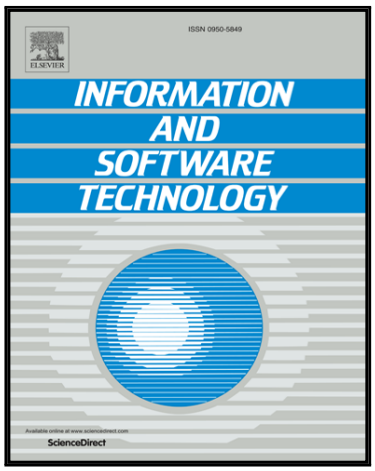


Impact of usability mechanisms: an experiment on efficiency, effectiveness and user satisfaction

Juan M. Ferreira , Silvia T. Acuña , Oscar Dieste , Sira Vegas , Adrián Santos , Francy Rodríguez , Natalia Juristo

PII: S0950-5849(19)30202-2
DOI: https://doi.org/10.1016/j.infsof.2019.106195
Reference: INFSOF 106195

To appear in: *Information and Software Technology*

Received date: 30 November 2018
Revised date: 17 July 2019
Accepted date: 20 September 2019

Please cite this article as: Juan M. Ferreira , Silvia T. Acuña , Oscar Dieste , Sira Vegas , Adrián Santos , Francy Rodríguez , Natalia Juristo , Impact of usability mechanisms: an experiment on efficiency, effectiveness and user satisfaction, *Information and Software Technology* (2019), doi: https://doi.org/10.1016/j.infsof.2019.106195

***Highlights**

- We study the effect of the Abort Operation, Progress Feedback and Preferences usability mechanisms on efficiency, effectiveness and user satisfaction.
- We conducted a between-subjects experiment with orthogonal array using 168 non-computer scientists performing one task for each adopted or non-adopted usability mechanism in an online shop.
- The adoption of Abort Operation has a significantly positive effect on efficiency (time taken), effectiveness and user satisfaction.
- 4. The adoption of Progress Feedback does not appear to have any impact on any of the usability characteristics.
- The adoption of Preferences has a significantly positive effect on effectiveness and user satisfaction but no influence on efficiency.
- We found that the adoption of all three mechanisms tends to improve system usability and does not degrade user performance.

Impact of usability mechanisms: an experiment on efficiency, effectiveness and user satisfaction

Juan M. Ferreira[1], Silvia T. Acuña[2], Oscar Dieste[3], Sira Vegas[3], Adrián Santos[4], Francy Rodríguez[2], Natalia Juristo[3]

[1]Facultad Politécnica, Universidad Nacional de Asunción, CC 2111, San Lorenzo, Paraguay
jmferreira1978@fpuna.edu.py
[2]Universidad Autónoma de Madrid, Calle Francisco Tomás y Valiente 11, 28049 Madrid, Spain
{silvia.acunna, francy.rodriguez}@uam.es
[3]Universidad Politécnica de Madrid, Campus de Montegancedo, 28660 Boadilla del Monte, Madrid, Spain
{odieste, svegas, natalia}@fi.upm.es
[4]M3S ITEE, University of Oulu, Finland
adrian.santos.parrilla@oulu.fi

Abstract

Context. As a software quality characteristic, usability includes the attributes of efficiency, effectiveness and user satisfaction. There are several recommendations in the literature on how to build usable software systems, but there are not very many empirical studies that provide evidence about their impact. **Objective**. We report an experiment carried out with users to understand the effect of three usability mechanisms —Abort Operation, Progress Feedback and Preferences— on efficiency, effectiveness and user satisfaction. Usability mechanisms are functionalities that should, according to the HCI community, be implemented within a software system to increase its usability. **Method**. The experiment was conducted with 168 users divided into 24 experimental groups. Each group performs three online shopping tasks. We measure efficiency variables (number of clicks and time taken), effectiveness (percentage of task completion) and user satisfaction gathered from a questionnaire. **Results**. The adoption of Abort Operation has a significantly positive effect on efficiency (time taken), effectiveness and user satisfaction. The adoption of Progress Feedback does not appear to have any impact on any of the variables. The adoption of Preferences has a significantly positive effect on effectiveness and user satisfaction but no influence on efficiency. **Conclusions**. We provide relevant evidence of the impact of the three usability mechanisms on efficiency, effectiveness and user satisfaction. In no case do the usability mechanisms degrade user performance. The effort to adopt Abort Operation and Preferences appears to be justified by the benefits in terms of effectiveness and user satisfaction. Also Abort Operation enables the user to be more productive. We believe that the effects on efficiency, effectiveness and satisfaction depend not only on mechanism functionality but also on the problem domain. The impact of a mechanism in other contexts could differ. Therefore, we need to conduct further experiments to gather more evidence and confirm these results.

Keywords. Usability, usability mechanism, empirical study, experimental design, efficiency, effectiveness, satisfaction, software engineering

1 Introduction

Software system quality is decomposed into several characteristics, including usability [36]. Usability is a critical characteristic of highly interactive systems [2, 16, 42, 68, 73]. Despite its importance, however, most software products still have limited usability [21, 23, 25, 33, 72, 78]. The ISO/IEC 25010 [34] standard defines usability as "the degree to which a product or system can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use". Using this definition, it is possible to measure usability perceived in terms of efficiency, effectiveness and user satisfaction [34].

Human-computer interaction (HCI) proposes recommendations for achieving desirable levels of software system usability [2, 6, 73, 76, 80, 10, 12, 16, 30, 47, 53, 55, 64]. For example, a user that enters an unwanted scenario (interface) should be able to navigate backwards out of that scenario. Backward navigation could be demanding [76]. Therefore, a Cancel button should be provided to close all the open windows and reject any possible changes made by the user. As there are a lot of HCI recommendations, standards and models have been put forward to unify the usability concept and account for all the usability features covered by HCI recommendations [27, 28]. One objective of unifying the concept of usability is to be able to evaluate a software product to establish its usability level. Another goal is to provide tools for adopting usability in the development process.

Software engineering experimentation conducts empirical studies to evaluate both final system usability and how to improve individual usability attributes and characteristics during the development process. There are empirical studies on: usability evaluation and inspection techniques [13, 32, 56], the use of special-purpose tools to evaluate usability [66, 77], the validation of usability in particular application types, like mobile applications [33], domain-specific languages [58] and security libraries [5], and problems with the use of interface design patterns [60]. There are also empirical studies focusing on web sites These studies evaluate the usability of different navigation designs [14, 18] or gather empirical evidence on the relationship between web site usability and brand perception [11].

There are works that provide empirical evidence on the relationship between software design and usability [44]. This research identifies usability recommendations with a high impact on software design. Such recommendations are different from guidelines that affect only the graphical user interface since recommendations with impact on software design involve the construction of additional software components. Juristo et al. [45] referred to these recommendations as usability mechanisms. Usability mechanisms represent particular functionalities that should, according to the HCI community, be implemented within a software system to increase its usability. There are many empirical studies on usability evaluation related to recommendations that affect graphical interface issues [19, 39, 40, 51]. However, there are very few empirical studies that address usability recommendations that affect software design and measure their benefits from the viewpoint of users [4]. Our experiment focuses on this type of HCI recommendations that have an impact on software design.

Based on the ISO/IEC 25010 definition [34] and usability recommendations with a high impact on design [12, 30, 47, 55, 64, 73, 76, 80], this research reports an experiment to evaluate the impact of three usability mechanisms —Abort Operation (ABR), Progress Feedback (PFB) and Preferences (PRF)— on the usability of a web application. ABR is a functionality helping users to safely quit or cancel an action before it is executed [76], PFB provides users with feedback about what the system or interface is doing and how long it is likely to take [55, 76, 81], and PRF is a functionality for adapting the appearance of the graphical interface to user preferences [81]. We chose these three mechanisms because they had a greater impact on software design than other mechanisms like Structured Text Entry or Alert. Another characteristic of the three selected mechanisms is that users can easily recognize their user interface components. This should facilitate their evaluation against HCI recommendations [61].

The functionality of each usability mechanism depends on a combination of use cases identified in previous related works [61, 62] and called application scenarios. For ABR, for example, we selected the FormCancelOpButtonUnsavedChanges, that is, a cancel button on the form that checks for changes and whether or not the changes should be saved. For PFB, we selected the Plw/oInfow/oCancelw/MSG scenario, that is, a noncancellable progress indicator that provides a completion message but does not provide progress information. For PRF, we selected font-size and font-family-based interface customization. Rodríguez et al. [61] also identified the common elements of any scenario implementation. They extracted the common elements as reusable patterns. Although there are potentially several different ways of implementing a usability mechanism, we are quite confident that they will not, in essence, differ much from the proposals in [61], since this was the very point of the reported research. Therefore, we believe that the results of our experiment are not closely dependent on the specific implementation used to program the scenarios used in the experiment.

Our research question is: Does the adoption of usability mechanisms have an impact on application usability? To answer this question, we conducted an empirical study with 168 subjects divided into 24 experimental groups. Each group had to perform one task per mechanism. The aim was to check whether the adoption/non-adoption of each mechanism improves application usability [34, 38].

The three usability variables evaluated in the experiment are efficiency, effectiveness and user satisfaction. Efficiency was measured using two metrics: speed, in terms of time taken by the user to complete a task, and interactiveness, in terms of number of clicks. Effectiveness was measured by percentage task completion, and a questionnaire was used to assess satisfaction. Each experimental subject was placed in a single group and sequentially performed randomly assigned tasks to interact with all three (adopted or non-adopted) usability mechanisms. Accordingly, a single value is output for each experimental subject with respect to each response variable for use in the statistical analysis.

The three usability mechanisms require the inclusion of additional components [44]. The inclusion of additional components leads to increased development costs and time for implementing each mechanism. Rodríguez et al. [61, 62] report that some mechanisms are more or less expensive to implement than others, leading to different impacts in terms of development time and cost. For example, ABR and PFB are more costly to implement, while ABR is more expensive than PFB. PRF has a low implementation cost. If there are large differences in the implementation cost of each mechanism, the experimental evaluation of the real benefit of the adoption of the usability recommendations for a software product can provide valuable information for prioritizing and deciding which mechanisms a system should include.

The major contribution of this paper is that it provides evidence of how the three HCI recommendations improve usability, as a starting point for understanding the impact of implementing usability recommendations that affect software design. As a result of the study, we arrived at several findings.

Key findings:

- The results suggest that the adoption of the three mechanisms improves system usability and does not undermine user performance.
- Implementation cost (studied in [61]) is not related to the resulting usability improvement:
  - Unlike Progress Feedback, Abort Operation and Preferences have a high impact on user satisfaction in the context of this experiment, while Progress Feedback is much more expensive to implement.
  - Abort Operation and Preferences improve user effectiveness, Abort Operation more than Preferences. If these two mechanisms are not enabled, users have to make a bigger effort to complete the tasks. We found that the non-adoption of these two mechanisms sometimes even prevents users from completing the tasks altogether. On the other hand, we did not find any evidence of improved effectiveness for Progress Feedback, despite it being the most expensive mechanism to implement.

- None of the three mechanisms improves user efficiency in terms of interactiveness. Unlike Progress Feedback and Preferences, however, Abort Operation does improve efficiency in terms of speed. Only Abort Operation is worth implementing from the speed point of view.

The remainder of the article is structured as follows. Section 2 describes the papers related to this research. Section 3 reports the experimental setting. Section 4 analyses the data, and discusses and interprets the results. Section 4.3, in particular, addresses our experience with remote experiment. Section 5 addresses the validity of the experiment. Finally, the conclusions and future lines are described in Section 6.

2 Related Work

HCI recommendations aim to achieve satisfactory levels of software systems usability [61]. Many of these recommendations are related to interface design [10, 12, 47] and/or user-system interaction [6, 76, 80]. However, Juristo et al. [44] established that some HCI recommendations affect system functionality beyond its interface. This means that they should be dealt with as functional requirements. There are numerous empirical studies related to software usability: design heuristics [55], usability rules [73], usability principles [16], ergonomics principles [68] and usability patterns [70, 81]. However, there is not much evidence about the final impact of HCI recommendations on software products, particularly, the impact of HCI recommendations that affect software design [4].

Usability-related empirical and experimental studies employ different approaches. As usability evaluation is important for assuring software system quality, it is one of the key issues addressed by experimentation. Some authors aim to empirically validate new methodologies to round out the standard usability evaluation techniques [13]. Other studies validate the use of remote tools to evaluate usability [3]. Still others address the use of tools to evaluate the usability of particular applications like mobile applications [33], specific domain languages [58], or application programming interfaces (APIs) [57]. For example, Sagar et al. compared the results of tool-based and end-user usability evaluation of the top fifty academic websites [65].

There are studies on usability problems regarding specific issues like security [5, 79] or papers addressing general system usability properties or characteristics, for example, comprehension and learnability [77]. Apart from usability methods, some studies propose optimization algorithms [26, 29]. Based on the hierarchical usability model, these algorithms can be used to select a small number of key usability attributes for a particular case and determine application usability without degrading system performance.

Besides the papers on final product usability evaluation, there are also many studies on how to adopt and assure usability during the systems development process. These studies report experiments on usability inspection, including, for example, an experiment comparing design process inspection techniques [32], an experiment exploring how novice evaluators' perceptions of usability heuristics affect their later use [56], a study of the adoption of usability evaluation activities using agile development techniques [66] or an experiment on interface design pattern applicability factors [60]. Other studies focus on web sites and applications and evaluate different user interface issues like navigation design [14, 18] or interface element structures [40].

There are very few empirical studies addressing HCI recommendations that affect software design, and even fewer papers validating the impact of HCI recommendations on user-perceived usability. We have found only one paper on this issue. Aveledo et al. [4] report a study to measure the benefits of usability, exploring the impact of some usability recommendations on software systems developed for a particular context. The experimental design focuses on a toy web application evaluating the attributes of efficiency, effectiveness and user satisfaction. This research addresses the usability evaluation of four mechanisms: Global Undo, Progress Feedback, Structured Text Entry and Go Back. It employs a case study rather than an experiment (the method used in our research).

With regard to HCI recommendations with an impact on software design, there are studies [61, 62] proposing reusable artefacts to implement three usability mechanisms (ABR, PFB and PRF). They use case studies to evaluate the solution. The empirical evidence suggests that the implementation of each mechanism has different costs. The ABR mechanism is found to affect a high percentage of use cases, that is, all or part of the ABR functionality is included in a high number of use cases, whereas the PFB and PRF mechanisms affect a small percentage of use cases. They found that the number of system classes increases most when the PFB mechanism is implemented, whereas there is a moderate increase for ABR, and the number of additional classes for the PRF mechanism is negligible. They also found that each mechanism couples differently with each application functionality: ABR and PFB have a high coupling level, whereas PRF can be regarded as an additional independent requirement. And, finally, the PFB mechanism is harder to program because multithreading is required. Therefore, the ABR mechanism can be said to be more costly at requirements analysis level, whereas the PFB mechanism appears to be the most costly mechanism at design and programming level. Finally, PFR is the least costly in both cases.

3 Experimental Setting

To study the effect of the three selected usability mechanisms, we conducted an experiment following Juristo and Moreno's recommendations [43]. Based on the statement of the research question to be addressed together with the research hypotheses, we applied a preliminary experimental design that evolved into the final design.

Before running the experiment, we conducted a pilot experiment. The pilot experiment was designed to check whether users manage to find the differences between one or more usability mechanisms within the application, evaluate the understanding of the set tasks and conduct a preliminary analysis of the collected data in order to validate the definition of the hypotheses and improve the design, if necessary. The pilot experiment details are outside the scope of this paper. In the following, we discuss the experimental design and the process enacted based on Jedlitschka and Pfahl's experiment reporting guidelines [41].

3.1 Goal, Research Questions and Hypotheses

The aim of our research was to [7]:

- **Experiment** with the mechanisms (ABR, PFB and PRF) in a web application **in order to** evaluate the impact of usability **with respect to** efficiency, effectiveness and satisfaction **from the viewpoint of** users **in the context of** users not trained in computer science.

To achieve our goal, our experiment aims to answer the following research question:

RQ: Does the adoption of usability mechanisms have an impact on application usability?

This research question is further divided into the following three specific research questions:

- RQ1: Does the adoption of the ABR usability mechanism have an impact on application usability? We aim to check whether or not the adoption of Abort Operation improves application usability in terms of efficiency, effectiveness and user satisfaction. The null hypothesis governing this research question is *H.1.x.0: There is no significant difference in user EFFICIENCY | EFFECTIVENESS | SATISFACTION with or without the adoption of ABR*. This hypothesis is broken down into three specific null hypotheses, one for each quality characteristic (where x represents 1. Efficiency, 2. Effectiveness and 3. Satisfaction). They are:
  - H.1.1.0: There is no difference in EFICIENCY with or without the adoption of ABR.
  - H.1.2.0: There is no difference in EFFECTIVENESS with or without the adoption of ABR.
  - H.1.3.0: There is no difference in SATISFACTION with or without the adoption of ABR.

There are another two research questions associated with the other two usability mechanisms:

- RQ2: Does the adoption of the PFB usability mechanism have an impact on application usability?
- RQ3: Does the adoption of the PRF usability mechanism have an impact on application usability?

Like RQ1, these two research questions each break down into three null hypotheses, one for each response variable.

3.2 Factors and Response Variables

The **factor** used in our study is the usability mechanism with two **levels:** adopted and not adopted. We aim to compare the effect of the adoption or non-adoption of a particular usability mechanism during task performance. We run three experiments with the same participants (i.e. sample) to simultaneously study all three factors or usability mechanisms —ABR, PFB and PRF—. Nonetheless, we evaluate the mechanism effect separately, as the interaction between mechanisms is irrelevant at this early stage of the research.

The **response variables** are efficiency, effectiveness and satisfaction. ISO/IEC 25010 [34] defines measures of quality in use; and efficiency, effectiveness and satisfaction are common attributes for evaluating product usability. An increase in these three quality characteristics is a measure of impact on usability, which can improve or degrade application usability. The standard defines each of the attributes as follows:

- Effectiveness: degree to which users correctly and completely achieve specified goals.
- Efficiency: resources expended by users to correctly and completely achieve specified goals.
- Satisfaction: degree to which user needs are satisfied by using a product or system in a specified context of use.

The response variables efficiency, effectiveness and satisfaction and their respective **metrics** are outlined below:

- Effectiveness: percentage task completion by a subject [35]. Effectiveness can be represented by:

$$Effectiveness = \frac{Number\ of\ successfully\ completed\ subtasks}{Total\ number\ of\ subtasks\ undertaken} * 100\%$$

- Efficiency is measured according to two metrics:
  a) Speed: time measured in seconds taken by a subject to complete the task [35, 37]. The clocked time represents the time taken by the subject to perform the task and, if necessary, to reread the instructions during task performance. Efficiency measured as user speed can be represented by:

$$Ef_{speed} = \frac{StopTime_{milliseconds} - StartTime_{milliseconds}}{1000}$$

  b) Interactiveness: number of clicks made by a subject to complete the task [31, 71]. We count separate clicks, where a double click is classed as two separate clicks. Efficiency measured as user interactiveness can be represented by:

$$Ef_{interactiveness} = count(separateClicks).$$

- Satisfaction: mean value of the responses to the post-task questionnaire questions. The values of the questionnaire responses are ordinal on a Likert scale (1 = Strongly disagree to 5 = Strongly agree) [67]. There are two satisfaction questions per mechanism. Satisfaction can be represented by:

$$s = \frac{questionValue_1 + questionValue_2}{2}$$

3.3 Context and Experimental Subjects

The experiment was conducted in two contexts:

a) Academic setting: undergraduate students from the Universidad Autónoma de Asunción (Paraguay) taking different degree programmes, including economic and business science, legal science, health science, human and communication science. In this context, the challenge was to carry out the experiment on a university distance learning platform (*e-campus*), where each professor gave additional points to any student who volunteered to participate.
b) Non-academic setting: practitioners and non-practitioners who were sent an invitation to participate by messaging applications or electronic mail.

The experiment was executed in each context at different non-overlapping time periods. The experimental subjects were not computer science specialists. This was an experimental design constraint introduced on two grounds: a) users do not need to be knowledgeable about computer science to use the system and appreciate usability, and b) computer specialists could alter the results of the experiment because they are familiar with the controls implementing some of the usability mechanisms in the experiment scenarios.

All the subjects had to perform the tasks set as part of the experiment. Participation was voluntary and, in the case of students, required the consent of the institutional authorities. Finally, all the subjects completed the familiarity questionnaire in Appendix A (Table 10).

The sample-related details are as follows:

- The final sample contained 168 subjects: 88 from the academic setting and 80 from the non-academic setting. Of the sample, 76 were men and 92, women.
- The biggest concentration of participants spanned two age groups: 18-30 years (61%) and 31-40 years (26%).
- The experimental subjects were regular Internet users that use the Web both at home (49%) and at work (46%). A smaller group uses the Internet in other environments (4-5%).
- With respect to subjects' online shopping habits, most participants had never purchased anything over the Internet (61%), whereas 13% had seldom placed orders, 17% had occasionally shopped online and 9% were regular (always or almost always) e-shoppers. It is an advantage to have users that are unfamiliar with the application domain (online shopping) in our experiment because are more sensitive to system usability.

3.4 Experimental Design

We use a between-subjects design with orthogonal array [15, 75]. Each experimental subject is assigned to a single group that accommodates a combination of factor levels (with or without usability mechanism adoption). In our experiment, each subject interacts with three adopted or non-adopted usability mechanisms by sequentially performing randomly assigned tasks. Accordingly, each experimental subject provides a single value for each response variable (efficiency, effectiveness, satisfaction), which will be used in the statistical analysis.

Our design is composed of a treatment matrix, a mechanism exposure order matrix and a group assignment matrix. Table 1.a shows the treatment matrix describing which usability mechanisms will and will not be adopted. The zeros denote a non-adopted usability mechanism, whereas the ones stand for the adopted mechanism. For example, when a subject is assigned treatment A, he or she will have to perform the ABR and PFB tasks without access to the usability mechanism and the PRF task with the enabled usability mechanism. Table 1.b shows the order of exposure for each factor. This matrix establishes all the possible task performance sequences for each factor (without repetitions). Therefore, if a subject is assigned to order O1, he or she will follow a sequence of actions starting with the ABR task, followed by the PFB task and ending with the PRF task. The exposure order matrix aims to cancel out the possibility of one mechanism having a learning effect on another.

Table 1: Between-subjects design

a) Treatment matrix (orthogonal)

Treatment	ABR	PFB	PRF
A	0	0	1
B	0	1	0
C	1	0	0
D	1	1	1

b) Mechanism order exposure matrix

Order	Task 1	Task 2	Task 3
O1	ABR	PFB	PRF
O2	ABR	PRF	PFB
O3	PFB	PRF	ABR
O4	PFB	ABR	PRF
O5	PRF	ABR	PFB
O6	PRF	PFB	ABR

Table 2: Group assignment matrix

			Task 1			Task 2			Task 3		
Treatment	**Order**	**Group**	**ABR**	**PFB**	**PRF**	**ABR**	**PFB**	**PRF**	**ABR**	**PFB**	**PRF**
A	**O1**	**GROUP1**	0				0				1
	O2	**GROUP2**	0					1		0	
	O3	**GROUP3**		0				1	0		
	O4	**GROUP4**		0		0					1
	O5	**GROUP5**			1	0				0	
	O6	**GROUP6**			1		0		0		
B	**O1**	**GROUP7**	0				1				0
	O2	**GROUP8**	0					0		1	
	O3	**GROUP9**		1				0	0		
	O4	**GROUP10**		1		0					0
	O5	**GROUP11**			0	0				1	
	O6	**GROUP12**			0		1		0		
C	**O1**	**GROUP13**	1				0				0
	O2	**GROUP14**	1					0		0	
	O3	**GROUP15**		0				0	1		
	O4	**GROUP16**		0		1					0
	O5	**GROUP17**			0	1				0	
	O6	**GROUP18**			0		0		1		
D	**O1**	**GROUP19**	1				1				1
	O2	**GROUP20**	1					1		1	
	O3	**GROUP21**		1				1	1		
	O4	**GROUP22**		1		1					1
	O5	**GROUP23**			1	1				1	
	O6	**GROUP24**			1		1		1		

Each row of the treatment matrix is combined with each row of the exposure order matrix to produce 24 groups (Table 2). A subject assigned to, for example, GROUP11 will perform the PRF task without the mechanism, followed by the ABR task without the mechanism and, finally, the PFB task with the mechanism. Note that the treatment matrix is useful for the data analysis process (contrasts), whereas the exposure order matrix aims to moderate any potential learning effect of one factor on another. The group

assignment matrix, generated based on the treatment and the exposure order matrices, is used to randomly assign each subject to a group during experiment execution.

3.5 Instrumentation and Tasks

The experiment participants will use a web application system. This application must have highly interactive features, that is, the software system should be dependent on the actions taken by the user to perform a task. In this respect, we developed an online shop called *QuickStore* [83], adopting the three usability mechanisms: ABR, PFB and PRF. We also developed the experiment administration interface [22], which provides the subject with guidance throughout the whole experiment in the form of instructions on the activities and actions he or she is to perform. Additionally, the administration interface formats and exports the collected empirical data (times, clicks, questionnaires, etc.) for analysing the impact on the efficiency, effectiveness and satisfaction attributes.

Each subject performs three tasks. The tasks are:

- Abort Operation: the subject will apply a cancel operation to his or her shopping cart. Upon login, the user's shopping cart will already contain several items. The user has to go to his or her shopping cart and modify data (for example, increase the number of any of the items, enter a promotional code, etc.) and then cancel the operation. If the usability mechanism is enabled, the user will have a quick cancel option and will merely have to confirm the cancellation of all the pending changes. If the usability mechanism is not enabled, the user will have to manually undo each change made since the start of the task one by one. Appendix C (Table 11) shows the design of the proposed task for ABR.
- Progress Feedback: the subject has to search for a specified item and add this item to the shopping cart. The subject starts the task from the *QuickStore* application home page, running a search using his or her preferred criteria, for example, item name. If the search is successful, he or she merely has to press the Add to Shopping Cart button. If the usability mechanism is enabled, a progress bar will be displayed while the search is running telling the user that the action is being executed and a message will be displayed at the end of the search specifying the number of items found. If the usability mechanism is not enabled, the user will not be informed during the search that the action is ongoing. Appendix C (Table 12) shows the design of the proposed task for PFB.
- Preferences: this task is divided into two parts. The user will perform first the basic task and then the fictitious task. Appendix C (Table 13) shows the design of the proposed task for PRF.
    a) Basic Preferences Task: the subject should customize the application user interface. The font size of the original interface is small and not very legible. If the usability mechanism is enabled, the user can customize some shop features to his or her liking. On the other hand, if the mechanism is not enabled, the user cannot modify the application interface appearance.
    b) Fictitious Preferences Task: the user is asked to search for information on the time limit for returns of purchased items provided by the application. If the subject has modified the system interface, he or she can easily find the link to the required information. However, if the user was not able or decided not to modify the application interface appearance, it will be very hard for him or her to find the required information.

The above three primary tasks make up the experiment execution scenario. The subject is assumed to be a customer, that is, he or she interacts with the final system interface performing non-administrative tasks. The order of execution of each task is random and determined by the group assignment matrix (Table 2). In the particular case of PRF, metrics are collected exclusively for the fictitious task, as we want to evaluate the effect of the mechanism on the specified task performance in the shop.

Finally, the subjects must complete the familiarity questionnaire outlined in Appendix A (Table 10) before performing the tasks. Additionally, each subject must fill in a post-task satisfaction questionnaire for the mechanism used to complete each task (Appendix C).

3.6 Operation

At the time of experiment execution, the subjects were not familiar with the aim of the study or with the research hypotheses. We referred to the experiment as an *evaluation* instead of *experiment* because we did not want the subjects to feel that they were being used as guinea pigs. We also informed the subjects that the results of this evaluation would be used to improve web application issues and we guaranteed their confidentiality.

The experiment was conducted over a five-month period from March to July 2016. Over the first four months, the experiment was executed within academia at the Universidad Autónoma de Asunción using the distance education platform (*e-campus*). Each professor published the experiment link on his or her course *e-campus* (see the screenshot in Appendix B). Over the last month (July), the experiment was conducted outside academia. Subjects were informed that participation was voluntary. The students that agreed to participate were encouraged to make their best effort to perform the tasks, although it was an optional challenge that had no bearing on their outcomes.

Apart from the link [83] that each participant was to use to log in and start the evaluation, we did not provide any additional material. Neither did we organize any preparatory or practice sessions with the subjects. The experiment was all executed remotely, that is, the subjects performed the experiment in their natural —home, office, university— surroundings. The experimental procedure was as follows. Subjects log on to the web application [83] link from a desktop computer Internet browser (mobile

devices were not allowed) and complete the familiarity questionnaire to start experiment configuration. The system randomly assigns the subject to one of 24 same-size groups to rule out any learning effects. The group number determines the task performance order and which usability mechanisms are or are not adopted (Table 2). At this point, a web page is displayed giving participants basic instructions on how to perform the tasks and consistently answer survey questions. After reading and confirming that they have understood the instructions, participants start to evaluate the *QuickStore* application. To do this, tasks are displayed one at a time by wizard navigation. Each task includes a task description, satisfaction questionnaire and the online store graphical interface equipped with the mechanism functionalities. A satisfaction questionnaire must be completed before moving on to the next task. During task performance, data on the number of clicks, time taken and percentage task completion were collected. No time limit was set. After completing all the experimental tasks, participants filled in an evaluation satisfaction questionnaire.

In principle, data were collected for a total of 182 subjects. However, the data for 14 subjects were removed because they did not correctly complete the tasks. In particular:

- two subjects did not answer the questionnaire questions.
- three subjects repeated the experiment.
- three subjects had a computer science background.
- six subjects did not finish the whole scenario, that is, they started the first and even the second but did not finish all the tasks. Remember that an experiment execution scenario is the combination of the above three tasks, and subjects are required to perform the tasks in their entirety.

Finally, 168 valid data remained for the statistical analysis and results interpretation.

4 Analysis Approach

We divided the analysis into three different parts (one per usability mechanism: ABR, PFB and PRF). In each part we assessed the effect of the usability mechanism on all response variables (clicks, time, percentage task completion and satisfaction).

Within each part (i.e., for each mechanism), we followed an identical procedure:

- Use violin and box plots to illustrate the score distributions for each response variable. Two groups were compared: the group in which the usability mechanism was adopted and the group in which the mechanism was not adopted. Summary statistics (mean, median and standard deviation) were provided in order to round out the violin and box plots.
- Assess the statistical significance (i.e., p-value) of the findings.
- Assess the effect size of the findings.

The response variables in this experiment were measured using count (click and time), proportion (percentage task completion) and ordinal (satisfaction) data. We used the Mann-Whitney U test [50] in order to increase data analysis consistency. The Mann-Whitney U test is a scale-free statistical test and can assess the statistical significance of all response variables. As a result, we did not have to use different statistical tests to analyse each response variable (e.g., count models [48] for click and time, beta regressions [20] for percentage task completion and ordinal logistic regressions [9] for satisfaction). This makes it easier to understand the overall results.

The Mann-Whitney U test is regarded as the non-parametric counterpart of the independent t-test [52]. This test uses ranks, instead of natural units, to make sample-to-population inferences. Contrary to common belief, if the two samples that are being compared are not similarly shaped, the Mann-Whitney U test is not just a test of medians, it is also a test of standard deviations [54]. In other words, the shape of each distribution matters in the overall result. Thus, if the result of the Mann-Whitney U test is found to be significant, the distributions are not identically shaped.

In order to round out the results provided by the Mann-Whitney U test, we computed Cliff's delta effect size [49]. Cliff's delta effect size was selected as it is unit-free (can be used for scores with unknown distribution shapes), compares the overall scores in terms of rank (like the Mann-Whitney U test) and is straightforward to interpret since it quantifies the superiority (or inferiority) of one group over another. Cliff's delta ranges from 1 to -1 (absolute superiority and inferiority of one group over another, respectively). A Cliff's delta close to 0 indicates similar distributions (i.e., exactly the same number of scores are superior in one group to another group). We interpret each Cliff's delta effect size magnitude (small, medium, large) following the conventions shown in [46]. In addition to Cliff's delta effect size, we use the probability of superiority effect size [24] because it is straightforward to interpret: the probability of one group being larger than another. Probabilities of superiority close to 0 indicate that both groups are equally likely to achieve a larger score for the outcome under assessment. The probability of superiority and Cliff's delta effect size should make the findings more interpretable and provide consistent units for comparing the benefits and drawbacks of all the usability mechanisms assessed during the experiment.

4.1 Analysis

Sections 4.1.1, 4.1.2 and 4.1.3 report the analysis of the ABR, PFB and PRF usability mechanisms, respectively.

4.1.1 RQ1: Abort Operation (ABR)

As Figure 1 shows, the spread of the click scores and time taken distributions appear to be greater for the non-adopted than the adopted mechanism. The task completion percentage appears to be noticeably larger for the adopted than the non-adopted

mechanism, whereas the mean satisfaction (marked with a + within the plot) appears to be greater for the adopted than the non-adopted mechanism.

Figure 1. Violin-plots for the number of clicks, elapsed time, percentage task completion and satisfaction with the ABR usability mechanism.

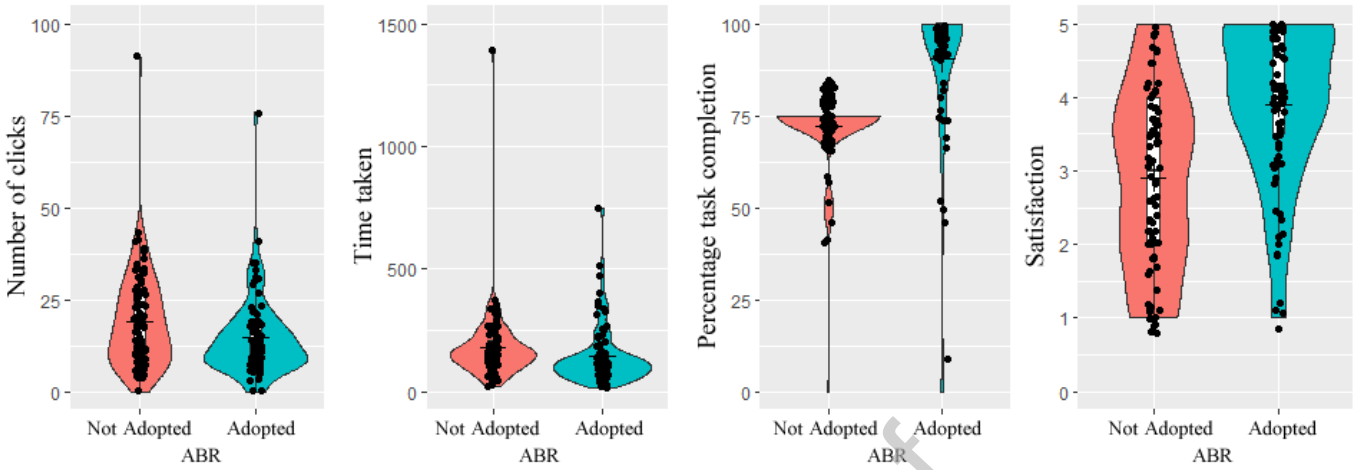


Table 3 shows the summary statistics for each response variable distribution (depending on whether the ABR usability mechanism is or is not adopted).

Table 3. Summary statistics (by response variable and group)

Response Variable	Group	Mean	Median	SD
Click	Not adopted	18.93	17	13.21
	Adopted	14.52	12	10.92
Time	Not adopted	179.68	157.22	154.69
	Adopted	142.67	110.23	120.12
Percentage	Not adopted	72.02	75	10.56
	Adopted	90.47	100	23.57
Satisfaction	Not adopted	2.90	3	1.23
	Adopted	3.90	4	1.10

Table 4 shows the statistical significance (Mann-Whitney U test p-value), effect size (Cliff's delta effect), statistical significance assessment, effect size assessment and probability of superiority of one group over another for each response variable for the Abort Operation mechanism.

Table 4. Effect size and statistical significance assessment

Response Variable	p-value	p-value Assessment	Effect Size (Cliff's delta)	Effect Size Assessment	Pr(Superiority)	
					(Adopted>Not adopted)	(Adopted>Not adopted)
Click	0.015	Significant	0.21	Small	0.377	0.594
Time	<0.001	Significant	0.31	Medium	0.344	0.656
Percentage	<0.001	Significant	-0.73	Large	0.809	0.079
Satisfaction	<0.001	Significant	-0.46	Large	0.674	0.213

In summary:

- There is a 37.7% probability of subjects with access to the Abort Operation clicking more often than subjects who do not have access to this usability mechanism. In other words, there is a 59.4% probability of subjects with access to the Abort Operation mechanism clicking less often than subjects who do not have access to this mechanism. The difference is statistically significant, and the effect size is small.

- There is a 34.4% probability of subjects with access to the Abort Operation taking longer than subjects who do not have access to this usability mechanism. In other words, there is a 65.6% probability of subjects with access to the Abort Operation taking less time than subjects who do not have access to this mechanism. The difference is statistically significant, and the effect size is medium.
- There is an 80.9% probability of subjects with access to the Abort Operation being more efficient than subjects who do not have access to this usability mechanism. In other words, there is a 7.9% probability of subjects with access to the Abort Operation being less efficient than subjects who do not have access to this mechanism. The difference is statistically significant, and the effect size is large.
- There is a 67.4% probability of subjects with access to the Abort Operation being more satisfied than subjects who do not have access to this usability mechanism. In other words, there is a 21.3% probability of subjects with access to the Abort Operation being less satisfied than subjects who do not have access to this usability mechanism. The difference is statistically significant, and the effect size is large.

4.1.2 RQ2: Progress Feedback (PFB)

As Figure 2 shows, the number of clicks and time taken distributions appear to be similarly shaped across both groups, although the standard deviation appears to be larger when the Progress Feedback mechanism is adopted. Also both the percentage task completion and satisfaction distributions appear to be similarly shaped across both groups, although the satisfaction scores look to be more variable when the Progress Feedback mechanism is not adopted.

Figure 2. Violin plots for the PFB usability mechanism with respect to number of clicks, time taken, and percentage task completion.

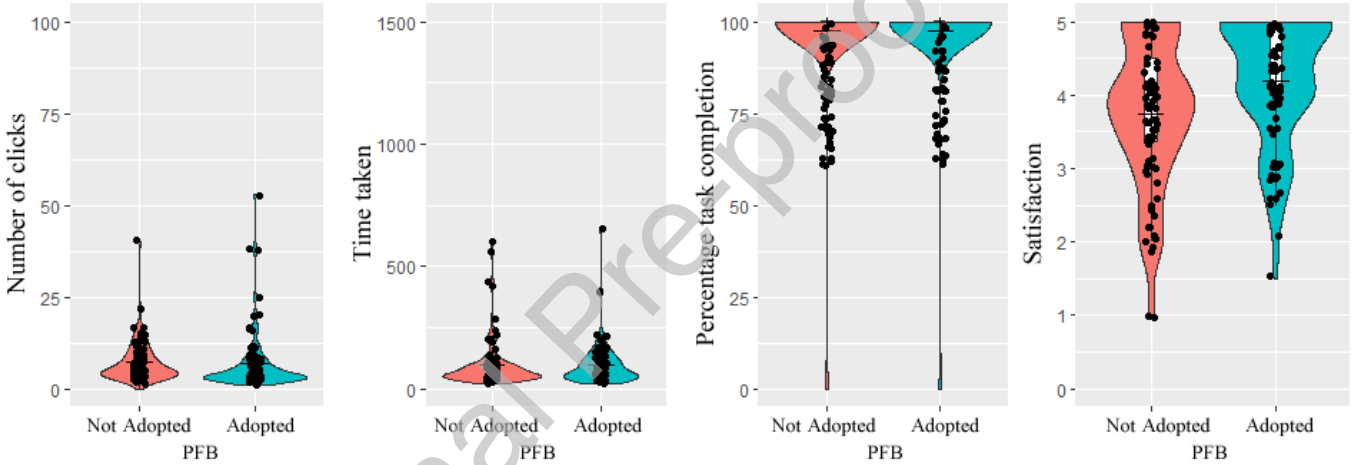


Table 5 shows summary statistics for each response variable distribution (depending on whether or not the Progress Feedback usability mechanism is adopted).

Table 5. Summary statistics (by response variable and group).

Response Variable	Group	Mean	Median	Sd
Click	Not adopted	7.10	5	5.56
	Adopted	6.93	3.5	8.45
Time	Not adopted	99	65.21	105.33
	Adopted	95.86	72.21	85.51
Percentage	Not adopted	97.61	100	15.33
	Adopted	97.61	100	15.337
Satisfaction	Not adopted	3.75	4	0.98
	Adopted	4.19	4.5	0.85

Table 6 shows the statistical significance (Mann-Whitney U test p-value), effect size (Cliff's delta effect), statistical significance assessment, effect size assessment and probability of superiority of one group over another with respect to each response variable for the Progress Feedback mechanism.

Table 6. Effect size and statistical significance assessment.

Response	p-value	p-value	Effect size	Effect size	Pr(Superiority)

Variable		assessment	(Cliff's delta)	assessment		
					(Adopted>Not adopted)	(Adopted>Not adopted)
Click	0.011	Significant	0.223	Small	0.338	0.562
Time	0.518	Not significant	-0.06	Small	0.528	0.471
Percentage	1	Not significant	<-0.001	Small	0.023	0.023
Satisfaction	0.001	Significant	-0.274	Small	0.548	0.273

In summary:

- There is a 33.8% probability of subjects with access to the Progress Feedback mechanism clicking more often than subjects who do not have access to this usability mechanism. In other words, there is a 56.2% probability of subjects with access to the Progress Feedback mechanism clicking less often than subjects who do not have access to this mechanism. The difference is statistically significant, and the effect size is small.
- There is a 52.8% probability of subjects with access to the Progress Feedback mechanism taking longer than subjects who do not have access to this usability mechanism. In other words, there is a 47.1% probability of subjects with access to the Progress Feedback mechanism taking less time than subjects who do not have access to this mechanism. The difference is not statistically significant, and the effect size is small.
- There is a 2.3% probability of subjects with access to the Progress Feedback mechanism being more efficient than subjects who do not have access to this usability mechanism. In other words, subjects are equally likely to finish the same task with or without the usability mechanism in place. The difference is not statistically significant, and the effect size is small.
- There is a 54.8% probability of subjects with access to the Progress Feedback mechanism being more satisfied than subjects who do not have access to this usability mechanism. In other words, there is a 27.3% probability of subjects with access to the Progress Feedback mechanism being less satisfied than subjects who do not have access to this mechanism. The difference is statistically significant, and the effect size is small.

4.1.3 RQ3: Preferences (PRF)

As Figure 3 shows, there appears to be a larger mean number and a wider spread of clicks for the non-adopted than the adopted Preferences mechanism. Less time appears to be taken when the Preferences mechanism is adopted, and the spread appears to be narrower when it is not adopted. Whenever the mechanism is adopted, the mean percentage task completion is noticeably greater than when it is not adopted, and the same applies to satisfaction.

Figure 3. Violin plots for the PRF usability mechanism with respect to number of clicks, time taken, percentage task completion and satisfaction.

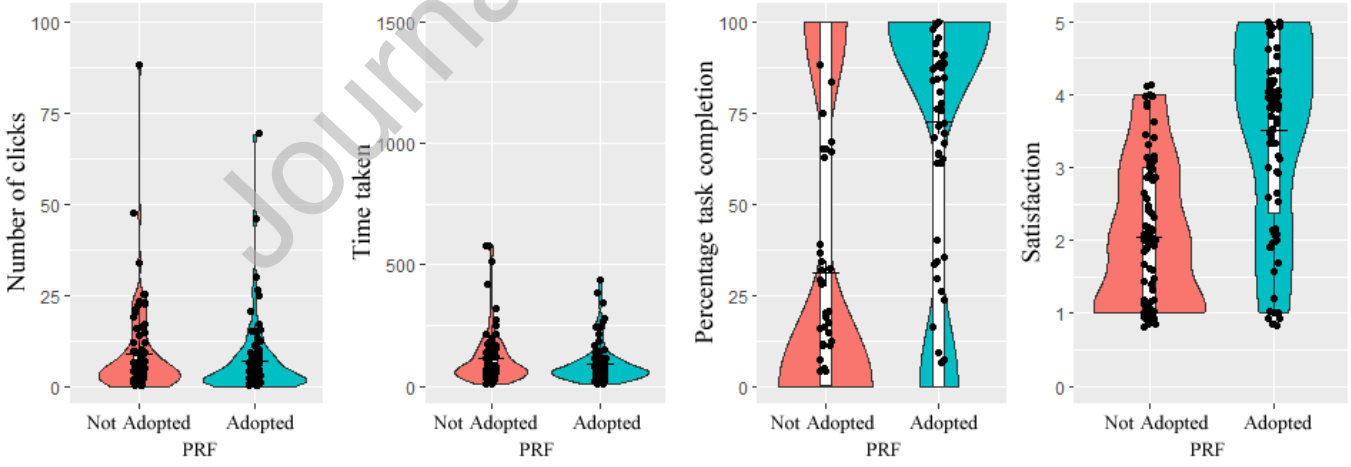


Table 7 shows the summary statistics of the distribution of each response variable (depending on whether or not the Preferences usability mechanism is adopted).

Table 7. Summary statistics (by response variable by group).

Response Variable	Group	Mean	Median	SD
Click	Not adopted	9.82	5	15.94
	Adopted	6.75	4	10.25
Time	Not adopted	113.59	76.24	112.25
	Adopted	90.20	66.89	80.92
Percentage	Not adopted	30.95	0	46.51
	Adopted	72.62	100	44.86
Satisfaction	Not adopted	2.04	2	0.99
	Adopted	3.50	4	1.35

Table 8 shows the statistical significance (Mann-Whitney U test p-value), effect size (Cliff's delta effect), statistical significance assessment, effect size assessment and probability of superiority of one group over another with respect to each response variable for the Preferences mechanism.

Table 8. Effect size and statistical significance assessment.

Response Variable	p-value	p-value assessment	Effect size (Cliff's delta)	Effect size assessment	Pr(Superiority)	
					(Adopted>Not adopted)	(Adopted>Not adopted)
Click	0.124	Not significant	0.137	Small	0.399	0.536
Time	0.141	Not significant	0.131	Small	0.434	0.566
Percentage	<0.001	Significant	-0.416	Medium	0.501	0.085
Satisfaction	<0.001	Significant	-0.593	Large	0.747	0.154

In summary:

- There is a 39.9% probability of subjects with access to the Preferences mechanism clicking more often than subjects who do not have access to this usability mechanism. In other words, there is a 53.6% probability of subjects with access to the Preferences mechanism clicking less often than subjects who do not have access to this usability mechanism. The difference is not statistically significant, and the effect size is small.
- There is a 43.4% probability of subjects with access to the Preferences mechanism taking longer than subjects who do not have access to this usability mechanism. In other words, there is a 56.6% probability of subjects with access to the Preferences mechanism taking less time than subjects who do not have access to this usability mechanism. The difference is not statistically significant, and the effect size is small.
- There is a 50.1% probability of subjects with access to the Preferences mechanism being more efficient than subjects who do not have access to this usability mechanism. In other words, there is an 8.5% probability of subjects with access to the Preferences mechanism being less efficient than subjects who do not have access to this usability mechanism. The difference is statistically significant, and the effect size is medium.
- There is a 74.7% probability of subjects with access to the Preferences mechanism being more satisfied than subjects who do not have access to this usability mechanism. In other words, there is a 15.4% probability of subjects with access to the Preferences mechanism being less satisfied than subjects who do not have access to this usability mechanism. The difference is statistically significant, and the effect size is large.

4.2 Results

In this section, we discuss the quantitative results in response to the research questions. The entire analysis was conducted using the R language [59]. We interpreted Cliff's delta according to Kitchenham's cut-off points [46]. Table 9 summarizes the experiment results. The effect size is specified between square brackets. The * symbol denotes that the difference is statistically significant.

Table 9: Summary of the experiment results

	Efficiency	Effectiveness	Satisfaction

Usability Mechanism	Clicks	Time	Percentage	s
ABR	[small] *	[medium] *	[large] *	[large] *
PFB	[small] *	[small]	[small]	[small] *
PRF	[small]	[small]	[medium] *	[large] *

4.2.1 RQ1: Abort Operation (ABR)

Efficiency. Considering efficiency measured in terms of interactiveness (clicks), there is a statistically significant small difference in the number of clicks between subjects working with and without the adopted mechanism. The number of clicks is greater if no cancel option has been enabled indicative of a bigger user effort to undo actions. This additional user effort degrades efficiency measured as interactiveness when the mechanism is not adopted. Note, however, that the difference in the mean number of clicks by subjects working with and without the adopted mechanism is not substantial. We conclude that **the adoption of ABR does not improve user efficiency in terms of interactiveness**.

In terms of speed, there is a difference between the mean time taken by subjects working with and without the adopted mechanisms, that is, subjects tend not to take as long when the mechanism is adopted. Users performing the ABR task are required to cancel the changes that they made to the shopping cart. Users are expected to use the cancel option if it is available and to spend more time trying to find out how to do undo changes if it is not. User speed appears to be better when the mechanism is adopted. According to this mean effect size, subjects working with the adopted mechanism are 25% faster at cancelling actions and safely quitting an unwanted state than subjects without access to the mechanism. Although this would appear to be a negligible difference (~37 s), we believe that the time taken decreases noticeably when the subjects have access to the mechanism (see Table 3). We conclude that the **adoption of ABR improves user efficiency in terms of speed**.

Effectiveness. The results confirm that there is a statistically significant big difference in percentage task completion between subjects with and without access to the mechanism. Many subjects who had access to the mechanism managed to undo their changes and go on to complete all the subtasks before giving up. This means that subjects managed to complete a greater percentage of the task thanks to the adopted mechanism. We conclude that **the adoption of ABR improves user effectiveness**.

Satisfaction. The results show a statistically significant big difference between the satisfaction questionnaire responses of subjects working with and without the adopted mechanism. The subjects felt that the system provided help for cancelling actions. This means that the subjects with access to the mechanism rated system change alerts favourably and had a positive perception of the ease of action cancellation. We conclude that the **adoption of ABR improves user satisfaction**.

Although the Abort Operation mechanism is costly for software developers to implement [61, 62], its usability benefits appear to justify its implementation. If the Abort Operation mechanism is adopted, user-system interaction improves significantly in terms of productivity (speed with which tasks are performed), effectiveness (task completion) and satisfaction (help perception and mechanism comfort).

4.2.2 RQ2: Progress Feedback (PFB)

Efficiency. In terms of interactiveness, subjects with access to the mechanism appear to employ slightly fewer clicks than users that do not have access to the mechanism. If the mechanism is not available, it fails to provide any indication of what the system is doing. This is likely to increase the number of user clicks due to a temporary loss of system control or user disorientation caused by the delay induced by the search task. However, we observe that the difference in the average number of clicks is negligible, and subjects perform similar numbers of clicks irrespective of whether or not the mechanism is adopted. We conclude that **the adoption of PFB does not improve user efficiency in terms of interactiveness**.

With respect to speed, there is also a small difference between the mean times taken by subjects with and without access to the mechanism. PFB adoption does not appear to improve user speed, as mechanism implementation only displays emerging information about the search progress. Ultimately, user execution times differ by a matter of seconds irrespective of whether or not the mechanism is adopted, and the observed improvement is more or less imperceptible. We believe that, as this mechanism informs users about what is happening in the system while they are performing the task, it is reasonable to assume that all user execution times will be similar. We conclude that the **adoption of the PFB does not improve user efficiency in terms of speed.**

Effectiveness. There is a very small difference in the percentage task completion between the subjects working with and without the adopted mechanism that is not statistically significant. In fact, the difference in mean percentage task completion tends to zero, that is, the adoption of the mechanism does not play a vital role in successful task completion. The role of this mechanism is in fact to make slow applications more acceptable by giving users appropriate feedback about their actions and letting them know that the system is doing what it should be doing. Our application response times are relatively fast as the PFB task is a simple search operation. The subjects searched for and found the specified item with or without the adopted mechanism, which was an animated progress indicator providing progress information. We conclude that **the adoption of PFB does not improve user effectiveness**.

Satisfaction. There is a statistically significant small difference in the responses to the user satisfaction questionnaires. When PFB was adopted, subjects positively rated the pop-up window displayed by the system reporting the status of the actions that were running. This means that user satisfaction was slightly better for subjects that had access to the mechanism than for others that did

not. As the difference is very small, the improvements perceived by the users are regarded as negligible. We believe that larger improvements could be observed in contexts involving the performance of high latency tasks. We conclude that the **adoption of PFB does not improve user satisfaction**.

According to implementation effort data, Progress Feedback tends to be costly to implement [61, 62]. On this ground, it should not be adopted in the software system if project resources are limited. We do not think it is worthwhile investing development effort in implementing Progress Feedback in the online shop, because it does not provide appreciable improvements in user performance. As there are pattern-based solutions [61], developers are of the opinion that the cost of implementing this mechanism accounts for only a very small proportion of the global system development effort. Therefore, we do not rule out the implementation of this mechanism, provided that there are sufficient project resources. Additionally, more and more development frameworks provide more flexible software components to facilitate usability adoption in systems.

4.2.3 RQ3: Preferences (PRF)

Efficiency. With this usability mechanism, no significant effects were observed in terms of speed or interaction. In terms of interactiveness, there is a negligible difference in the number of clicks between subjects with and without access to the mechanism that is not statistically significant. The PRF task requires users to customize the application interface and visually search for specified information on screen. It takes at most one click to locate this information. Therefore, it is not possible to observe an improvement in terms of user clicks. We conclude that **the adoption of PRF does not improve user efficiency in terms of interactiveness**.

In terms of speed, there is a small difference in the mean time between subjects with and without access to the mechanism access that is not statistically significant. Subjects appear to be slightly faster working with than without the PRF mechanism. Accordingly, user interface customization using this mechanism does not appear to help improve user performance because the task (browsing the shop visually in search of the returns policy) is too simple. We think that a customizable interface could have different effects on efficiency for more complex tasks. We conclude that **the adoption of PRF does not improve user efficiency in terms of speed**.

Effectiveness. There is a statistically significant sizeable difference in the percentage task completion between the subjects with and without access to the mechanism. On the one hand, many subjects managed to visualize the specified information in the customized interface before the mechanism was adopted. On the other hand, a sizeable number subjects gave up this task because they had difficulty seeing the information specified in the original illegible interface. The mean effect size is a key indicator that subjects managed to advance further with the task when the mechanism was enabled. With the adoption of configurable options, we believe that users perform better as they can customize some system features. We conclude that **the adoption of PRF improves user effectiveness**.

Satisfaction. There is a statistically significant big difference in the satisfaction responses between subjects with and without access to the mechanism. Subjects were not happy with the original illegible interface. However, they thought that the system was easy and agreeable to interact with after they used the PRF mechanism to improve the interface. User satisfaction with the application improved, and they positively rated the adoption of configurable system options. They also claimed that such configurations improved task performance. We conclude that **the adoption of PRF improves user satisfaction**.

Considering that implementation effort data suggest that the Preferences mechanism is not costly to develop [61, 62] and that the adoption of this mechanism significantly improves effectiveness and satisfaction, we believe that it is worthwhile building this mechanism into a software system. Its low implementation cost, together with the potentially positive impact appreciated by users, mean that this is an attractive mechanism for improving software system usability.

4.3 Our Experience with Remote Experiment

On top of the experimental results, we should also discuss remote testing experiences [8, 63]. We decided to use unmoderated remote usability testing [69], where participants complete all the sessions on their own. This should provide more realistic data than laboratory testing. This is a feasible approach for remote usability evaluation, but is perhaps less practicable for experiments evaluating other task types. A number of interesting questions emerged during the execution of the experiment that we had not taken into account initially. We gathered the opinions of the participants using questionnaires. Additionally, we questioned randomly selected subjects and some subjects with outlying behaviour (atypical times and clicks, unexpected responses to satisfaction questions, etc.).

Firstly, no connection problems were reported during the experiment execution. Even so, several subjects failed to complete all the tasks. Some of the reasons given by subjects are: they had to do a lot of reading to perform a task, they were not motivated or interested enough to learn how to do the tasks, they broke off the experiment for work-related reasons or another activity, they accidently quit the web browser during the test, etc. The incomplete subject data (mentioned in Section 3.6) were removed from the results analysis and interpretation.

We also identified outlying behaviour with respect to:

- Task interpretation and performance: We found subjects that misinterpreted the set task. For example, even when the ABR mechanism was enabled, subjects emptied the shopping cart by closing the application window, refreshing the web page or even buying the item instead of undoing the changes. For PFB, some users performed a manual search instead of using

the search engine. For PRF, some subjects failed to complete the task from the original interface (small and illegible font size). They thought this was due to their eyesight.

- Graphical interface usability components: Some subjects did not appear to be acquainted with or notice the implemented PRF mechanism components, even though there was a My Preferences button on the application home page. The ABR mechanism cancel alert prompted other users to quit the page and start the evaluation over again.
- Satisfaction questions: many subjects stated that they were satisfied with system interaction. Even though the mechanisms were not adopted, they felt that the graphical interface was intuitive, easy to use and provided immediate responses to their actions.
- Additional comments: some subjects were averse to completing questionnaires. Others stated that they were not interested in performing the experiment. A few defined the search criteria incorrectly. They, therefore, failed to find the item, irrespective of mechanism adoption. Some subjects managed to perform the PRF task with the original illegible interface merely by using browser options (zoom, copy & paste, etc.).

Only the subjects that correctly completed the entire task were included in the results analysis and interpretation.

5 Threats to Validity

An experiment is valid insofar as the results can be attributed to an independent variable and can be generalized outside the experiment. If the results of an experiment can be definitely attributed to the independent variable, the experiment is said to be internally valid [82]. External validity is related to the generalizability of the results. In other words, internal validity is the extent to which the association relationships can be established between the independent and dependent variables. External validity establishes the conditions under which the results can be generalized outside the context of the experiment [1]. In this section, we discuss the threats with respect to internal and external validity.

5.1 Internal Validity

Internal validity is related to experiment quality. Threats to internal validity are issues that can, unbeknown to the researcher, have an impact on the cause effect of the independent variable. Wohlin et al. [82] describe categories for the factors that have an impact on internal validity, including history, maturation, instrumentation, etc.

This experiment accounts for potential threats to internal validity related to:

1. Knowledge of the technology: although all the participants share the same level of experience with this type of experiments (they are novices), not all subjects have the same level of knowledge of the activity to be performed. Additionally, the familiarity questionnaire revealed that a high percentage of subjects are familiar with web pages, whereas online shopping frequency among subjects is low.
2. Task performance order: there could be bias caused by the learning effect, as the tasks associated with each mechanism are performed sequentially.
3. Low user experience: all subjects are novice application users, and therefore there is a risk that they will not put in all the effort required to understand the instructions, will not understand the procedure, etc.
4. Remote usability test: all subjects performed the experiment remotely. Therefore, we could not identify or interact with participants in real time. As a result, participants could do things wrong or drop out of the experiment because they misinterpreted the task instructions and did not have the opportunity to ask what to do when they were unsure.
5. Motivation: each participant will predictably have a different reaction to the experiment and subjects could be affected negatively, especially if they are alternating experiment performance with other activities.

To mitigate any possible influence of the first two threats, the subjects were randomly assigned to balanced groups. According to Suresh [74], this randomization procedure constitutes an experimental guarantee, as interferences may or may not occur, irrespective of their impact. It is worthwhile making the effort to randomize in order to counteract any potential bias. The third threat is overcome by introducing order as a factor within the design, as detailed in Section 3.4. To try to mitigate the fourth threat, we captured the IP address of each subject and an additional contact address (for example, telephone number, email address or chat ID). We used the IP address to exclude any subjects that performed the experiment more than once. We used the contact information to gather feedback from the subject. We set simple and direct tasks in our experiment. However, we cannot guarantee that all subjects understood and correctly followed the task instructions. Finally, we cannot avoid the fifth threat [17], and we interviewed subjects at random to find out if they were affected by the factors of fatigue, boredom or similar. They should be taken into account during the analysis and interpretation of results in order to lower their impact.

5.2 External Validity

External validity refers whether and how the results can be generalized in other contexts. We believe that the experimental results cannot be generalized to all users. To prevent any potential bias caused by familiarity with the technology, the selected participants are non-computer scientists. However, the participants are regular Internet users (Section 3.3). Additionally, these subjects are

members of a sizeable part of the user population that tend not to use online shopping web applications. Therefore, we can gather quite reliable empirical evidence about the impact of the usability mechanisms analysed at non-computer scientist user level.

Also a probable threat is the generalization to other applications other than the selected domain. This threat could be dealt with by executing the experiment in other application domains.

6 Conclusions and Future Work

We studied usability mechanisms in a web application in order to experimentally evaluate the impact of usability from the viewpoint of non-computer scientist users. The tested mechanisms were: Abort Operation, Progress Feedback and Preferences. Each mechanism was evaluated with respect to three quality characteristics: efficiency, effectiveness and satisfaction.

To do this, we carried out an experiment. We developed an application and application interface. The interface automatically assigned subjects to groups and tasks and collected data. The users participating in this experiment were non-computer scientists. They were responsible for performing one task for each usability mechanism. Experimental subjects were randomly assigned to equally sized groups. Accordingly, we were able to check whether or not the group with access to the mechanism had a better perception of usability than the group without access to the mechanism.

In the following, we list some suggestions based on our experience with respect to experimental studies.

Actionable results

- The impact of a usability mechanism on the efficiency, effectiveness and satisfaction attributes is determined by the functionality proper to the mechanism and its relationship to the problem domain. Additional experiments in other domains are required to confirm whether these results apply in other domains.
- Participants' knowledge and experience both pose threats to the internal validity of the experiment. The subjects should be randomly assigned to groups.
- By evaluating the attributes with the same percentage of adopted and non-adopted usability mechanisms within each group, we can check whether or not the group with access to the mechanism has a better perception of usability than the group without access to the mechanism.

Overall, the gathered evidence tends to support an improvement in system usability. We have found no evidence that usability detracts from user performance. In this respect, the results suggest the following:

- The usability mechanisms analysed in this research have significant effects on user satisfaction. Users positively rated the benefits of the Abort Operation and Preferences usability mechanisms, whereas the impact of Progress Feedback on satisfaction is low. We believe that it might have a greater effect in contexts with high latencies.
- We found that the non-adoption of some usability mechanisms, like Abort Operation and Preferences, plays a more or less decisive role with respect to user effectiveness. On the one hand, the non-adoption of Abort Operation required additional effort on the part of users who had to undo their changes manually, as there was no direct cancellation option. On the other hand, the non-adoption of Preferences meant that the users were unable to perform the task because the interface was visually unintelligible. The non-adoption of both mechanisms tends to prevent task completion because users quit the application prematurely, are discouraged from looking for other alternatives, get frustrated early on, etc. On the other hand, their adoption facilitated successful task performance and thus effective interaction with the system.
- We gathered different evidence about the effect of usability on user speed. For the Abort Operation, we observed a 25% drop in the time taken by users. This is an appreciable improvement in the effort required by users to cancel actions. In the case of Preferences, there was no improvement in user speed because the task of searching for information on screen is easy and does not require complex interactions with the system. Progress Feedback serves the purpose of providing information on what the system is doing and aims to maintain the visibility of the system state during potential user delays produced by the search process. It should not interfere with or prevent task performance. Consequently, the times taken by users tend to be similar irrespective of mechanism adoption/non-adoption.
- The results show that Progress Feedback has a small effect on all quality characteristics. In this analysis, we set a simple task for Progress Feedback. However, we believe that this mechanism might have a different effect if the task included batch transactions involving long user waiting times.

The above conclusions are not generally applicable. However, we can say that the effort to adopt Abort Operation and Preferences is justified by their benefits with respect to effectiveness and user satisfaction. We think that the improvements with respect to efficiency, effectiveness and satisfaction are determined not only by the functionality provided by the mechanism but also by the problem domain. In this respect, a usability mechanism implemented in another domain may have a different effect on the same quality characteristics. Therefore, the adoption of usability features should be evaluated according to the benefits provided for each attribute.

With regard to the perception that usability adoption during software development is costly in terms of software engineer time and effort, our research highlights two important points. Firstly, the data from our experiment on the impact on software development and user performance of usability mechanisms provide software engineers with valuable information for prioritizing which mechanisms should be included. In this respect, we consider Preferences to be the first potential mechanism to be considered

for adoption in a system because of its low-cost implementation and its significantly positive effect on users. Abort Operation improves efficiency, effectiveness and satisfaction and should be considered in second place because it is not as cheap to implement. Finally, Progress Feedback is a desirable mechanism provided it does not compromise project resources bearing in mind that it is costly to implement and its impact on user-system interaction is low. Secondly, the results of this experiment justify the additional effort invested in improving software system usability levels. They also examine the benefits of adopting some usability recommendations, which are directly perceived by users in terms of efficiency, effectiveness and satisfaction.

Practically speaking, we admit that data about the real effect of three usability mechanisms on user performance provided by our experiment provides are preliminary. Although the results show that usability improves the performance of users on several points, they must be interpreted carefully. They are valid in the context that we defined (online shop), where we implemented a subset of scenarios for each usability mechanism. We cannot guarantee that the same results will be easily extrapolated to all mechanism scenarios and other online shopping contexts. We need to conduct further experiments to gather more evidence and confirm these results. A total of 168 subjects participated in the experiment. Most of the subjects were undergraduate students with similar knowledge and habits. This is a potential limitation, as the results may not be representative of a real population, although more and more people use the Internet and interact with such applications. Another potential limitation is that it was an unmoderated remote experiment. Although the selected domain is generally well known, most participants had never made online purchases. They might not have had a good understanding of the shopping cart philosophy and have found it difficult to perform the tasks. Although we defined rather easy-to-understand tasks, subjects might have had questions and been unable to continue with the experiment without the help of the researcher. Additionally, motivation was not good, and we cannot guarantee that all the subjects made their best effort to perform the experiment.

This research is just a first step in the empirical analysis of the impact of usability from the user viewpoint. Besides, the effect size, albeit small, is reason enough to continue the experiment and further explore the following promising lines:

- Replicate the experiment with different subjects to the sample described in this research.
- Study the impact of other usability mechanisms on the same quality characteristics from the viewpoint of users.
- Study the impact of the usability mechanisms on other quality characteristics.
- Study the impact of usability mechanisms in mobile environments.
- Redesign the experiment instrumentation for application to a domain other than online shopping.
- Redesign the tasks to create more complex scenarios.

Finally, empirical studies are gaining in importance in software engineering, generating evidence to support the evolution of knowledge in order to improve theories as a result of empirical results. Our research targets this aim.

Acknowledgments

We would like to thank all the participants in the experiments, along with the Empirical Software Engineering Research Group at the School of Computer Engineering, UPM. The reported research was partially funded by Spanish Ministry of Science, Innovation and Universities research grant RTI2018-095255-B-I00, the PGC2018-097265-B-I00 project and the FORTE-CM project (P2018/TCS-4314).

References

[1] Acuña ST, Gómez M, Juristo N (2008) Towards understanding the relationship between team climate and software quality - A quasi-experimental study. Empir Softw Eng 13:401–434. doi: 10.1007/s10664-008-9074-8

[2] Ammar L Ben, Trabelsi A, Mahfoudhi A (2016) A model-driven approach for usability engineering of interactive systems. Softw Qual J 24:301–335

[3] Andreasen MS, Nielsen HV, Schrøder SO, Stage J (2007) What happened to remote usability testing? An empirical study of three methods. Proc 25th SIGCHI Conf Hum Factors Comput Syst 1405–1414. doi: 10.1145/1240624.1240838

[4] Aveledo M, Curtino DM, la Rosa A De, Moreno AM (2012) Measuring the effect of usability mechanisms on user efficiency, effectiveness and satisfaction. In: SEKE. pp 599–604

[5] B MU (2013) Topics in Cryptology – CT-RSA 2013. Springer International Publishing

[6] Baecker RM (2014) Readings in Human-Computer Interaction: toward the year 2000. Morgan Kaufmann

[7] Basili VR, Caldiera G, Rombach HD (1994) The goal question metric approach. Encycl Softw Eng 2:528–532. doi: 10.1.1.104.8626

[8] Baxter K, Courage C, Caine K (2015) Understanding Your Users: A Practical Guide to User Research Methods, 2nd ed. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA

[9] Bender R, Grouven U (1997) Ordinal logistic regression in medical research. J R Coll Physicians Lond 31:546–551

[10] Bevan N (2001) International Standards for HCI and Usability. Int J Hum Comput Stud 55:533–552. doi: 10.1006/ijhc.2001.0483

[11] Bolchini D, Garzotto F, Sorce F (2009) Does branding need web usability? A value-oriented empirical study. In: Human-Computer Interaction -- INTERACT 2009: 12th IFIP TC 13 International Conference. Springer Berlin Heidelberg, Berlin, Heidelberg, Heidelberg, pp 652–665

[12] Brighton (2003) Usability Pattern Collection. http://www.cmis.brighton.ac.uk/research/patterns/. Accessed 1 Sep 2014

[13] Cassino R, Tucci M, Vitiello G, Francese R (2015) Empirical validation of an automatic usability evaluation method. J Vis Lang Comput 28:1–22. doi: 10.1016/j.jvlc.2014.12.002

[14] de Castro V, Genero M, Marcos E, Piattini M, Piattni M (2011) Empirical study to assess whether the use of routes

facilitates the navigability of web information systems. IET Softw 5:528–542. doi: 10.1049/iet-sen.2010.0062
[15] Charness G, Gneezy U, Kuhn MA (2012) Experimental methods: Between-subject and within-subject design. J Econ Behav Organ 81:1–8. doi: 10.1016/j.jebo.2011.08.009
[16] Constantine LL, Lockwood LAD (1999) Software for Use: A Practical Guide to the Models and Methods of Usage-centered Design. ACM Press/Addison-Wesley Publishing Co.
[17] España S, Condori N, Wieringa R, González A, Pastor Ó (2011) Model-driven system development: Experimental design and report of the pilot experiment. Comput Res Repos abs/1111.0:83
[18] Fang X, Holsapple CW (2007) An empirical study of web site navigation structures' impacts on web site usability. Decis Support Syst 43:476–491. doi: 10.1016/j.dss.2006.11.004
[19] Fernandez A, Insfran E, Abrahão S (2011) Usability evaluation methods for the web: A systematic mapping study. Inf Softw Technol 53:789–817. doi: 10.1016/J.INFSOF.2011.02.007
[20] Ferrari S, Cribari-Neto F (2004) Beta regression for modelling rates and proportions. J Appl Stat 31:799–815
[21] Ferré X, Juristo N, Windl H, Constantine L (2001) Usability basics for software developers. IEEE Softw 18:22–29. doi: 10.1109/52.903160
[22] Ferreira JM, Acuña ST (2017) A software application for collecting usability empirical data about user efficiency, effectiveness and satisfaction. In: XII Iberoamerican Conference on Software Engineering and Knowledge Engineering JIISIC'2017. p 11
[23] Golden E (2010) Early-Stage Software Design for Usability. Human-Computer Interaction: Human-Computer Interaction Institute, Carnegie Mellon University
[24] Grissom RJ (1994) Probability of the superior outcome of one treatment over another. J Appl Psychol 79:314
[25] Groth A, Haslwanter D (2016) Efficiency, effectiveness, and satisfaction of responsive mobile tourism websites: a mobile usability study. Inf Technol Tour 16:201–228. doi: 10.1007/s40558-015-0041-0
[26] Gupta D, Ahlawat AK (2017) Usability feature selection via MBBAT: A novel approach. J Comput Sci 23:195–203. doi: 10.1016/j.jocs.2017.06.005
[27] Gupta D, Ahlawat AK (2019) Taxonomy of GUM and usability prediction using GUM multistage fuzzy expert system. Int Arab J Inf Technol 16:357–363
[28] Gupta D, Ahlawat AK, Sagar K (2017) Usability prediction & ranking of SDLC models using fuzzy hierarchical usability model. Open Eng 7:161–168. doi: 10.1515/eng-2017-0021
[29] Gupta D, Rodrigues JJPC, Sundaram S, Khanna A, Korotaev V, de Albuquerque VHC (2018) Usability feature extraction using modified crow search algorithm: a novel approach. Neural Comput Appl 6. doi: 10.1007/s00521-018-3688-6
[30] Hix D, Hartson HR (1993) Developing User Interfaces: Ensuring Usability Through Product & Process. John Wiley & Sons, NY, USA
[31] Hornbæk K (2006) Current practice in measuring usability: Challenges to usability studies and research. Int J Hum Comput Stud 64:79–102. doi: 10.1016/j.ijhcs.2005.06.002
[32] Hornbæk K, Frøkjær E (2004) Two psychology-based usability inspection techniques studied in a diary experiment. In: Proceedings of the Third Nordic Conference on Human-computer Interaction. ACM, pp 3–12
[33] Idri A (2016) Usability evaluation of mobile applications using ISO 9241 and ISO 25062 standards. Springerplus 5:548. doi: 10.1186/s40064-016-2171-z
[34] ISO/IEC-25010 (2010) Systems and Software Engineering - Systems and Software Quality Requirements and Evaluation (SQuaRE) - System and Software Quality Models
[35] ISO/IEC (2004) ISO/IEC 9126-4 Software engineering -Product quality- part4: Quality In Use metrics
[36] ISO/IEC 9126 (1991) Information Technology - Software Product Evaluation - Quality Characteristics and Guidelines for Their Use
[37] ISO 9241-11 (1998) Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs)–Part II Guidance on Usability
[38] Issa A, Bong CH (2015) Measuring software quality in use: State-of-the-art and research challenges. ASQ Softw Qual Prof 17:4–15
[39] Ivory MY, Hearst MA (2001) The state of the art in automating usability evaluation of user interfaces. ACM Comput Surv 33:470–516. doi: 10.1145/503112.503114
[40] Ivory MY, Hearst MA (2001) An Empirical Foundation for Automated Web Interface Evaluation. University of California, Berkeley
[41] Jedlitschka A, Pfahl D (2005) Reporting guidelines for controlled experiments in software engineering. In: 2005 International Symposium on Empirical Software Engineering, 2005. pp 10 pp.-
[42] Johnson J, Henderson A (2012) Usability of interactive systems: It will get worse before it gets better. J Usability Stud 7:88–93
[43] Juristo N, Moreno AM (2010) Basics of Software Engineering Experimentation. Springer Publishing Company
[44] Juristo N, Moreno AM, Sanchez-Segura M-I (2007) Analysing the impact of usability on software design. J Syst Softw 80:1506–1516. doi: 10.1016/j.jss.2007.01.006
[45] Juristo N, Moreno AM, Sanchez-Segura MI (2007) Guidelines for eliciting usability functionalities. IEEE Trans Softw Eng 33:744–758. doi: 10.1109/TSE.2007.70741

[46] Kitchenham B, Madeyski L, Budgen D, Keung J, Brereton P, Charters S, Gibbs S, Pohthong A (2017) Robust statistical methods for empirical software engineering. Empir Softw Eng 22:579–630
[47] Laakso SA (2003) User Interface Design Patterns. https://www.cs.helsinki.fi/u/salaakso/patterns/
[48] Long JS, Freese J (2006) Regression models for categorical dependent variables using Stata. Stata press
[49] Macbeth G, Razumiejczyk E, Ledesma RD (2011) Cliff's Delta Calculator: A non-parametric effect size program for two groups of observations. Univ Psychol 10:545–555
[50] Mann HB, Whitney DR (1947) On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 50–60
[51] Mazumder F, Das U (2014) Usability guidelines for usable user interface. Int J Res Eng Technol 3:79–82
[52] McKnight PE, Najab J (2010) Mann-Whitney U Test. Corsini Encycl Psychol
[53] Moreno AM, Seffah A, Capilla R, Sánchez-Segura MI (2013) HCI practices for building usable software. Computer (Long Beach Calif) 46:100–102. doi: 10.1109/MC.2013.133
[54] Nachar N, others (2008) The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution. Tutor Quant Methods Psychol 4:13–20
[55] Nielsen J (1993) Usability Engineering. Morgan Kaufmann Publisher Inc.
[56] Paz F, Villanueva D, Rusu C, Roncagliolo S, Pow-Sang JA (2013) Experimental evaluation of usability heuristics. In: 2013 10th International Conference on Information Technology: New Generations. pp 119–126
[57] Piccioni M, Furia CA, Meyer B (2013) An empirical study of API usability. Int Symp Empir Softw Eng Meas 5–14. doi: 10.1109/ESEM.2013.14
[58] Poltronieri I, Zorzo AF, Bernardino M, de Borba Campos M (2018) Usa-DSL: Usability evaluation framework for domain-specific languages. In: Proceedings of the 33rd Annual ACM Symposium on Applied Computing. ACM, NY, USA, pp 2013–2021
[59] R Core Team (2009) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing
[60] Rivero L, Conte T (2016) How novice software engineers apply user interface design patterns: An empirical study. Proc Int Conf Softw Eng Knowl Eng SEKE 2016-Janua:600–604. doi: 10.18293/SEKE2016-122
[61] Rodríguez FD, Acuña ST, Juristo N (2015) Design and programming patterns for implementing usability functionalities in web applications. J Syst Softw 105:107–124. doi: 10.1016/j.jss.2015.04.023
[62] Rodríguez FD, Acuña ST, Juristo N (2015) Reusable solutions for implementing usability functionalities. Int J Softw Eng Knowl Eng 25:727–755. doi: 10.1142/S0218194015500084
[63] Rubin J (1994) Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests, 1st ed. John Wiley & Sons, Inc., New York, NY, USA
[64] Rubinstein R, Hersh H (1990) Human Factor: Designing Computer Systems for People, 24th ed. Butterworth-Heinemann, Newton, MA, USA
[65] Sagar K, Gupta D, Sangaiah AK (2018) Manual versus automated qualitative usability assessment of interactive systems. Concurr Comput 1–13. doi: 10.1002/cpe.5091
[66] Salman FA, Deraman AB, Jalil MBA (2017) The value of T-GIUESE-supporting usability evaluation practices during development product: A controlled experiment. J Theor Appl Inf Technol 95:1808–1817
[67] Sauro J, Kindlund E (2005) A method to standardize usability metrics into a single score. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp 401–409
[68] Scapin DL, Bastien JMC (1997) Ergonomic criteria for evaluating the ergonomic quality of interactive systems. Behav Inf Technol 16:220–231. doi: 10.1080/014492997119806
[69] Schade A (2013) Remote usability tests: moderated and unmoderated. Evidence-Based User Exp Res Training, Consult NN/g Nielsen Norman Gr
[70] Seffah A (2010) The evolution of design patterns in HCI: From pattern languages to pattern-oriented design. ACM Int Conf Proceeding Ser 4–9. doi: 10.1145/1824749.1824751
[71] Seffah A, Donyaee M, Kline RB, Padda HK (2006) Usability measurement and metrics: A consolidated model. Softw Qual J 14:159–178. doi: 10.1007/s11219-006-7600-8
[72] Seffah BA, Metzker E (2004) The obstacles and myths of usability and software engineering. Commun ACM 47:70–76. doi: 10.1145/1035134.1035136
[73] Shneiderman B, Plaisant C, Cohen M, Jacobs S, Elmqvist N, Diakopoulos N (2016) Designing the User Interface: Strategies for Effective Human-Computer Interaction, 6th ed. Pearson
[74] Suresh K (2011) An overview of randomization techniques: An unbiased assessment of outcome in clinical research. J Hum Reprod Sci 4:8–11. doi: 10.4103/0974-1208.82352
[75] Taguchi G, Konishi S, Konishi S (1987) Taguchi Methods, Orthogonal Arrays and Linear Graphs. American Supplier Institute Dearborn, MI
[76] Tidwell J (2010) Designing Interfaces. Patterns for Effective Interaction Design, 2nd ed. O'Reilly Media, Inc., USA
[77] Tiwari S, Gupta A (2017) Investigating comprehension and learnability aspects of use cases for software specification problems. Inf Softw Technol 91:22–43. doi: 10.1016/j.infsof.2017.06.003
[78] Torrente MCS, Prieto ABM, Gutiérrez DA, de Sagastegui MEA (2013) Sirius: A heuristic-based framework for measuring web usability adapted to the type of website. J Syst Softw 86:649–663. doi: 10.1016/J.JSS.2012.10.049

[79] Weber S, Coblenz M, Myers B, Aldrich J, Sunshine J (2017) Empirical studies on the security and usability impact of immutability. Proc - 2017 IEEE Cybersecurity Dev Conf SecDev 2017 50–53. doi: 10.1109/SecDev.2017.21
[80] van Welie M (2008) The Amsterdam Collection of Patterns in User Interface Design. http://www.welie.com/patterns/
[81] Welie M Van, Trætteberg H (2000) Interaction patterns in user interfaces. 7th Pattern Lang Programs Conf 2000 13–16. doi: 10.1.1.172.2238
[82] Wohlin C, Runeson P, Höst M, Ohlsson MC, Regnell B, Wesslén A (2012) Experimentation in Software Engineering: An Introduction. Springer Berlin Heidelberg
[83] QuickStore (Experimental Application). http://webadm.senado.gov.py/tesisweb/

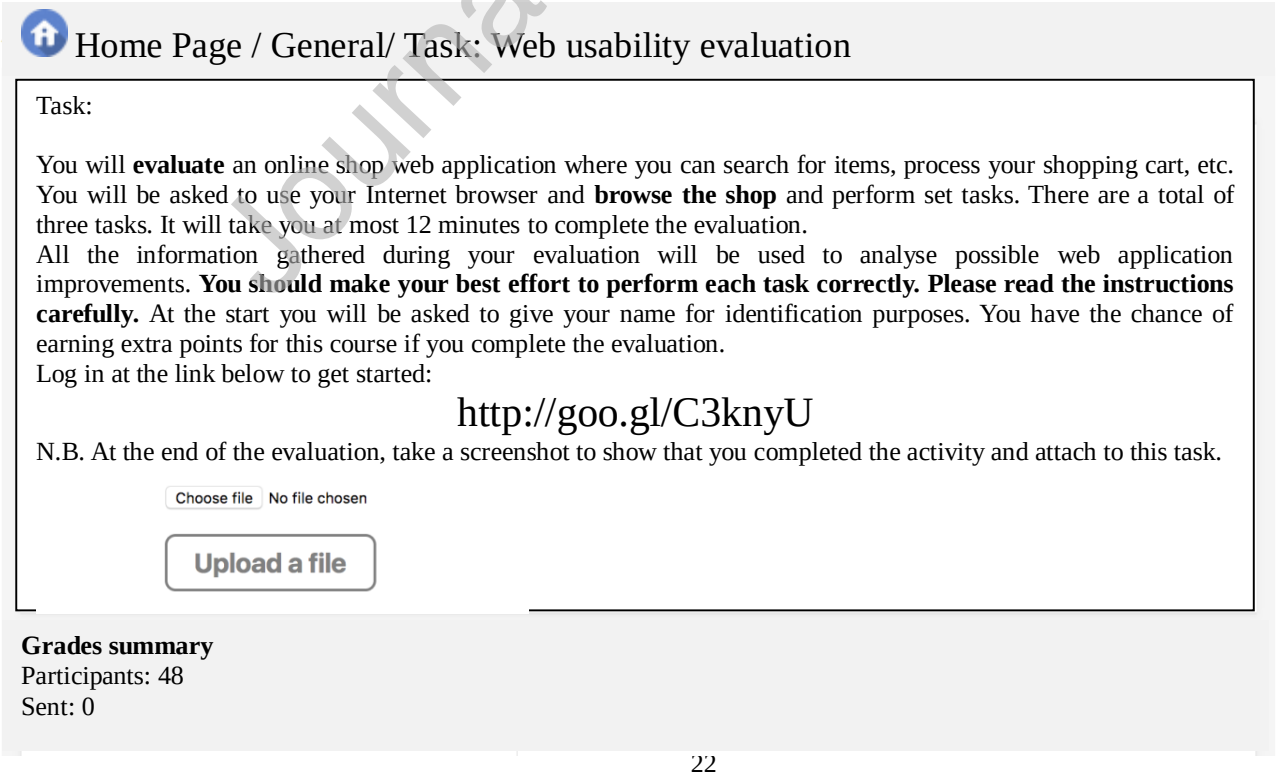
Appendices

Appendix A

Table 10: Familiarity questionnaire

Interests questionnaire	
Name or alias	
Optional data	Email: Telephone no.:
Sex	○ Male ○ Female
Age (years)	○ Under 18 ○ 18-30 ○ 31-40 ○ 41-50 ○ Over 50
Languages	□ Spanish □ English □ French □ Others
Are you a computer science professional?	○ Yes ○ No
Profession or occupation/trade	
Where do you usually use the Internet?	○ At home ○ At work ○ I do not use Internet ○ Others
What do you use the Internet for?	□ Shopping □ Work □ Education □ Entertainment □ Others
Which web applications do you use most often?	□ Facebook □ Twitter □ Google □ Yahoo □ Others
Have you ever personally shopped online?	○ Never ○ Rarely ○ Sometimes ○ Very often ○ Always
Have you shopped online with the help of middlemen?	○ Never ○ Rarely ○ Sometimes ○ Very often ○ Always
Which shops on the Internet have you used to search for or purchase items?	□ Amazon □ eBay □ Others □ None
What problems do you come across when you shop or want to shop online?	○ Displayed items are subject to unexpected charges (taxes, shipping, etc.) ○ Physical shops offer better prices and promotions ○ I don't feel safe paying online ○ I have no hang-ups ○ Others
Generally speaking, do you consider online shopping to be an appealing option?	○ Yes ○ No

Appendix B



Appendix C

Table 11: Set task and post-task questionnaire for ABR

Task #1
Imagine that you visited the shop yesterday and you have items or articles in your shopping cart. Log in and open your shopping cart containing an existing list of items, **change the number** of the first item/article on the list and enter any number of your choice, then **enter the promotional code "2015"** in the specified box and confirm the code to get the final price. You now realize that the total price is more than you expected, you change your mind and decide not to go ahead with the purchase. **Cancel the actions and undo the changes that you have just made to your shopping cart**.
Click on **[Perform task]** to log into the online shop and complete this task.

Satisfaction questionnaire for Task #1						
Items	**Strongly agree**	**Agree**	**Neutral**	**Disagree**	**Strongly disagree**	**Comments**
I managed to **abort** the operations.						
I felt confident **undoing the changes** to my shopping cart.						
I found the **system to be helpful** because it provides alerts about the changes made to my shopping cart and prompts for confirmation to continue.						

Table 12: Set task and post-task questionnaire for PFB

Task #2
Imagine that a friend has recommended that you visit the shop because you are interested in buying the book titled "The Girl on the Train". **Search for this item** and add it to your shopping cart.
Click on **[Perform task]** to log in to the online shop and complete this task.

Satisfaction questionnaire for Task #2						
Items	**Strongly agree**	**Agree**	**Neutral**	**Disagree**	**Strongly disagree**	**Comments**
I managed to **find the item** I was looking for.						
I received **feedback** while I was performing the search, that is, the system informed me about what was going on at all times.						
The **system** provided **guidance/help** while searching for the item.						

Table 13: Set task and post-task questionnaire for PRF

Task #3 (Basic)
Inspect the appearance of the virtual shop and check that you are happy with its features: font size, font type... **If you need to change something, now is the time**. Go ahead and do it.
Click on **[Perform task]** to log in to the online shop and complete this task.

Subtask #3 (Fictitious)
Now imagine that you are interested in returning an item purchased from the shop. To do this, you need to know the **how long** you have to return the item. **Browse the shop** to search for a link to this information.

Click on **[Perform task]** to log into the online shop and complete this task.

Satisfaction questionnaire for Task #3						
Items	**Strongly agree**	**Agree**	**Neutral**	**Disagree**	**Strongly disagree**	**Comments**
I managed to adapt the system interface to **my preferences (font type and size)**.						
I managed to find the **information on the returns policy.**						
It was **easy** to find information on the returns policy with the online shop **visual settings**.						
The online shop **visual settings** were helpful for finding the information on the returns policy **faster**.						

GRAPHICAL ABSTRACT