Impact of Usability Mechanisms: A Family of Experiments on Efficiency, Effectiveness and User Satisfaction

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

Abstract—*Context*: The usability software quality characteristic aims to improve system user performance. In a previous study, we found evidence of the impact of a set of usability features from the viewpoint of users in terms of efficiency, effectiveness and satisfaction. However, the impact level appears to depend on the usability feature and suggest priorities with respect to their implementation depending on how they promote user performance. *Objectives*: We use a family of three experiments to increase the precision and generalization of the results in the baseline experiment and provide findings regarding the impact on user performance of the Abort Operation, Progress Feedback and Preferences usability mechanisms. *Method*: We conduct two replications of the baseline experiment in academic settings. We analyse the data of 366 experimental subjects and apply aggregation (meta-analysis) procedures. *Results*: We find that the Abort Operation and Preferences usability mechanisms appear to improve system usability a great deal with respect to efficiency, effectiveness and user satisfaction. *Conclusions*: We find that the family of experiments further corroborates the results of the baseline experiment. Most of the results are statistically significant, and, because of the large number of experimental subjects, the evidence that we gathered in the replications is sufficient to outweigh other experiments.

Index Terms—Usability mechanism, efficiency, effectiveness, satisfaction, experimental software engineering, family of experiments

1 Introduction

Usability is a quality characteristic of a software system, which plays a more important role in highly interactive systems [1], [2], [3]. According to ISO/IEC 25010 [4], usability is defined as "the degree to which a product or system can be used by specified users to achieve specific goals with effectiveness, efficiency and satisfaction in a specified context of use". From the viewpoint of human-computer interaction (HCI), usability is related primarily to user interface design and user-system interaction [5]. HCI researchers propose recommendations for achieving a proper usability level in software systems [6], [7], [8], [9], [10], [11]. However, there is evidence that some of these recommendations also affect system functionality and not only its interface [12].

Software engineering (SE) studies how to include these HCI recommendations in software development [13], and SE experimentation aims to find empirical evidence on both final system usability and how to implement and improve usability during the software development process. There are many studies on usability evaluation related to recommendations that affect graphical interface issues, but there are very few empirical studies that address usability recommendations that affect software design and measure their benefits from the viewpoint of users [14].

In order to extend empirical evidence on the impact of HCI recommendations that affect software design, the results of an experiment studying the effect on efficiency, effectiveness and user satisfaction of three usability mechanisms —Abort Operation (ABR), Progress Feedback (PFB) and Preferences (PRF)— was reported in [14]. Usability mechanisms are functionalities that should, according to the HCI recommendations, be implemented within a software system to increase its usability. We have conducted two replications of this baseline experiment to build a family of three experiments. This paper illustrates how the results evolve from the baseline to the family of experiments.

More and more replications of experiments are being conducted in SE [15]. Different authors have analysed the process of experiment replication [16] and data aggregation techniques [17] in order to identify the best techniques for use in the field of SE. Moreover, there is unanimous agreement within the scientific community that one-off experiments are, with

- *Juan M. Ferreira is with Facultad Politécnica, Universidad Nacional de Asunción, San Lorenzo 2111, Paraguay. E-mail: jferreira@pol.una.py.*
- *Francy D. Rodríguez is with Universidad Católica de Ávila, 05002 Ávila, Spain. E-mail: fdiomar.rodriguez@ucavila.es.*
- *Adrián Santos is with M3S (M-Group), ITEE University of Oulu, 90014 Oulu, Finland. E-mail: adrian.santos.parrilla@oulu.fi.*
- *Oscar Dieste is with Universidad Politécnica de Madrid, 28660 Boadilla del Monte, Spain. E-mail: odieste@fi.upm.es.*
- *Silvia T. Acuña is with Universidad Autónoma de Madrid, 28049 Madrid, Spain. E-mail: silvia.acunna@uam.es.*
- *Natalia Juristo is with Escuela Técnica Superior de Ingenieros Informáticos, Universidad Politécnica de Madrid, 28660 Boadilla del Monte, Spain. E-mail: natalia@fi.upm.es.*

Manuscript received 15 July 2021; revised 17 December 2021; accepted 31 January 2022. Date of publication 8 February 2022; date of current version 9 January 2023. This work was funded in part by the FEDER/Spanish Ministry of Science and Innovation-Research State Agency (MASSIVE) under Grants RTI2018-095255-B-I00 and PGC2018-097265-B-I00, and in part by the R&D programme of Madrid FORTE, under Grant P2018/TCS-4314.
(Corresponding author: Juan M. Ferreira.)
Recommended for acceptance by P. Runeson.
Digital Object Identifier no. 10.1109/TSE.2022.3149586

few exceptions, of little value. The truth is that the results of a baseline experiment can only be confirmed through replication and results comparison.

A family of experiments is a set of experimental replications with access to the raw (or aggregated) data of each of at least three experiments with at least two different technologies testing the same response variable according a known experimental design and protocol [17]. The aim of replication is to provide a family of experiments to aggregate separate experiments and get more reliable results, as well as to be able to analyse aspects that individual experiments have overlooked, providing accurate information for decision making and/or more in-depth knowledge of the issue under investigation.

In this study, the goal of the baseline experiment was to evaluate the impact of three usability mechanisms (ABR, PFB and PRF) on an online shopping web application. We chose these three mechanisms because they had a greater impact on software design [12]. This impact was identified in earlier studies on usability mechanisms [5], [18]. In both cases, there is evidence that a significant design-level effort is required to include the functionalities associated with the mechanisms. Another reason for selecting these three mechanisms is that users can easily recognize their user interface components. This should facilitate their evaluation against HCI.

The evaluation was carried out using three response variables taken from the usability definition set out in ISO/IEC 25010 [4]: efficiency, effectiveness and user satisfaction. The baseline experiment was conducted with 168 users divided into 24 experimental groups. Each group performs three online shopping tasks. Efficiency was measured using number of clicks and time taken, effectiveness was gauged by percentage task completion, and user satisfaction was gathered from a questionnaire.

In this paper, we report a family of experiments that build upon a previously published baseline experiment [14]. This paper extends our previous research [14] by replicating the experiment for two new scenarios. This extension has an impact on the contents of the research, as further experimentation, calculations and analysis have to be conducted. The family of experiments that we ran compared the impact of three usability mechanisms (ABR, PFB and PRF) on efficiency, effectiveness and user satisfaction in order to increase the reliability of the results [15] of the baseline experiment [14]. In addition, the three usability mechanisms require the inclusion of additional components [12]. The inclusion of additional components leads to increased development costs and time for implementing each mechanism. Rodríguez *et al.* [5], [13] report that some mechanisms are more or less expensive to implement than others, leading to different impacts in terms of development time and cost. If there are large differences in the implementation cost of each mechanism, our family of experiments can provide valuable information for prioritizing and deciding which mechanisms a system should include.

The replications were as similar as possible to the baseline experiment. Strict replications increase sample size and thus statistical power [19], [20]. The three experiments have the same between-subjects experimental design, the same goal, the same research hypotheses and the same two-level factors: adopted and not adopted. Replication 1 evaluates two of the three baseline experiment response variables: efficiency and satisfaction. Effectiveness data are missing from Replication 1 due to a technical error concerning metric configuration in the experiment administration interface where we collected binary effectiveness data (metric initially implemented in a pilot experiment not reported in the literature) instead of percentage task completion by a subject. Replication 2 evaluates the same three response variables as the baseline experiment. There were 100 experimental subjects in Replication 1 and 98 in Replication 2, amounting to a total number of 366 subjects for the family of experiments.

The *major contribution* of this paper is that it reports a family of experiments that provides evidence of how three HCI recommendations that have an impact on design, that is, three usability mechanisms (Abort Operation, Progress Feedback and Preferences) improve the usability of a system. Data from three different samples can be aggregated by the family of experiments, leading to several findings.

Key findings

- The baseline experiment finding that ABR significantly improves efficiency, effectiveness and user satisfaction is confirmed.
- The results corroborate the fact that PFB has a negligible impact compared with the other two mechanisms, even though it is the costliest to implement.
- The family of experiments reveals that, like ABR, PRF also has a positive effect on efficiency, effectiveness and satisfaction. This contradicts the baseline experiment finding suggesting that PRF did not improve user efficiency (speed and interactivity).
- The aggregated data of the family of experiments again suggest that the three mechanisms improve system usability and does not undermine user performance.

Paper Organization. Section 2 describes work related to this research. Section 3 shows the design of the baseline experiment. Section 4 describes the replications and the results of their analysis. Section 5 analyses the results aggregation, discusses the joint results and explores the influence of demographic variables. Section 6 describes the conclusion, internal, external and construct validity threats. Finally, Section 7 presents the conclusions and future work.

2 Related Work

Even though usability is recognized as a software product quality characteristic [4], [21], many systems still do not achieve an acceptable level of usability [5], [22], [23]. A systematic literature review [24] on design patterns for mobile device interface design found that, although there are many studies on how to improve usability, there are topics or areas where information on how to adapt HCI recommendations to SE is missing. The empirical community has been studying usability from different viewpoints [14]. Some studies focus on the software process and lifecycle activities, whereas others focus on the end products. The former study how to implement or evaluate usability characteristics in the different software development lifecycle activities [12], [25], [26]. The latter focus on validating the usability of products, technologies and applications [27], [28], [29], [30]. Below, we describe empirical studies on usability, making a distinction between single experiments and families of experiments.

2.1 Usability Experiments

Some of the experimental studies evaluate or validate applications or final products focus on web applications [31], [32], [33], [34], some address mobile applications [35], [36], and some evaluate specific properties like security [37], comprehension and learnability [38] or application programming interfaces (APIs) [39]. A case study of the impact on mobile application architecture and implementation design of seven usability mechanisms, including ABR and PFB, is reported in [18]. They conclude that the mechanisms affect the overall design, and end-user satisfaction can be improved through different combinations of mechanisms. They also refer to the need to continue studying these combinations.

Reusable artefacts for implementing the three usability mechanisms addressed in this study (ABR, PFB and PRF) in web applications are proposed in [5], [13]. The empirical evidence gathered from the evaluation suggests that the implementation of each mechanism has different costs. The ABR mechanism was 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 the increase is moderate for ABR and negligible for PRF implementation. They also found that the mechanisms couple differently with application functionalities: the coupling level is high for ABR and PFB, whereas PRF can be regarded as an additional independent requirement. The PFB mechanism is harder to program because multithreading is required. Therefore, ABR can be said to be the costliest mechanism at requirements analysis level, whereas PFB appears to be the costliest mechanism at design and programming level. Finally, PRF is the least costly mechanism in both cases.

We ran the baseline experiment [14] of the family of experiments analysed in this paper. This baseline experiment evaluated the effect of three usability mechanisms on a web application. The examined usability mechanisms were ABR, PFB and PRF. We evaluated three quality characteristics that, according to ISO/IEC 25010 [4], are useful for determining product usability: efficiency, effectiveness and satisfaction. An increase in these three quality characteristics is a measure of their impact on usability, which can improve or degrade application usability. The results of the baseline experiment [14] showed that the adoption of ABR has a significantly positive effect on efficiency, effectiveness and user satisfaction, the adoption of PFB does not appear to have any impact on any of the variables, and the adoption of PRF has a significantly positive effect on effectiveness and user satisfaction, but no impact on efficiency. In no case do the usability mechanisms degrade user performance.

Separate experiments provide useful data for generating empirical evidence and improving existing knowledge, but a larger sample size can lead to new discoveries that are not observed when running a single experiment [20]. Statistical methods perform more efficiently with larger samples [17]. Although SE experimentation has increased over recent years [40], it still has the pitfall of using sample sizes that are too small to be representative [19], [41]. To overcome this shortcoming, researchers have resorted to experiment replication, which, through data aggregation, provides more evidence and increases the quality of the findings based on more evidence [40].

2.2 Families of SE Experiments

The aim of replications is to validate and round out the results of the baseline experiment [42]. Some papers focus on how to evaluate similarities or differences between the results of different replications [42], whereas others aim to reproduce the results of families of experiments and evaluate process validity [43]. In this respect, one study [44] focuses on the reproducibility of experiments. This study highlights the difficulty of running replications within SE experimentation. They concluded that, despite the use of replication packages, the communication process between researchers is still informal, costly and time consuming. They also found that, while there are robust platforms to support the technological part of the process, they are too specialized for transfer from one domain to another.

Other papers focus on identifying and analysing the techniques used to aggregate the results [17], [19]. During data aggregation, the effect sizes of all the replications are calculated first based on descriptive statistics, like means, variances or sample sizes or results of the experiment statistical tests, and are then combined using a meta-analysis model [19], [45]. Techniques like aggregated data (AD), narrative synthesis, individual participant data stratified (IPD-S) and aggregation of p-values are used to analyse families of experiments. IPD-S and AD were found to be the best techniques for analysing families of SE experiments [17], and all the data of the experiments that are part of the family have to be analysed jointly, recognizing their source experiment [19].

There is one family of experiments that compares three requirements elicitation methods [46]. Its aim was to help developers select the best method. Other experimental studies analyse model-driven development (MDD) in terms of final software quality [20] and maintainability [47]. We also found studies that define a framework or evaluate MDD tool usability [20], [48], [49], which could be used to conduct families of experiments.

Another family of experiments evaluated whether the use of test-driven development (TDD) improves software product quality [50]. The family is composed of 12 separate experiments and aims to improve the accuracy and generalizability of the results. The study evaluates whether the characteristics of the experiments affect the results of TDD performance in terms of quality.

Finally, the only study that we have found using families of experiments to evaluate usability-related aspects plans a family of experiments to empirically evaluate a web usability evaluation process (WUEP) proposed by the authors [51] within the framework of MDD use. There were 64 participants in the family of experiments, including PhD and MS computer science students. The objective of the experiments was to evaluate the participants' effectiveness, efficiency, perceived ease of use and perceived satisfaction when using WUEP compared to heuristic evaluation.

Our review of the related work retrieved only one paper [51] using families of experiments to evaluate application

usability according to HCI recommendations. Considering the importance of families of experiments and research into the best techniques to analyse results [17], we conducted this study on a family of experiments generated by two replications of the baseline experiment reported in [14].

Our family of experiments is designed to improve the accuracy of the results regarding the implementation of specific usability mechanisms and their effect on the final usability of a web application. The results should provide software engineers with criteria for evaluating and prioritizing usability mechanisms and making more reliable decisions on which usability mechanisms to best implement within a specific system or web application.

3 Baseline Experiment

This section reports the definition, design and settings of the baseline experiment. We published the details in a previous paper [14]. We conducted two strict replications, which, together with the baseline experiment, constitute a family of three experiments.

3.1 Goal, Research Questions and Hypotheses

The research goal of this experiment is to evaluate the impact of three usability mechanisms (ABR, PFB and PRF) on a web application. The research question (RQ) is: Does the adoption of usability mechanisms improve application usability in terms of efficiency, effectiveness and user satisfaction?

The research question is further divided into three specific research questions:

- RQ1: Does the adoption of the ABR usability mechanism improve application usability in terms of efficiency, effectiveness and user satisfaction?
- RQ2: Does the adoption of the PFB usability mechanism improve application usability in terms of efficiency, effectiveness and user satisfaction?
- RQ3: Does the adoption of the PRF usability mechanism improve application usability in terms of efficiency, effectiveness and user satisfaction?

The null hypothesis governing these three specific research questions is H.1.x.0: There is no significant difference in user EFFICIENCY | EFFECTIVENESS | SATISFACTION with or without the adoption of the usability mechanism. 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). For RQ1, the three hypotheses are:

- H.1.1.0: There is no difference in EFFICIENCY 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.

The three null hypotheses for RQ2 and RQ3 are formulated similarly.

3.2 Factors and Response Variables

The factor or independent variable [31] defined for the family of experiments is the usability mechanism with two levels: adopted and not adopted. *Adopted* means that a specified usability mechanism is adopted during task performance. *Not adopted* indicates that a specified usability mechanism is not adopted during task performance.

The baseline experiment aimed to evaluate the effect of the usability mechanism through the efficiency, effectiveness and user satisfaction response variables. According to ISO/IEC 25010 [4], efficiency refers to resources expended by users to correctly and completely achieve specific goals, effectiveness is the degree to which users correctly and completely achieve specific goals, and satisfaction is the degree to which users' needs are satisfied by using a product or system in a specified context of use.

In the following, we describe the metrics used for each response variable —efficiency, effectiveness and satisfaction—:

- Efficiency is measured according to two metrics:
  a) Speed: time measured in seconds taken by a subject to complete the task [52]. The elapsed 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) Interactivity: number of clicks made by a subject to complete the task [53], [54]. We count separate clicks, where a double click is classed as two separate clicks. Efficiency measured as user interactivity can be represented by:

$$Ef_{interactivity} = \ count(separateClicks)$$

- Effectiveness: percentage task completion by a subject [55]. Effectiveness can be represented by:

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

- Satisfaction: mean value of the responses to the post-task questionnaire questions. The questionnaire responses are ordinal values on a Likert scale (1 = Strongly disagree to 5 = Strongly agree) [56]. 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 baseline experiment was conducted in two contexts: academic setting and non-academic setting [14]. The academic setting included undergraduate students from different degree programmes (economic and business science, legal science, health science, etc.). The non-academic setting included practitioners and non-practitioners who were sent an invitation to participate via messaging applications or electronic mail. The experiment was executed in each context at different non-overlapping time periods. The experimental

TABLE 1
Treatment Matrix

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

TABLE 2
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

subjects were not computer science specialists. The experiment had a total of 168 participants: 88 from the academic setting and 80 from the non-academic setting.

The biggest concentration of participants spanned two main age groups: 18–30 years (61%) and 31–40 years (26%). 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.

3.4 Experimental Design

The family of experiments uses a between-subjects design with orthogonal array [57], [58]. Each experimental subject was placed in one group and sequentially performed randomly assigned tasks to interact with all three (adopted or non-adopted) usability mechanisms. Thus, each subject interacts with only one mechanism (adopted or non-adopted) at any one time during task performance. Accordingly, when the experimental subject completes the task, we measure a single value for each metric measured for each response variable: efficiency, effectiveness and satisfaction. Based on the statistical analysis, we can then compare whether or not the usability perceived by the group with the adopted mechanism is greater than for the group with the mechanism disabled.

Our design is composed of a treatment matrix, a mechanism exposure order matrix and a group assignment matrix. Table 1 shows the treatment matrix describing which usability mechanisms will 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 2 shows the order of exposure for each factor. This matrix establishes all the possible task performance sequences for each factor (without repetitions).

Finally, each row of the treatment matrix is combined with each row of the exposure order matrix to produce 24 groups (group assignment matrix). The group assignment matrix is available in Appendix A, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TSE.2022.3149586.

3.5 Instrumentation and Tasks

The family of experiments uses web application software: an online shop called QuickStore [59], [60]. The application and user interface design include automatic group and task assignment, as well as data collection. Therefore, the adoption of usability mechanisms is assigned randomly in the experiment by the application. Each subject performs three tasks. The tasks are:

- Abort Operation: the subject applies 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 has been adopted, 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 has not been adopted, the user will have to manually undo each change made since the start of the task one by one.
- 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 [59], 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 has not been adopted, the user will not be informed during the search that the action is ongoing.
- Preferences: this task is divided into two parts. The user will perform first the basic task and then the fictitious task. Basic Task: the subject should customize the application user interface. The font size of the original interface is small and not very legible. On the one hand, if the usability mechanism has been adopted, the user can customize some shop features to his or her liking. On the other hand, if the mechanism has not been adopted, the user cannot modify the application interface appearance. Fictitious 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.

3.6 Operation

The baseline 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)[1]. Each professor published the experiment link on his or her course e-campus. Over the last month (July), the experiment was conducted outside academia. We informed subjects that participation was voluntary. We encouraged the students that agreed to participate to do their best to perform the tasks, although it was an optional challenge that had no bearing on their learning outcomes.

At the time of experiment execution, the subjects were not familiar with the aim of the study or with the research hypotheses. Apart from the link [59] that each participant was to use to log in and start the evaluation, we did not provide any additional material. Originally, we collected data for a total of 182 subjects. However, we removed data for 14 subjects because they did not correctly complete the tasks. Finally, 168 valid data remained for the statistical analysis and results interpretation.

4 Replications

This section describes the replications conducted and the results of the analysis of each replication, highlighting the similarities and differences to the baseline experiment. The baseline experiment concluded that the impact of a mechanism may depend on other factors and vary depending on the context [14]. Therefore, it was necessary to conduct further experiments to gather more evidence and confirm the results.

The two replications are experiments executed in a realistic environment (web application software executing real user actions) conforming to a between-subjects design. They have the same goal, research questions, hypotheses and instrumentation. The replications differ as to the experimental subjects. Like the baseline experiment, none of the subjects who participated in the replications were computer science specialists. This experimental constraint underpinned the idea that subjects with little or no computer expertise can use the system and appreciate the benefits of usability. Additionally, it rules out the influence of computer-literate users who may be familiar with this type of applications.

To describe each replication of the original empirical research, we apply the guidelines defined for reporting experimental replications proposed by Carver [61]. To analyse the family of experiments, we apply Steps 1 to 4 of the guidelines recommended by Santos *et al.* [19]:

- Step 1: Describe the participants.
- Step 2: Analyse individual replications.
- Step 3: Aggregate the results.
- Step 4: Conduct exploratory analyses.

Throughout this section, we describe the participants in each replication according to Step 1 of the guidelines published by Santos *et al.* [19].

4.1 Replication 1

The experimental subjects of Replication 1 are students from Rodeira Secondary School in Galicia (Spain), who volunteered to participate in the experiment with the consent of their teachers. We conducted this replication in August 2016. To rule out the learning effect, we did not hold any informative or practice sessions beforehand. All the subjects completed a familiarity questionnaire. The details of the sample are as follows:

- The sample was composed of 100 subjects, of which 43 were males and 57, females.
- With regard to age, 85% of subjects were aged under 18 years, 10% were members of the 18 to 30 age group, and 5% were aged over 30.
- The subjects connect to the Internet at home. Some also use the Web at work or elsewhere. The primary uses are for entertainment and education.
- With respect to subjects' online shopping habits, most participants had never purchased anything over the Internet (37%), whereas 32% shopped online occasionally, 17% rarely, 9% almost always and only 5% always. As with our baseline experiment, this is an advantage as most subjects are not acquainted with the application domain and are therefore more sensitive to system usability.

4.2 Replication 2

We ran this experiment in an academic setting with first-year students of accountancy, law, sport sciences and health sciences at the Universidad Autónoma de Asunción (Paraguay), all of whom participated on a voluntary basis. We conducted this replication over a two-month period from November to December 2016.

Like the baseline experiment and Replication 1, we did not hold any informative session beforehand, again to rule out the learning effect. We gave participants an overview of the application, introducing the structure of the experiment to assure that they were able to successfully perform the tasks. The results of the familiarity questionnaire completed by subjects before the start of the experiment were as follows:

- The final sample included 98 subjects, of which 39 were males and 59, females.
- Most participants were aged from 18 to 30 years (89%), except nine that were members of the 31 to 40 age group (9%) and two within the 41 to 50 age group (2%).
- The experimental subjects are regular Internet users. They connect to the Web at home (73%), at work (19%) and to a lesser extent elsewhere (8%).
- Most participants had never shopped online (63%), whereas 18% rarely, 15% occasionally, and 4% more often (always or almost always) shopped online.

Table 3 provides a detailed summary of the subjects that participated in the family of experiments, specifying the differences between the baseline experiment, and Table 4 describes the profile of the researchers that participated in each experiment from design to results analysis.

4.3 Analysis of Replications

Following Step 2 of the guidelines published by Santos *et al.* [19], we describe and analyse the data of each replication with consistent statistical techniques. We analyse the replications following the same procedure as enacted in the baseline

1. http://e.uaa.edu.py/

TABLE 3
Summary of Subjects

	Baseline experiment	Strict Rep1	Strict Rep2
Subjects type	Academic and non-academic	Academic	Academic
Number of participants	168	100	98
Men	76	43	39
Women	92	57	59
Age range with the largest number of participants	18-30	< 18	18-30

Rep means Replication: both terms are used indistinctly hereinafter.

TABLE 4
Summary of Experimenters

Experimenters	Baseline experiment	Strict Rep1	Strict Rep2
Designer	Academic staff from UPM-UAM-UNA	Academic staff from UPM-UAM-UNA	Academic staff from UPM-UAM-UNA
Monitor	Academic staff from UNA	Academic staff and student from UAM	Academic staff from UNA
Measurer	Academic staff from UNA	Academic staff and student from UAM	Academic staff from UNA
Analyst	Academic staff from UPM-UAM-UNA	Academic staff from UPM-UAM-UNA	Academic staff from UPM-UAM-UNA

UPM: Universidad Politécnica de Madrid | UAM: Universidad Autónoma de Madrid | UNA: Universidad Nacional de Asunción

experiment [14]. Briefly, we divide the analysis into three parts, one for each usability mechanism: ABR, PFB and PRF. In each part, we evaluate the impact of usability, measured by clicks, times, percentage task completion and satisfaction. We provide the violin and box plots and evaluate the statistical significance (p-value). We use violin and box plots to illustrate the score distributions for each response variable and to show the data distribution shape (which varies enormously from one distribution to another). We compared two groups: one group in which the usability mechanism was adopted and another group in which the mechanism was not adopted. According to [62], a violin plot synergistically combines the box plot and the density trace and should be interpreted as follows: wider sections of the violin plot represent a higher probability that members of the population will take on the given value; the skinnier sections represent a lower probability. This visualizes where more points are clustered within the box plot range. Clusters of data appear as bumps in density estimators [62]. Therefore, the peaks, valleys, and tails of each group's density curve can be compared to see similarities and differences between groups. We reported summary statistics (mean, median and standard deviation, p-value) in order to round out the violin and box plots.

We use the Mann-Whitney U test [63] to evaluate statistical significance. Note that the Mann-Whitney U test is a scale-free statistical test and can assess the statistical significance of all response variables irrespective of the data type (continuous, discrete, ordinal, etc.).

We have not removed any outliers because they are regarded as legitimate experiment values. For readability and reasons of space, we report the statistical analysis of ABR only. The analysis for PFB and PRF is described in Appendix B, available in the online supplemental material.

4.3.1 Abort Operation

Table 5 shows the summary statistics for each response variable distribution (depending on whether the ABR usability mechanism is or is not adopted) in all the replications. The respective violin and box plots are show in Fig. 1. The line between the two boxes connects the means.

Efficiency. As Fig. 1 shows, the wider section of adopted ABR is further down than for the non-adopted ABR, meaning that the subjects using the system with adopted ABR appear to be more efficient in terms of clicks and time. Table 5 shows that the difference is statistically significant (in terms of clicks and time) for Replication 2.

Effectiveness. The result shows that there is a considerable difference between adopted and non-adopted ABR in Replication 2 (Fig. 1). Adopted ABR data clusters are higher up than for non-adopted ABR. Table 5 shows that this difference is significant and appears to be greater when ABR is adopted. There are no data on effectiveness for Replication 1.

Satisfaction. Fig. 1 indicates that there is an observable increase in the mean satisfaction across replications: the subjects appear to be more satisfied when ABR is adopted. The shape of the non-adopted ABR shows more disperse data than for adopted ABR. The difference in user satisfaction is found to be statistically significant in all replications.

Comparing the results of the replications with the baseline experiment for ABR, we find that Replications 1 and 2 return similar results to the baseline experiment. The adoption of ABR improves efficiency (speed), effectiveness and user satisfaction. However, the adoption of ABR does not appear to improve user efficiency in terms of interactivity.

5 Analysis Approach

Following Step 3 of the guidelines by Santos *et al*. [19], this section analyses the results aggregation and discusses the results. Due to the heterogeneity of the resulting impacts in the three experiments considering the three usability mechanisms, it would be premature to draw conclusions based on the separate results of each experiment. Besides, aggregation procedures would mitigate the threat to the generalization and reliability of the results of the individual experiments [14], [19]. Note that rather than reproduce the published baseline experiment results, our aim is to pool together the different experiments in order to

TABLE 5
Summary Statistics and Statistical Significance Assessment for ABR: Rep1 and Rep2

Response Variable	Rep.	Group	Mean	Median	SD	p-value
Click	Rep1	Not adopted	20.41	15	16.76	0.11
		Adopted	16.80	11	16.85	
	Rep2	Not adopted	17.13	15	12.12	0.04 *
		Adopted	12.49	12	6.93	
Time	Rep1	Not adopted	160.01	141.072	100	0.03 *
		Adopted	124.73	103.2	83.28	
	Rep2	Not adopted	191.61	174.02	123.54	0.006 *
		Adopted	122.37	103.64	77.17	
Percentage	Rep1	Not adopted	-	-	-	-
		Adopted	-	-	-	
	Rep2	Not adopted	68.09	75	19.30	<0.001 *
		Adopted	86.27	100	28.86	
Satisfaction	Rep1	Not adopted	2.36	2	1.22	<0.001 *
		Adopted	3.68	4	1.27	
	Rep2	Not adopted	2.90	3	1.30	<0.001 *
		Adopted	3.96	4	1.02	

understand the effect of the usability mechanism in a broader setting [42].

In our case, the family of experiments is composed of three experiments: (a) the baseline experiment, with 168 subjects; (b) Replication 1, with 100 subjects, and (c) Replication 2, with 98 subjects. The baseline experiment and Replication 2 measure all three response variables: effectiveness, efficiency and satisfaction. Replication 1 measures two of the response variables: efficiency and satisfaction. The experimental data on effectiveness are missing on technical grounds. The power analysis reported in Appendix C, available in the online supplemental material, shows that we need 125 subjects to achieve 80% power for effectiveness (percentage task completion). We double this value for all three usability mechanisms. Therefore, the number of subjects in our family of experiments does not appear to constitute a validity threat. Altogether, the three experiments comprised 366 subjects.

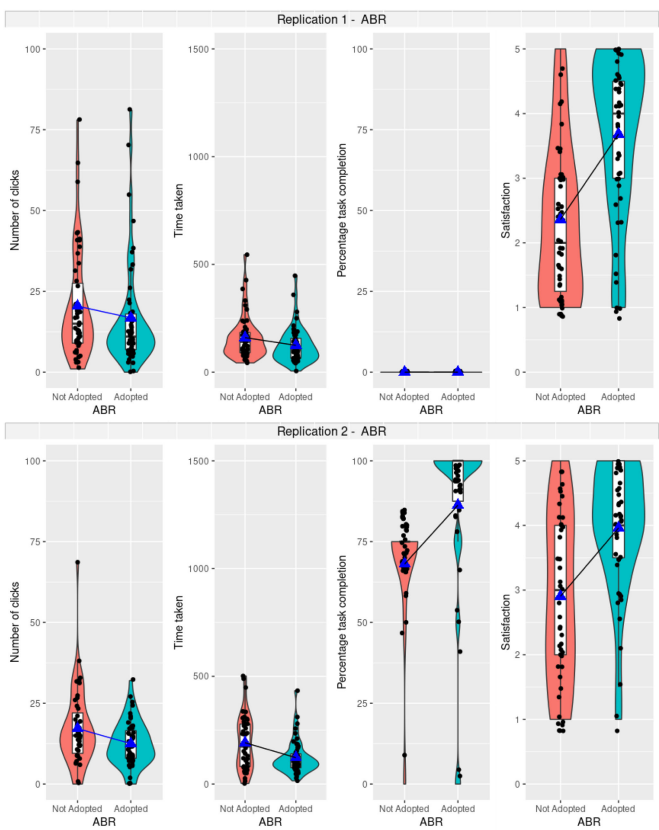

Fig. 1. Violin plots for the number of clicks, elapsed time, percentage task completion and satisfaction with the ABR usability mechanism: Rep1 and Rep2.

We divided the analysis of the family of experiments into three different parts, one per usability mechanism: Abort Operation, Progress Feedback and Preferences. We assessed four response variables according to each usability mechanism: CLICK (i.e., number of clicks), ELAPSED_TIME (i.e., time taken to perform the task), PERCENTAGETASK (i.e., percentage task completion) and VALUE (i.e., satisfaction score on a 1-to-5 Likert scale). We did not measure the PERCENTAGETASK response variable in Replication 1.

We followed an identical analysis procedure for each usability mechanism (i.e., within each part):

- We provided a profile plot showing the average score of the subjects for each response variable divided by the adoption/non-adoption of the usability mechanism across the replications. The lines linking points show whether the estimated marginal means are increasing or decreasing across levels (adopted and non-adopted) [19], [64]. We made preliminary observations with respect to the differences in the results across the experiments.
- Following the conventions used in medicine to analyse groups of interrelated experiments, we fitted fixed-effects linear regression models with the main factor TREATMENT and EXPERIMENT to analyse the data [65], [66]. We chose linear regression over meta-analysis of effect sizes [45], as: (1) access to the raw data is guaranteed within the family, and (2) all the replications have identical response variable operationalizations. We fitted a fixed-effects linear

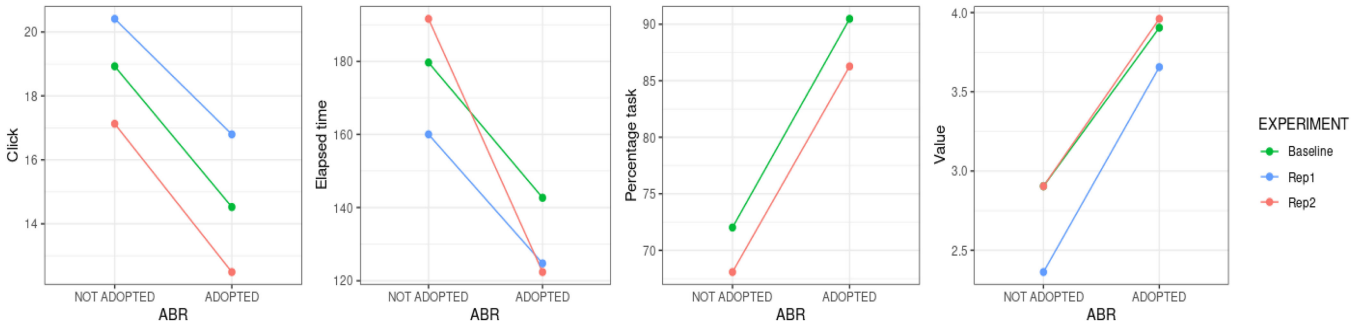

Fig. 2. Profile-plot for ABR.

regression instead of a random-effects model (i.e., linear mixed model with EXPERIMENT as a random factor and TREATMENT as a random effect) because: (1) experiment operationalizations are identical, and (2) populations are similar across the replications. Two important assumptions need to be met within fixed-effects linear models: the normality assumption and the homoscedasticity assumption (i.e., the equality of variances across the treatment groups [66]). The normality assumption is tenable due to the relatively large sample size achieved at the family level (i.e., sample size in the hundreds [67], [68]). With regard to the homoscedasticity assumption, we fitted generalized least squares models [69] accommodating different variances across treatment groups and experiments (i.e., allowing for heteroscedasticity) to assess the robustness of the linear regression results. As the linear regression and generalized least square results were similar, we chose to interpret the statistical significance and practical significance of results using the most parsimonious model (i.e., the linear regression model).

- We assessed the statistical significance of results according to the p-value of the TREATMENT estimate. We assessed the practical significance of the results according to: (1) the sign of the TREATMENT estimate, and (2) the magnitude of the TREATMENT estimate with respect to the intercept term (i.e., the non-adopted condition in the baseline experiment, since the non-adopted condition in the baseline experiment is taken as the reference class in all the fitted fixed-effects regressions). If the control conditions differ markedly across the replications (and, thus, the estimate of the control condition in the baseline experiment is uninformative for assessing the magnitude of the TREATMENT estimate), we assess the magnitude of the TREATMENT estimate considering the control estimates of the other experiments also.
- To ease the integration of results in future meta-analyses [45], we provide Cohen's d effect sizes [70], alongside their interpretation (i.e., small, medium, large) and their corresponding variances for all the pairwise comparisons made (i.e., the adoption/non-adoption of each usability mechanism for each response variable) for all the experiments. We used the R package effsize [71] to compute the effect sizes and their respective variances.

Throughout this section, we analyse the data of each usability mechanism one by one (i.e., Abort Operation, Progress Feedback and Preferences).

5.1 Abort Operation Analysis

Fig. 2 shows the profile plot for CLICK, ELAPSED_TIME, PERCENTAGETASK and VALUE by adoption/non-adoption of the Abort Operation usability mechanism across all experiments.

As shown in Fig. 2, the averages are consistent across the replications: the adoption of the Abort Operation decreases (1) the number of clicks and (2) the elapsed time across all the experiments, and increases (1) the percentage of task completion and (2) subject satisfaction. Therefore, the subjects appear to be relatively more efficient (fewer clicks and less time), more effective and more satisfied when ABR is adopted across experiments.

Table 6 shows the results of the linear regression models fitted to analyse the adoption or non-adoption of Abort Operation across experiments.

As Table 6 shows, the Abort Operation usability mechanism appears to have a remarkable impact on the number of clicks (i.e., a drop in the number of clicks of around 23% (i.e., 4.25/18.85) with respect to the intercept: the average score calculated for the non-adopted abort operation condition in the baseline experiment). This drop appears to be larger for elapsed time (i.e., a drop of around 24% in time). We also find an increase of around 25% in percentage task completion, and a larger increase in satisfaction (i.e., an increase of almost 39%). Thus, overall, the adoption or non-adoption of the Abort Operation usability mechanism appears to have a major impact on system usability.

Table 7 shows Cohen's d effect sizes, interpretations (i.e., small, medium, large), and respective variances for each replication.

TABLE 6
Linear Regression Coefficients for ABR

Coefficient	Click	Time	Percentage	Satisfaction
Intercept	18.85 (1.21)***	183.76 (10.99)***	72.07 (2.05)***	2.85 (0.11)***
Adopted	-4.25 (1.36)**	-45.16 (12.33)***	18.36 (2.56)***	1.10 (0.12)***
Experiment = = Rep1	1.87 (1.64)	-18.90 (14.90)		-0.38 (0.15)*
Experiment = = Rep2	-1.93 (1.65)	-4.68 (14.99)	-4.07 (2.65)	0.03 (0.15)

Significance levels: *** *(p<0.001)*, ** *(p<0.01)*, * *(p<0.05)*, '.' *(p<0.1)*.

TABLE 7
Replication Effect Sizes for ABR

Response variable	Experiment	d	vi	Interpretation
Click	Baseline	-0.3634	0.0242	small
	Rep1	-0.2151	0.0402	small
	Rep2	-0.4748	0.0420	small
Time	Baseline	-0.2672	0.0240	small
	Rep1	-0.3827	0.0407	small
	Rep2	-0.6786	0.0432	medium
Percentage	Baseline	1.0102	0.0268	large
	Rep1	-	-	-
	Rep2	0.7352	0.0436	medium
Satisfaction	Baseline	0.8543	0.0260	large
	Rep1	1.0398	0.0459	large
	Rep2	0.9053	0.0451	large

5.2 Progress Feedback Analysis

Fig. 3 shows the profile plot for CLICK, ELAPSED_TIME, PERCENTAGETASK and VALUE by adoption/non-adoption of the Progress Feedback usability mechanism across all experiments.

As Fig. 3 shows, the sign of the effects appears to be consistent across all experiments, except for percentage task completion (where the baseline average appears to be unchanged irrespective of whether or not the mechanism is adopted). Of all the experiment participants, Replication 2 subjects appear to experience the largest drop in number of clicks, elapsed time, and percentage task completion. Note that mean differences do not appear to be significant. However, the profile plot scale creates the impression that there is big drop between Replication 2 and the other two experiments, where there really there is none. Besides, with respect to Replication 1 and baseline experiment subjects, there is a smaller increase in satisfaction among Replication 2 participants when the mechanism is adopted.

Table 8 shows the results of the linear regression models fitted to analyse the adoption or non-adoption of the Progress Feedback mechanism across experiments.

As shown in Table 8, the Progress Feedback usability mechanism appears to affect the response variables to a smaller extent than the other usability mechanisms. Specifically, the number of clicks and elapsed time are reduced by around only 18% and 7%, respectively, and the percentage task completion by a negligible amount (i.e., around 2%). Besides, the satisfaction scores do not appear to increase much either (just 10%). In view of these findings, the adoption or non-adoption of the Progress Feedback mechanism does not appear to have much of an impact on system usability. Interactions may be in operation, and, therefore, moderators should be identified in order to explain the heterogeneity of results.

TABLE 8
Linear Regression Coefficients for PFB

Coefficient	Click	Time	Percentage	Satisfaction
Intercept	9.86 (1.10)***	111.29 (8.80)***	30.74 (4.49)***	2.05 (0.11)***
Adopted	-3.15 (1.08)**	-18.79 (8.61)*	42.10 (5.59)***	1.44 (0.11)***
Experiment == Rep1	0.09 (1.23)	5.86 (9.79)		-0.11 (0.12)
Experiment == Rep2	-0.09 (1.58)	2.39 (12.64)	-1.19 (5.79)	0.38 (0.15)*

Significance levels: *** *(p<0.001),* ** *(p<0.01),* * *(p<0.05), '.' (p<0.1).*

Table 9 shows Cohen's d effect sizes, interpretations (i.e., small, medium, large), and respective variances for each replication.

5.3 Preferences Analysis

Fig. 4 shows the profile plot for CLICK, ELAPSED_TIME, PERCENTAGETASK and VALUE by adoption/non-adoption of the Preferences usability mechanism across all experiments.

As shown in Fig. 4, the direction of the effects is consistent across the experiments: while the Preferences mechanism decreases the number of clicks and the elapsed time, it increases percentage task completion and user satisfaction.

Table 10 shows the results of the linear regression models fitted to analyse the adoption or non-adoption of the Preferences mechanism across experiments.

As Table 10 shows, the Preferences usability mechanism appears to have a considerable effect on the number of clicks (i.e., leading to a drop in the number of clicks of around 32%), a smaller effect on the elapsed time (i.e., leading to a drop of around 17% in time), a larger effect on task completion (i.e., an increase of almost 137%) and satisfaction (i.e., an increase of around 70%). Therefore, the adoption or non-adoption of the Preferences usability mechanism appears to have a substantial impact on system usability, especially percentage task completion and user satisfaction.

Table 11 shows Cohen's d effect sizes, interpretations (i.e., small, medium, large), and respective variances for each replication.

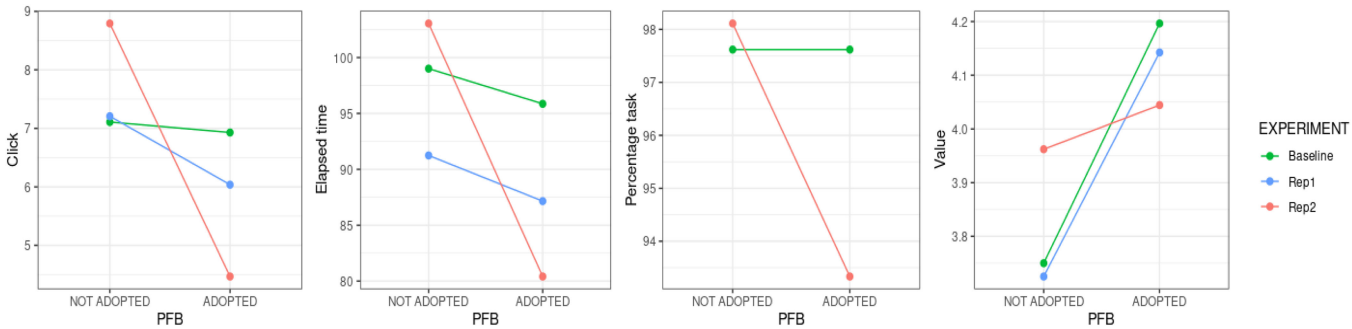

Fig. 3. Profile plot for PFB.

TABLE 9
Replication Effect Sizes for PFB

Response variable	Experiment	d	vi	Interpretation
Click	Baseline	-0.0250	0.0238	small
	Rep1	-0.1910	0.0150	small
	Rep2	-0.5272	0.0425	medium
Time	Baseline	-0.0328	0.0238	small
	Rep1	-0.0511	0.0149	small
	Rep2	-0.2445	0.0414	small
Percentage	Baseline	-0.0000	0.0238	small
	Rep1	-	-	-
	Rep2	-0.2408	0.0414	small
Satisfaction	Baseline	0.4831	0.0245	small
	Rep1	0.4762	0.0154	small
	Rep2	0.0794	0.0411	small

TABLE 10
Linear Regression Coefficients for PRF

Coefficient	Click	Time	Percentage	Satisfaction
Intercept	7.73 (0.61)***	101.02 (7.74)***	98.50 (1.69)***	3.79 (0.08)***
Adopted	-1.43 (0.60)*	-7.18 (7.58)	-1.75 (2.10)	0.37 (0.08)***
Experiment == Rep1	-0.39 (0.68)	-8.20 (8.61)		-0.04 (0.09)
Experiment == Rep2	-0.27 (0.88)	-5.07 (11.13)	-1.77 (2.18)	0.04 (0.12)

Significance levels: *** *(p<0.001),* ** *(p<0.01),* * *(p<0.05), '.' (p<0.1).*

5.4 Summary of the Results

In this section, we discuss the quantitative results in response to the research questions. Table 12 summarizes the experiment results. The percentage values signify the ratio of the adopted factor (i.e., 4.25/18.85 for clicks) to the intercept, that is, the average score calculated for the non-adopted condition in the baseline experiment. The (-) sign means that the adoption of a mechanism has a negative effect on the response variable. The $*$ symbol denotes that the input is significant.

5.4.1 RQ1 Abort Operation

The adoption of Abort Operation has a favourable impact on efficiency, effectiveness and user satisfaction. Table 12 shows, for efficiency, an improvement of around 23% in user speed and interactivity. Effectiveness and user satisfaction are higher when this mechanism is adopted (25% and 39%, respectively). Additionally, the input of this mechanism is statistically significant for all response variables. *We conclude that the adoption of ABR improves efficiency, effectiveness and user satisfaction.*

5.4.2 RQ2 Progress Feedback

Table 12 shows that, on the one hand, user interactivity and satisfaction were slightly better (i.e., 18% and ∼10%) for subjects that had access to the mechanism than for others that did not. In both cases, the input is statistically significant. On the other hand, the adoption of PFB does not play a key role in either effectiveness or efficiency in terms of user speed. *We conclude that the adoption of PFB does not improve efficiency, effectiveness and user satisfaction.*

5.4.3 RQ3 Preferences

The results shown in Table 12 confirm that PRF has a positive impact on all response variables. In particular, the adoption of PRF was a decisive factor for improving effectiveness and user satisfaction (137% and 70%, respectively). Also, there is evidence that users are 17% faster (take less time) and interact less with the system (32%) (employ fewer clicks) to perform the specified tasks. The input of this mechanism is statistically significant. *We conclude that the adoption of PRF improves user efficiency, effectiveness and satisfaction.*

The findings on ABR and PFB are consistent with the baseline experiment. However, the family of experiments has revealed a change in the impact of PRF on user efficiency. In the baseline, we hypothesized that PRF appears to have a neutral effect on efficiency (speed and interactivity). Our family of experiments provides new findings, suggesting that PRF has a positive effect on efficiency, effectiveness and user satisfaction.

5.5 Influence of Demographic Variables

In this section, we address Step 4 of the guidelines published by Santos *et al.* [19] with respect to conducting an exploratory analysis to identify moderators. Although the findings tend to be stable and consistent, there is still an observable heterogeneity of outcomes within our family. For example, unlike the ABR and PRF mechanisms, PFB does not appear to have an appreciable impact on system usability, which suggests that there may be plausible moderators that could be influencing the results. Moderator variables like age, gender and online shopping experience are the focus of this section.

We conducted an exploratory analysis to understand the extent to which age, gender and shopping experience affect efficiency, effectiveness and user satisfaction for each

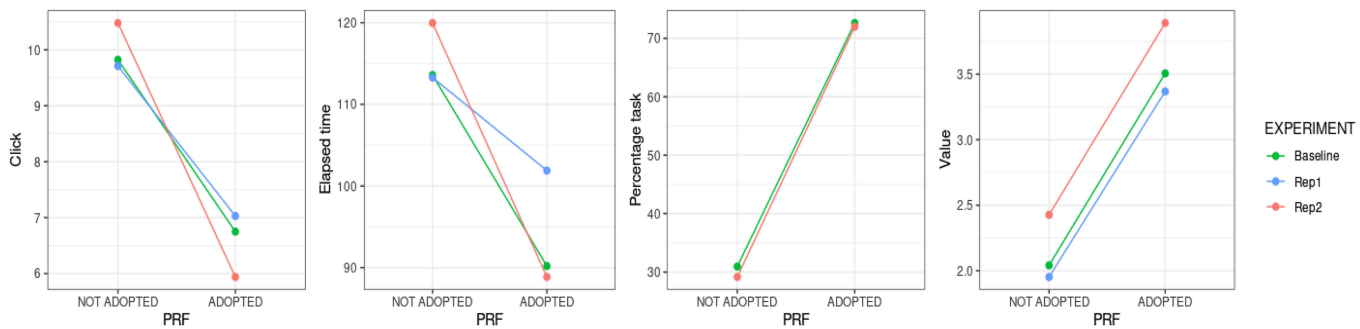

Fig. 4. Profile plot for PRF.

TABLE 11
Replication Effect Sizes for PRF

Response variable	Experiment	d	vi	Interpretation
Click	Baseline	-0.2292	0.0240	small
	Rep1	-0.2162	0.0150	small
	Rep2	-0.4136	0.0417	small
Time	Baseline	-0.2391	0.0240	small
	Rep1	-0.1184	0.0150	small
	Rep2	-0.2805	0.0412	small
Percentage	Baseline	0.9119	0.0263	large
	Rep1	-	-	-
	Rep2	0.9385	0.0453	large
Satisfaction	Baseline	1.2363	0.0284	large
	Rep1	1.1333	0.0174	large
	Rep2	1.2307	0.0486	large

usability mechanism considering the interactions and main effects. The full exploratory analysis of the moderator variables is available in Appendix D, available in the online supplemental material.

Main effects: The demographic variables did not meaningfully affect any response variable/usability mechanism, with one exception.

The significant results, both related to age, are reported in Tables D.5 and D.22 of Appendix D, available in the online supplemental material. They suggest that older subjects employ more clicks and take longer to complete the tasks. However, this does not apply to all mechanisms, and there are exceptions to that rule (see Table D.30, where older subjects give different satisfaction values). The interpretation looks plausible, because, as we will see later, a similar consistent and statistically significant pattern shows up when analysing the interactions. Nevertheless, we cannot rule out the possibility of these significant results being due to type-I errors, and they should, therefore, be regarded with due caution.

The exception is Table D.16 of Appendix D, available in the online supplemental material. It shows that females complete more tasks (4.2%) than males when using the PFB mechanism. This could be a genuine effect because other, albeit non-significant, tables point in a similar direction: Table D.4 suggests that females take less time, and Table D.10 suggests they achieve higher satisfaction levels. There is scant, albeit potentially convincing, evidence to support this.

Interaction effects: The key outcomes of the demographic analysis occur when considering the interactions. Due to the number of significant results, we have prepared a summary table (Table 13). Note that some tables with significant results, e.g., Table D.31 and D.32, are not discussed here because they do not offer any new insights. These tables have been moved to Appendix D, available in the online supplemental material.

Table 13 indicates that age and purchasing experience have a significant impact on the response variables. In the case of age, the pattern is relatively stable and suggests that:

1. Older subjects employ proportionally more clicks and a longer elapsed time when the usability mechanisms are not adopted.
2. Older subjects complete proportionally more tasks and with a proportionally higher satisfaction when the usability mechanisms are enabled.

Points 1 and 2 above hold in most cases, but there are some exceptions. Table 13 again shows that these results do not apply for all mechanisms. However, the pattern is appealing and convincing.

Table 13 also shows that, for two out of the three mechanisms, subjects are more satisfied when the usability mechanisms are enabled. Interestingly, the pattern is non-linear. The higher satisfaction levels show up at both ends of the scale (subjects that are inexperienced or very experienced online shoppers). Values are lower at the centre of the scale (subjects that rarely/sometimes buy online). We have no explanation for this, but, clearly, subjects with some experience are less affected by the usability mechanisms than first-time shoppers or subjects who have been using shopping portals for a long time.

6 Validity Threats

In this section, we discuss the threats with respect to the statistical conclusion validity, internal validity, external validity and construct validity.

6.1 Conclusion Validity

The threats at the level of the family of experiments that are related to statistical conclusion validity appear when replicating the experiment and combining the results. We relied upon parametric statistical tests (i.e., LMM [72]) to analyse the data of our family of experiments. We ensured the robustness of the results that we provided by meta-analysing the data with the one-stage IPD model and an extra factor that accounts for the difference between results across experiments [19], [20]. In order to ensure the transparency of the results, the original data and statistical analyses carried out are provided in the supplementary material, available online. All the supplementary material is also available at figshare[2].

Another threat is related to the reliability of the treatment implementation. The tasks that we use for the mechanisms only account for the scenario within the experimental evaluation. We expect to do things in a particular way, but we know that there may be deviations in the performance of a specific task. For example, the subjects may use tricks (e.g., zoom, full-screen mode, large screens, copy & paste from browser) instead of the Preferences implemented in the application to complete the task. Besides, some subjects may not even manage to perform the tasks even with the adopted mechanism. However, we used the contact information to gather feedback from subjects and follow up suspect behaviour.

6.2 Internal Validity

To increase internal validity, we did not inform participants about the tasks that they were to perform beforehand. In the following, we discuss the five identified threats to internal validity and the actions taken to mitigate these threats.

The first two threats are related to technological expertise and the order of task performance. With regard to

2. https://doi.org/10.6084/m9.figshare.13148117

TABLE 12
Summary of the Experiment Results

Usability Mechanism	Efficiency		Effectiveness	Satisfaction
	Clicks	Time	Percentage	Value
ABR	(-) 23% *	(-) 24% *	25% *	39% *
PFB	(-) 18% *	(-) 7.1%	(-) 1.8%	9.8% *
PRF	(-) 32% *	(-) 17% *	137% *	70% *

TABLE 13
Summary of Interaction Effects (the Referenced Tables are Available in Appendix D, available in the online supplemental material)

	Usability mechanism		
Response variable	Abort Operation	Progress Feedback	Preference
CLICK	Table D.2. Subjects performing the task without the mechanism employ more clicks, and the number of clicks increases with age.	-	-
ELAPSED_TIME	Table D.7. Subjects performing the task without the mechanism take longer, and the elapsed time increases with age.	-	-
PERCENTAGETASK	-	-	Table D.28. Subjects performing the task with the mechanism complete more tasks, and the percentage increases with have.purchased.by. yourself response variable.
VALUE	Table D.12. Subjects performing the task with the mechanism are more satisfied. Satisfaction peaks at the ends of the have. purchased.by.yourself response variable scale(NEVER and ALWAYS). The lowest value appears in the centre (SOMETIMES).	Table D.20. Subjects performing the task with the mechanism are more satisfied, and satisfaction increases with age.	Table D.33. Subjects performing the task with the mechanism are more satisfied. Satisfaction peaks at the ends of the have.purchased.by.yourself response variable scale (NEVER and ALWAYS/ALMOST ALWAYS). The lowest values appear in the centre (RARELY/SOMETIMES).

technological expertise, although all the experiment participants are novices with regard to their level of experience with this type of experiments, they do not all have the same expertise regarding the activity to be performed. Besides the familiarity questionnaire revealed that a large percentage of subjects were familiar with the use of web pages, although the online shopping rate among subjects is low. As far as the order of task performance is concerned, there could be bias caused by the learning effect, as the tasks associated with each mechanism are performed sequentially.

To mitigate the above two threats, we randomly assigned subjects to balanced groups. This randomization procedure is an experimental guarantee [73], as interferences may or may not occur irrespective of their impact. It is worthwhile making the effort to randomize experiments to offset any potential bias.

A third internal validity threat is low user experience, where there is a risk of users not making the effort it takes to understand the instructions, comprehend the procedure, etc. We overcome this threat by introducing the order as a design factor.

The fourth threat is related to the fact that subjects perform the usability test remotely, and it is not possible to interact with participants in real time. As a result, the participants could perform the experiment more than once, do things wrong or drop out of the experiment because they misunderstand the task instructions and do not have the chance to ask what to do when they are unsure. To try to mitigate this 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.

Finally, there is a fifth internal validity threat related to motivation. Each participant will, foreseeably, react differently to the experiment, and subjects may perform poorly, especially if they are alternating experiment performance with other activities. This threat cannot be mitigated. Nonetheless, we interviewed subjects at random to find out if they suffered from fatigue, boredom or similar. The responses should be considered during the analysis and interpretation of the results to reduce the impact of this threat.

6.3 External Validity

We identified two threats to external validity. The first threat is that experimental results cannot be generalized to all users. To prevent any potential bias caused by familiarity with the technology, the participants selected to participate in all three experiments are not computer scientists. Nevertheless, all the participants are regular Internet users. Additionally, the subjects are members of a sizeable user population group that tends not to use online shopping web applications. However, we can gather quite reliable

empirical evidence about the impact of the usability mechanisms analysed at lay user level.

Another probable threat is the generalization to applications from other domains. This threat could be dealt with by executing the experiment in other application domains.

6.4 Construct Validity

This threat is concerned with generalizing the results of the experiment to the concept or theory underlying the experiment. This is mainly related to how we measure the capability of a subject to perform a task. We chose tasks that were as representative as possible of realistic environments. Also, the measurements that we chose—clicks, elapsed time, percentage task completion, satisfaction questionnaire—are based on standard metrics and scales [52], [53], [56]. Clearly, the subjects who do not have access to the experimenters may misunderstand the task specification.

The use of questionnaires may have biased the results of the satisfaction response variable. However, this approach has been used in other studies to measure satisfaction, and we do not see any other more reliable mechanism of this measure.

Finally, different experimenters (baseline experiment/ Replication 2 vs Replication 1) may affect the results. The general instructions were given by different experimenters (even by the instructors of the course within which the experiment was executed). However, as Shull *et al.* [74] pointed out, the independence of the replicators from the original experimenters boosts confidence in the original results not being the result of experimental bias.

7 Conclusion and Future Work

Having run a set of three experiments (the baseline experiment, and two replications), we conducted a meta-analysis applied to the family of experiments. To do this, we used the linear regression model at family level with reference to the baseline experiment.

The aggregate data of the family of experiments again suggest that the adoption of the three mechanisms improves system usability and does not undermine user performance. In particular, a decisive improvement is not always observed in the case of the efficiency response variable, although the adoption of the usability mechanism never detracts from user efficiency. In the case of Abort Operation, the improvement in user efficiency is conclusive. In the case of Progress Feedback, the difference between adoption and non-adoption of the mechanism is appreciable only for number of clicks. Finally, for Preferences, the difference between mechanism adoption and non-adoption is conclusive for number of clicks and quite large for time reduction.

In the case of the effectiveness response variable for Abort Operation and Preferences, the difference between the adoption and non-adoption of each mechanism improves system usability conclusively. There is no significant improvement in the case of Progress Feedback.

For the satisfaction response variable, there is a significant difference between adoption and non-adoption for both Abort Operation and Preferences, which should therefore be considered, whereas the improvement is almost non-existent for the Progress Feedback mechanism.

Based on the meta-analysis results, we can therefore conclude that the adoption and non-adoption of the Abort Operation and Preferences mechanisms appear to have a major impact on system usability with respect to both user efficiency, effectiveness and satisfaction. Progress Feedback appears to affect the response variables less, and, ultimately, has a negligible impact on system usability.

With our family of experiments, we have verified that most of the values are statistically significant. The effect of the values that are not significant is so small as to render the pursuit of further research worthless, because, thanks to the high number of experimental subjects, the evidence that we have gathered from the replications would overrule the results of other experiments [42]. On this ground, due to the stability of the findings after applying data aggregation, we should make systematic changes to the experimental design (domain, tasks, usability mechanism, quality attributes, etc.) in search of new findings.

In sum, the family of experiments endorses the result of the baseline experiment and further ratifies the finding that Progress Feedback does not lead to appreciable improvements in user performance (at least with respect to the task implemented within the study domain). Additionally, we confirmed the preliminary findings of the baseline experiment [14] on the order of priority for the adoption of the usability mechanisms:

1. 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.
2. Abort Operation improves efficiency, effectiveness and satisfaction and should be considered in second place because it is not as cheap to implement.
3. 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.

These adoption priorities should be interpreted in the context that we defined (online shopping). In the particular case of PFB, we believe that this mechanism implemented in another domain may have a different effect. The role of PFB 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 because: a) the PFB task is a simple search operation, b) the Internet connection is fast, and c) the server overhead was practically negligible. We believe that larger improvements could be observed in contexts involving the performance of high latency tasks.

This research is another step forward in the empirical analysis of usability from the user viewpoint. Additionally, it pinpoints research gaps concerning the impact of other HCI recommendations (usability mechanisms) and their possible combinations on software systems, as well as the effect of including usability mechanisms on the different software development process activities. Families of experiments with their respective replication packages depend on the analysis of these impacts. This, together with the statistically significant findings of our family of

experiments, are reason enough to further pursue experimentation and further explore the following promising lines:

- Modifying the experimental design by altering the instrumentation for application in a different domain from online shopping.
- Redesign the tasks to create more complex scenarios.
- Researching the combinations of usability mechanisms addressed in this or other papers on the same quality characteristics from the viewpoint of users.
- Researching the usability mechanisms, and their combinations, on other quality characteristics.
- Research the impact of implementing usability mechanisms and their combinations at architecture, design and coding level.

Finally, families of experiments are becoming increasingly important in SE, generating evidence underpinning the evolution of knowledge on the impact of the usability recommendations implemented through usability mechanisms in web environments.

References

[1] J. Johnson and A. Henderson, "Usability of interactive systems: It will get worse before it gets better," *J. Usability Stud.*, vol. 7, no. 3, pp. 88–93, 2012.

[2] L. B. Ammar, A. Trabelsi, and A. Mahfoudhi, "A model-driven approach for usability engineering of interactive systems," *Softw. Qual. J.*, vol. 24, no. 2, pp. 301–335, 2016.

[3] B. Shneiderman, C. Plaisant, M. Cohen, S. Jacobs, N. Elmqvist, and N. Diakopoulos, *Designing the User Interface: Strategies for Effective Human-Computer Interaction*, 6th ed. London, U.K.: Pearson, 2016.

[4] ISO/IEC-25010, "Systems and software engineering - systems and software quality requirements and evaluation (SQuaRE) - System and software quality models," International Organization for Standardization, Geneva, Switzerland, 2011.

[5] F. D. Rodríguez, S. T. Acuña, and N. Juristo, "Design and programming patterns for implementing usability functionalities in web applications," *J. Syst. Softw.*, vol. 105, no. C, pp. 107–124, Jul. 2015, doi: 10.1016/j.jss.2015.04.023.

[6] R. M. Baecker, *Readings in Human-Computer Interaction: TOWARD the Year.*, San Mateo, CA, USA: Morgan Kaufmann, 2000.

[7] A. M. Moreno, A. Seffah, R. Capilla, and M. I. Sánchez-Segura, "HCI practices for building usable software," *Computational*, vol. 46, no. 4, pp. 100–102, 2013, doi: 10.1109/MC.2013.133.

[8] J. Tidwell, *Designing Interfaces. Patterns for Effective Interaction Design*, 2nd ed. Newton, MA, USA: O'Reilly Media, 2010.

[9] D. Hix and H. R. Hartson, *Developing User Interfaces: Ensuring Usability Through Product & Process*. Hoboken, NJ, USA: Wiley, 1993.

[10] X. Ferré, N. Juristo, H. Windl, and L. Constantine, "Usability basics for software developers," *IEEE Softw.*, vol. 18, no. 1, pp. 22–29, Jan./Feb. 2001.

[11] M. V. Welie and H. Trætteberg, "Interaction patterns in user interfaces," in *Proc. 7th Pattern Lang. Programs*, 2000, pp. 13–16.

[12] N. Juristo, A. M. Moreno, and M.-I. Sanchez-Segura, "Analysing the impact of usability on software design," *J. Syst. Softw.*, vol. 80, no. 9, pp. 1506–1516, Sep. 2007, doi: 10.1016/j.jss.2007.01.006.

[13] F. D. Rodríguez, S. T. Acuña, and N. Juristo, "Reusable solutions for implementing usability functionalities," *Int. J. Softw. Eng. Knowl. Eng.*, vol. 25, no. 04, pp. 727–755, 2015, doi: 10.1142/S0218194015500084.

[14] J. M. Ferreira *et al.*, "Impact of usability mechanisms: An experiment on efficiency, effectiveness and user satisfaction," *Inf. Softw. Technol.*, vol. 117, pp. 106195–106195, Jan. 2020, doi: 10.1016/j.infsof.2019.106195.

[15] A. Santos, S. Vegas, F. Uyaguari, O. Dieste, B. Turhan, and N. Juristo, "Increasing validity through replication: An illustrative TDD case," *Softw. Qual. J.*, vol. 28, pp. 1–25, 2020.

[16] M. Solari, S. Vegas, and N. Juristo, "Content and structure of laboratory packages for software engineering experiments," *Inf. Softw. Technol.*, vol. 97, pp. 64–79, 2018, doi: 10.1016/j.infsof.2017.12.016.

[17] A. Santos, O. Gómez, and N. Juristo, "Analyzing families of experiments in SE: A systematic mapping study," *IEEE Trans. Softw. Eng.*, vol. 46, no. 5, pp. 566–583, May 2020.

[18] R. Capilla, R. Kazman, C. Romera, and C. Carrillo, "Usability implications in software architecture: The case study of a mobile app," *Softw. Pract. Exp.*, vol. 50, no. 12, pp. 2145–2168, 2020, doi: 10.1002/spe.2883.

[19] A. Santos, S. Vegas, M. Oivo, and N. Juristo, "A procedure and guidelines for analyzing groups of software engineering replications," *IEEE Trans. Softw. Eng.*, vol. 47, no. 9, pp. 1742–1763, Sep. 2021.

[20] J. I. P. Navarrete *et al.*, "Evaluating model-driven development claims with respect to quality: A family of experiments," *IEEE Trans. Softw. Eng.*, vol. 47, no. 1, pp. 130–145, Jan. 2011.

[21] T. Komiyama, S. Fukuzumi, M. Azuma, H. Washizaki, and N. Tsuda, "Usability of software–Intensive systems from developers' point of view," in *Proc. Hum.-Comput. Interaction. Des. User Experience*, Copenhagen, Denmark, 2020, pp. 450–463.

[22] B. A. Seffah and E. Metzker, "The obstacles and myths of usability and software engineering," *Commun. ACM*, vol. 47, no. 12, pp. 70–76, 2004, doi: 10.1145/1035134.1035136.

[23] M. C. S. Torrente, A. B. M. Prieto, D. A. Gutiérrez, and M. E. A. de Sagastegui, "Sirius: A heuristic-based framework for measuring web usability adapted to the type of website," *J. Syst. Softw.*, vol. 86, no. 3, pp. 649–663, Mar. 2013, doi: 10.1016/J.JSS.2012.10.049.

[24] L. Punchoojit and N. Hongwarittorrn, "Usability studies on mobile user interface design patterns: A systematic literature review," *Adv. Hum.-Comput. Interact.*, vol. 2017, Nov. 2017, Art. no. 6787504, doi: 10.1155/2017/6787504.

[25] N. Juristo, A. M. Moreno, and M. I. Sanchez-Segura, "Guidelines for eliciting usability functionalities," *IEEE Trans. Softw. Eng.*, vol. 33, no. 11, pp. 744–758, Nov. 2007.

[26] D. Gupta, A. K. Ahlawat, A. Sharma, and J. J. P. C. Rodrigues, "Feature selection and evaluation for software usability model using modified moth-flame optimization," *Computational*, vol. 102, no. 6, pp. 1503–1520, Jun. 2020, doi: 10.1007/s00607-020-00809-6.

[27] K. Sagar, D. Gupta, and A. K. Sangaiah, "Manual versus automated qualitative usability assessment of interactive systems," *Concurr. Comput. Pract. Exp.*, vol. 33, no. 12, pp. 1–13, Nov. 2018, doi: 10.1002/cpe.5091.

[28] D. Gupta and A. K. Ahlawat, "Usability feature selection via MBBAT: A novel approach," *J. Comput. Sci.*, vol. 23, pp. 195–203, 2017, doi: 10.1016/j.jocs.2017.06.005.

[29] D. Gupta and A. K. Ahlawat, "Taxonomy of GUM and usability prediction using GUM multistage fuzzy expert system," *Int. Arab J. Inf. Technol.*, vol. 16, no. 3, pp. 357–363, 2019.

[30] A. I. Martins, A. Queirós, and N. P. Rocha, "Validation of a usability assessment instrument according to the evaluators' perspective about the users' performance," *Universal Access Inf. Soc.*, vol. 19, no. 3, pp. 515–525, Aug. 2020, doi: 10.1007/s10209-019-00659-w.

[31] A. Fernandez, E. Insfran, and S. Abrahão, "Usability evaluation methods for the web: A systematic mapping study," *Inf. Softw. Technol.*, vol. 53, no. 8, pp. 789–817, Aug. 2011, doi: 10.1016/J.INFSOF.2011.02.007.

[32] X. Fang and C. W. Holsapple, "An empirical study of web site navigation structures' impacts on web site usability," *Decis. Support Syst.*, vol. 43, no. 2, pp. 476–491, Mar. 2007, doi: 10.1016/j.dss.2006.11.004.

[33] M. Y. Ivory and M. A. Hearst, *An Empirical Foundation for Automated Web Interface Evaluation*. Berkeley, CA, USA: Univ. California, 2001.

[34] A. Karani, H. Thanki, and S. Achuthan, "Impact of university website usability on satisfaction: A structural equation modelling approach," *Manage. Labour Stud.*, vol. 46, no. 2, pp. 119–138, May 2021, doi: 10.1177/0258042X21989924.

[35] A. Idri, "Usability evaluation of mobile applications using ISO 9241 and ISO 25062 standards," *SpringerPlus*, vol. 5, pp. 548–548, 2016, doi: 10.1186/s40064-016-2171-z.

[36] H. H. Larsen, A. N. Scheel, T. Bogers, and B. Larsen, "Hands-free but not eyes-free: A usability evaluation of SIRI while driving," in *Proc. Conf. Hum. Inf. Interact. Retrieval*, 2020, pp. 63–72, doi: 10.1145/3343413.3377962.

[37] S. Weber, M. Coblenz, B. Myers, J. Aldrich, and J. Sunshine, "Empirical studies on the security and usability impact of immutability," in *Proc. IEEE Cybersecurity Develop. Conf.*, 2017, pp. 50–53, doi: 10.1109/SecDev.2017.21.

[38] S. Tiwari and A. Gupta, "Investigating comprehension and learnability aspects of use cases for software specification problems," *Inf. Softw. Technol.*, vol. 91, pp. 22–43, 2017, doi: 10.1016/j.infsof.2017.06.003.
[39] M. Piccioni, C. A. Furia, and B. Meyer, "An empirical study of API usability," in *Proc. ACM / IEEE Int. Symp. Empirical Softw. Eng. Meas.*, 2013, pp. 5–14, doi: 10.1109/ESEM.2013.14.
[40] O. S. Gómez, N. Juristo, and S. Vegas, "Understanding replication of experiments in software engineering: A classification," *Inf. Softw. Technol.*, vol. 56, no. 8, pp. 1033–1048, Aug. 2014, doi: 10.1016/j.infsof. 2014.04.004.
[41] F. G. de Oliveira Neto, R. Torkar, R. Feldt, L. Gren, C. A. Furia, and Z. Huang, "Evolution of statistical analysis in empirical software engineering research: Current state and steps forward," *J. Syst. Softw.*, vol. 156, pp. 246–267, Oct. 2019, doi: 10.1016/j.jss.2019.07.002.
[42] A. Santos, S. Vegas, M. Oivo, and N. Juristo, "Comparing the results of replications in software engineering," *Empir. Softw. Eng.*, vol. 26, no. 2, Feb. 2021, Art. no. 13, doi: 10.1007/s10664-020-09907-7.
[43] B. Kitchenham, L. Madeyski, and P. Brereton, "Meta-analysis for families of experiments in software engineering: A systematic review and reproducibility and validity assessment," *Empir. Softw. Eng.*, vol. 25, no. 1, pp. 353–401, Jan. 2020, doi: 10.1007/s10664-019-09747-0.
[44] C. E. Anchundia and E. R. C. Fonseca, "Resources for reproducibility of experiments in empirical software engineering: Topics derived from a secondary study," *IEEE Access*, vol. 8, pp. 8992–9004, 2020.
[45] M. Borenstein, L. V. Hedges, J. P. T. Higgins, and H. R. Rothstein, *Introduction to Meta-Analysis*. Hoboken, NJ, USA: Wiley, 2009, doi: 10.1002/9780470743386.
[46] S. Rueda, J. I. Panach, and D. Distante, "Requirements elicitation methods based on interviews in comparison: A family of experiments," *Inf. Softw. Technol.*, vol. 126, 2020, Art. no. 106361, doi: 10.1016/j.infsof.2020.106361.
[47] F. Ricca, M. Torchiano, M. Leotta, A. Tiso, G. Guerrini, and G. Reggio, "On the impact of state-based model-driven development on maintainability: A family of experiments using unimod," *Empir. Softw. Eng.*, vol. 23, no. 3, pp. 1743–1790, Jun. 2018, doi: 10.1007/s10664-017-9563-8.
[48] J. I. Panach, N. Condori-Fernández, A. Baars, T. Vos, I. Romeu, and Ó. Pastor, "Towards an experimental framework for measuring usability of model-driven tools," in *Proc. IFIP Conf. Hum.-Comput. Interact.*, 2011, pp. 640–643.
[49] N. Condori-Fernández, J. I. Panach, A. I. Baars, T. Vos, and Ó. Pastor, "An empirical approach for evaluating the usability of model-driven tools," *Sci. Comput. Program.*, vol. 78, no. 11, pp. 2245–2258, Nov. 2013, doi: 10.1016/j.scico.2012.07.017.
[50] A. Santos *et al.*, "A family of experiments on test-driven development," *Empir. Softw. Eng.*, vol. 26, no. 3, Mar. 2021, Art. no. 42, doi: 10.1007/s10664-020-09895-8.
[51] A. Fernandez, S. Abrahão, and E. Insfran, "Empirical validation of a usability inspection method for model-driven web development," *J. Syst. Softw.*, vol. 86, no. 1, pp. 161–186, Jan. 2013, doi: 10.1016/J.JSS.2012.07.043.
[52] ISO/IEC, "ISO/IEC 9126-4 software engineering -product quality-Part4: Quality in use metrics," International Organization for Standardization, Geneva, Switzerland, 2004.
[53] K. Hornbæk, "Current practice in measuring usability: Challenges to usability studies and research," *Int. J. Hum. Comput. Stud.*, vol. 64, no. 2, pp. 79–102, 2006, doi: 10.1016/j.ijhcs.2005.06.002.
[54] A. Seffah, M. Donyaee, R. B. Kline, and H. K. Padda, "Usability measurement and metrics: A consolidated model," *Softw. Qual. J.*, vol. 14, no. 2, pp. 159–178, 2006, doi: 10.1007/s11219-006-7600-8.
[55] ISO 9241-11, "Ergonomic requirements for office work with visual display terminals (VDTs)–Part II guidance on usability," International Organization for Standardization, 1998.
[56] J. Sauro and E. Kindlund, "A method to standardize usability metrics into a single score," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2005, pp. 401–409, doi: 10.1145/1054972.1055028.
[57] G. Charness, U. Gneezy, and M. A. Kuhn, "Experimental methods: Between-subject and within-subject design," *J. Econ. Behav. Organ.*, vol. 81, no. 1, pp. 1–8, 2012, doi: 10.1016/j.jebo.2011.08.009.
[58] G. Taguchi, S. Konishi, and S. Konishi, *Taguchi Methods, Orthogonal Arrays and Linear Graphs*. Dearborn, MI, USA: American Supplier Institute Dearborn, 1987.
[59] "QuickStore (experimental application)," Mar. 2016. Accessed: Oct. 03, 2020. [Online]. Available: http://webadm.senado.gov.py/tesisweb/
[60] J. M. Ferreira and S. T. Acuña, "A software application for collecting usability empirical data about user efficiency, effectiveness and satisfaction," in *Proc. 12th Iberoamerican Conf. Softw. Eng. Knowl. Eng.*, 2017, Art. no. 11.
[61] J. C. Carver, "Towards reporting guidelines for experimental replications: A proposal," in *Proc. 1st Int. Workshop Replication Empirical Softw. Eng. Res.*, 2010, Art. no. 4.
[62] J. L. Hintze and R. D. Nelson, "Violin plots: A box plot-density trace synergism," *Amer. Statist.*, vol. 52, no. 2, pp. 181–184, 1998, doi: 10.1080/00031305.1998.10480559.
[63] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Ann. Math. Statist.*, vol. 18, pp. 50–60, 1947.
[64] W. Jacoby, *Statistical Graphics for Visualizing Multivariate Data*. Thousand Oaks, CA, USA: Sage, 2021, doi: 10.4135/9781412985970.
[65] T. P. A. Debray *et al.*, "Get real in individual participant data (IPD) meta-analysis: A review of the methodology," *Res. Synth. Methods*, vol. 6, no. 4, pp. 293–309, Dec. 2015, doi: 10.1002/jrsm.1160.
[66] A. Whitehead, *Meta-Analysis of Controlled Clinical Trials*. Hoboken, NJ, USA: Wiley, 2002.
[67] T. Lumley, P. Diehr, S. Emerson, and L. Chen, "The importance of the normality assumption in large public health data sets," *Annu. Rev. Public Health*, vol. 23, no. 1, pp. 151–169, 2002.
[68] A. J. Vickers, "Parametric versus non-parametric statistics in the analysis of randomized trials with non-normally distributed data," *BMC Med. Res. Methodol.*, vol. 5, no. 1, 2005, Art. no. 35.
[69] V. Carey and W. You-Gan, "Mixed-effect models in S and S-Plus," *J. Am. Statist. Assoc.*, vol. 96, no. 455, 2001, Art. no. 1135.
[70] C. Jacob, "A power primer," *Tutor. Quant. Methods Psychol.*, vol. 112, pp. 155–159, 2007, doi: 10.1037/0033-2909.112.1.155.
[71] M. Torchiano, "Effsize - A package for efficient effect size computation," to be published, doi: 10.5281/zenodo.1480624.
[72] B. T. West, K. B. Welch, and A. T. Galecki, *Linear Mixed Models: A Practical Guide Using Statistical Software*. Boca Raton, FL, USA: CRC Press, 2014.
[73] M. Kang, B. G. Ragan, and J.-H. Park, "Issues in outcomes research: An overview of randomization techniques for clinical trials," *J. Athletic Training*, vol. 43, no. 2, pp. 215–221, 2008, doi: 10.4085/1062-6050-43.2.215.
[74] F. J. Shull, J. C. Carver, S. Vegas, and N. Juristo, "The role of replications in empirical software engineering," *Empir. Softw. Eng.*, vol. 13, no. 2, pp. 211–218, 2008, doi: 10.1007/s10664-008-9060-1.

Juan M. Ferreira received the MSc degree in software and systems from the Technical University of Madrid in 2011 and the MSc degree in ICT majoring in software engineering from the Universidad Nacional de Asunción in 2018. He is currently an assistant professor of software engineering with the Universidad Nacional de Asunción. His research interests include usability and experimental software engineering.

Francy D. Rodríguez received the PhD degree from the Technical University of Madrid in 2015. She is currently an associate professor of computer science with the University of Ávila. Her research interests include software development, design and programming patterns, and software usability.

Adrián Santos received the MSc degree in software and systems and the MSc degree in software project management from the Technical University of Madrid, Spain, the MSc degree in IT auditing, security and government from the Autonomous University of Madrid, Spain, and the PhD degree in software engineering from the University of Oulu, Finland. He is currently a software engineer in industry.

Oscar Dieste received the BS and MS degrees in computing from the Universidad da Coruña and the PhD degree from the Universidad de Castilla La Mancha. He is currently a researcher with the UPM's School of Computer Engineering. He was previously with the University of Colorado, Colorado Springs, as a fulbright scholar, Universidad Complutense de Madrid, and Universidad Alfonso X el Sabio. His research interests include empirical software engineering and requirements engineering.

Silvia T. Acuña received the PhD degree from the Technical University of Madrid in 2002. She is currently an associate professor of software engineering with the Computer Science Department, Autonomous University of Madrid. She has coauthored *A Software Process Model Handbook for Incorporating People's Capabilities* (Springer, 2005), and edited *Software Process Modeling* (Springer, 2005) and *New Trends in Software Process Modeling* (World Scientific, 2006). Her research interests include experimental software engineering, software usability, software process modeling, and software team building. She is deputy conference co-chair of the ICSE 2021 organizing committee. She is a member of IEEE Computer Society and a member of ACM.

Natalia Juristo received the PhD degree from the Technical University of Madrid (UPM) in 1991. She is currently a full professor of software engineering with UPM. Her research interests include experimental software engineering, requirements, and testing. She was the recipient of a Finland Distinguished Professor Program (FiDiPro) professorship, starting in January 2013.