. Mann-Whitney U test between the EG and CG showed a significant difference ($U = 160$, $p = .005$), meaning that these participants found it more satisfying to use eGLU-box from a computer than a smartphone. In the EG, eight participants (33.3%) were assigned Grade F, eight participants (33.3%) were assigned Grade D, four participants (16.7%) were assigned Grade C, one (4.2%) was assigned Grade C+, one (4.2%) was assigned Grade B, one (4.2%) was assigned Grade A-, and one (4.2%) was assigned Grade A+. In the CG, two participants (8%) were assigned Grade F, six (24%) were assigned Grade D, four (16.7%) were assigned Grade C, three (12%) were assigned grade C+, two (8%) were assigned grade B, one (4%) was assigned grade B+, two (8%) were assigned grade A-, three (12%) were assigned grade A, and two (8%) were assigned grade A+. Figure 3 shows a graph of the data divided based on the device used.

NASA TLX. The NASA TLX results showed the following weighted average values for the six scales: mental 210.20, physical 15.83, temporal 155.26, performance 124.48, effort 128.85, frustration 95.57. For the EG, the results for the weighted average values were: mental 214.37, physical 1.66, temporal 140, performance 144.37, effort 108.95 and frustration 95.41. For the CG, the results for the weighted average values were: mental 206.20, physical 2.20, temporal 101.60, performance 100.40, effort 142.80, frustration 76.60. For each scale, the Mann-Whitney U test was applied to identify a difference between the two groups. The results were not significant for any of the scales.

Net Promoter Score. At a general level, 25 detractors (51.0%), 20 passive (40.8%) and four promoters (8.2%) were identified, giving a final value for the NPS of -42.8. In the EG, 15 detractors (62.5%), eight neutrals (33.3%) and one promoter (4.2%) were identified, with an NPS value of -58.3. In the CG, 10 detractors (40%), 12 neutral (48%), and three promoters (12%) were found, giving an NPS value of -28. The data suggest that the participants would be more likely to recommend the desktop version over the smartphone version, although neither of the results were very high. Figure 4 shows the

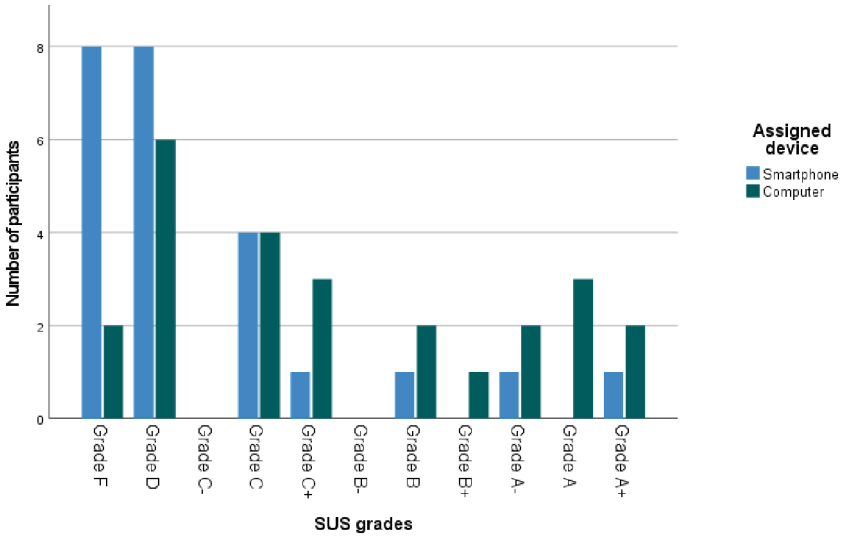


Fig. 3. Graph showing the number of participants with each grade on the SUS questionnaire. Participants were divided based on the device assigned, i.e., smartphone (EG) and computer (CG).

percentages of detractors, passives, and promoters for each device. Mann-Whitney U test was applied to the NPS scores for the CG and the EG, but no significant difference was found.

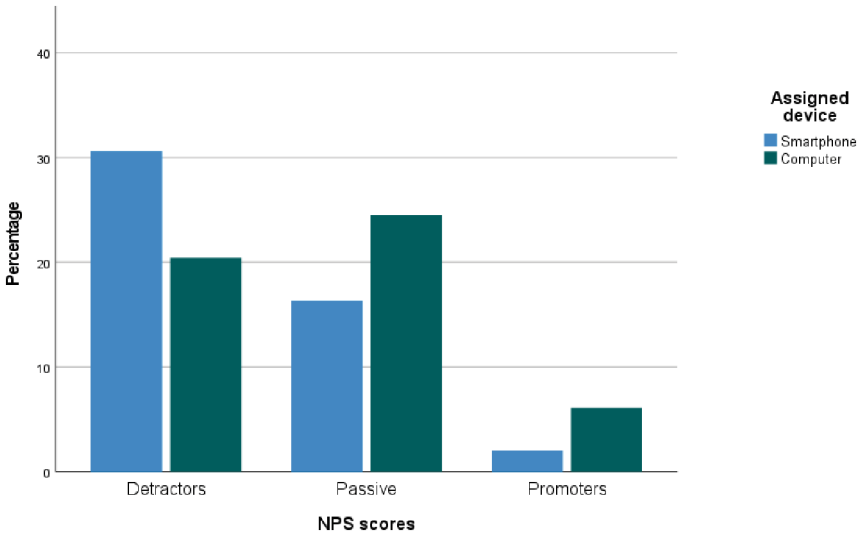


Fig. 4. Graph showing the percentages of detractors, passives, and promoters for each device, i.e., smartphone (EG) and computer (CG).

3.4 Eye Movement Analysis

Several AOIs were inserted for the smartphone device: one in the application bar, one around the help button, one around the button used to go to the next task, one around the exit button, and finally, one around the task button, which showed the time remaining and the task instructions if it was clicked. In the computer version, the following AOIs were inserted: one in the bar where the task instructions were displayed, one where the remaining time was visible, one around the task number button, one around the options button, one around the button to switch to the next task, one around the help button, one around the exit button, and finally, one around the web page. Participants looked at the webpage screen for an average of 11 min, with an average of 3,311 fixations. The eGLU-box bar was viewed for a mean total of 20 s, with a mean of 68 fixations. The exit button was viewed for only 2 s on average, with an average total of six fixations, while the help button was viewed for an average of 4 s and an average of 14 fixations. The next task button was looked at for an average of 3 s and an average number of fixations of 10, while the task number button was viewed for an average of 8 s and an average of 30 fixations.

All data from the CG met the requirement to be included in the analyses. CG participants looked at the web page for an average of 12 min with an average of 2,595 fixations. The help button was viewed for an average of 1 s with an average of three fixations. The button used to move to the next task was looked at for an average of 5 s with an average of 13 fixations, while the options button was viewed for an average of 3 s with an average of 12 fixations. The exit button was looked at for an average of 1 s and an average of two fixations, and the task instructions button was looked at for an average of 2 min and an average of 316 fixations. The task number button was looked at for an average of 4 s and an average of 14 fixations, whereas the time remaining button was viewed for an average of 4 s and an average of 14 fixations.

4 Discussion and Conclusion

In this study, we have reported pilot data from the development of eGLU-box Mobile, a version of eGLU-box for Android and iOS systems. This was designed to guide a PA webmaster to perform a semiautomatic evaluation directly from a smartphone. The aim of this study was to evaluate the user experience with eGLU-box Mobile compared to the desktop version (eGLU-box), by observing users' implicit (eye movement) and explicit (satisfaction, cognitive workload, and promotability) behaviors. From the results of the questionnaires, it can be seen that that the two applications are interchangeable.

Most participants (90%) reported spending more time interacting with a smartphone than a computer, even when browsing the Internet. In regard to which device they would be more motivated to use to answer a web questionnaire, the participants expressed a greater preference for computers (55.1%), but smartphones also achieved a significant proportion of the vote (44.9%). The same result was found when we asked how likely they would be to agree to participate in a study with a smartphone or a computer. The results were very similar and positive, indicating that a good percentage would agree to participate in both cases. It therefore seems that the participants are inclined to use smartphones as well as computers.

For the tasks performed using eGLU-box, a regression analysis did not reveal significant differences between the EG and CG participants in terms of success or failure. This indicates that both applications allow for correct use of the product. Furthermore, the results from the NPS and NASA TLX questionnaires were shown to be fairly similar, and the Mann-Whitney U test showed no significant differences between the participants in the EG and CG. This suggests that both applications gave the same results in terms of both promotability and cognitive load. Only the results of the SUS questionnaire were different, and the Mann-Whitney U test showed that the results were statistically significant. In this case, better results were found for those who had performed the tasks from the computer (CG) in terms of satisfaction with using the application.

From the PCTA interview, we discovered that one of the problems with the smartphone application may have been related to the absence of a button that clearly indicated where to find the task instructions, as these are clearly visible in the desktop application without the need to click anything. It could be deduced from the average viewing data for the AOIs that the participants in the CG looked at the area showing the instructions for the task more often, while the smartphone group almost never looked at these. This may be because it was not clear to the participants that clicking the task number button would also display the instructions. We can conclude that although the new application can be officially launched, there is still a need to make some improvements to the usability and satisfaction, such as the implementation of a clearer button that allows the user to read the task instructions again.

This study has some limitations that should be highlighted, such as the size of the sample, the homogeneity of the sample in terms of gender, and the limitations imposed by the collection of eye movements for mobile devices. For participants in the EG, recording of eye data started at the moment they logged into the eGLU-box application, while for CG participants eye data recording started immediately after device assignment. A decision was made to proceed in this way because data collection from a smartphone is much less sensitive than from a computer, which requires that the participant does not move from the position in which the calibration was performed. Therefore, in order not to require too much effort from the participants, it was decided to record the eye movements only in the part of our interest, i.e., when the eGLU-box was being used. However, this required the participant to stop the test to calibrate the eye-tracker. This procedure was not carried out for the CG, in which calibration was performed at the beginning of the experiment. This was because the sensitivity of the eye-tracker when using a computer is very high and the participant must remain motionless. It is also necessary to consider that the smartphone experiment was much more stressful than the computer one, since the capture of eye movements is much more rigid in this mode and the participant must move as little as possible.

In future studies, we intend to expand the sample size, and to include not only students but also the rest of the population.

References

1. Oulasvirta, A., Rattenbury, T., Ma, L., Raita, E.: Habits make smartphone use more pervasive. *Pers. Ubiquit. Comput.* **16**, 105–114 (2011). <https://doi.org/10.1007/s00779-011-0412-2>

2. Federici, S., et al.: Heuristic evaluation of eGLU-box: a semi-automatic usability evaluation tool for public administrations. In: Kurosu, M. (ed.) HCII 2019. LNCS, vol. 11566, pp. 75–86. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-22646-6_6
3. AGID (Agenzia per l'Italia Digitale): Linee guida di design per i servizi web della pubblica amministrazione. AGID, Rome, IT (2022)
4. Desolda, G., Gaudino, G., Lanzilotti, R., Federici, S., Cocco, A.: UTAssistant: a web platform supporting usability testing in Italian public administrations. In: 12th Edition of CHIItaly: CHIItaly 2017, pp. 138–142 (2017)
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: VISIGRAPP 2019: Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. Volume 2: HUCAPP, pp. 19–27 (2019)
6. Federici, S., et al.: UX evaluation design of UTAssistant: a new usability testing support tool for Italian public administrations. In: Kurosu, M. (ed.) HCI 2018. LNCS, vol. 10901, pp. 55–67. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91238-7_5
7. Borsci, S., Federici, S., Lauriola, M.: On the dimensionality of the System Usability Scale (SUS): a test of alternative measurement models. *Cogn. Process.* **10**, 193–197 (2009). <https://doi.org/10.1007/s10339-009-0268-9>
8. Lewis, J.R., Utesch, B.S., Maher, D.E.: UMUX-Lite: when there's no time for the SUS. In: Conference on Human Factors in Computing Systems: CHI 2013, pp. 2099–2102 (2013). <https://doi.org/10.1145/2470654.2481287>
9. Reichheld, F.F.: The one number you need to grow. *Harv. Bus. Rev.* **82**, 133 (2004)
10. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (task load index): results of empirical and theoretical research. In: Hancock, P.A., Meshkati, N. (eds.) *Human Mental Workload*, pp. 139–184. North-Holland, Amsterdam (1988)
11. Federici, S., Borsci, S., Mele, M.L.: Usability evaluation with screen reader users: a video presentation of the PCTA'S experimental setting and rules. *Cogn. Process.* **11**, 285–288 (2010). <https://doi.org/10.1007/s10339-010-0365-9>
12. Borsci, S., Federici, S.: The partial concurrent thinking aloud: a new usability evaluation technique for blind users. In: Emiliani, P.L., Burzagli, L., Como, A., Gabbanini, F., Salmiinen, A.-L. (eds.) *Assistive Technology from Adapted Equipment to Inclusive Environments: AAATE 2009*, vol. 25, pp. 421–425. IOS Press, Amsterdam (2009). <https://doi.org/10.3233/978-1-60750-042-1-421>
13. Borsci, S., Kurosu, M., Federici, S., Mele, M.L.: Computer systems experiences of users with and without disabilities: an evaluation guide for professionals. CRC Press, Boca Raton (2013). <https://doi.org/10.1201/b15619-1>
14. Borsci, S., Federici, S., Bacci, S., Gnaldi, M., Bartolucci, F.: Assessing user satisfaction in the era of user experience: comparison of the SUS, UMUX and UMUX-Lite as a function of product experience. *Int. J. Hum.-Comput. Interact.* **31**, 484–495 (2015). <https://doi.org/10.1080/10447318.2015.1064648>
15. Sauro, J., Lewis, J.R.: *Quantifying the User Experience: Practical Statistics for User Research*. Morgan Kaufmann, Burlington (2012)