



Heuristic Evaluation and Usability Testing as Complementary Methods: A Case Study

Braulio Murillo^{1,2(✉)}, Jose Pow Sang¹, and Freddy Paz¹

¹ Pontificia Universidad Católica del Perú, Lima 32, Peru
bmurillov@pucp.edu.pe

² DHL Express Perú, Callao 01, Peru

Abstract. Given the relevance that is taking the usability evaluation of software systems and especially web-based systems, the present work seeks to provide new evidence in the results obtained by applying combined techniques in usability evaluations. In this work, the process to schedule shipments using a web application was evaluated. First, a heuristic evaluation was performed and then a usability test with user. However, for this specific purpose, both tests will be applied independently to evaluate the results obtained. The evaluations were developed in an academic context.

Keywords: Human-computer interaction · Heuristic evaluation
Usability evaluation · Usability study · Transactional web applications

1 Introduction

Usability is a quality attribute that measures how easy the user interface is to use. It also includes methods to improve ease of use during the software design process [1]. Nowadays on the web, usability is a necessary condition for survival. If a website is difficult to use, people will stop using it. If the page does not clearly state what a company offers and what users can do on the site, people will stop using it. If users get lost on a website, they will stop using it. If the information on a website is difficult to read or does not answer the key questions of users, they will stop using it [2].

The first e-commerce law is that if users cannot find the product, they cannot buy it either [1]. In this paper, we will evaluate the web application of a logistic company using heuristic evaluation and a usability test with users. The results obtained in both tests will be shown to compare them and provide some conclusions.

2 Related Work

There are two types of methods to perform a usability assessment: inspection methods and test methods. The difference between them lies in the person who applies them. In the first case the inspectors perform and in the second the users participate [3].

In the present work, we make the evaluation of usability to a web application using both methods. As an inspection method, a heuristic evaluation is executed. Heuristic

evaluation is a well-known inspection technique that is widely used by usability specialists. It was developed by Nielsen as an alternative to user testing [4].

On the other hand, as a test method the usability test with users is executed. The main purpose of this method is the identification of usability problems during the interaction of users and the system [2].

In previous works, [2, 5], it is argued that applying heuristic and user evaluations as complementary studies provide advantages in the evaluation of usability. The heuristic evaluation and the usability test with users are executed independently. In this way, we seek to avoid any bias in the application of the tests.

The present work has been developed under an academic context. All the participants have developed the tests with professionalism and ethical values.

3 Research Design

In order to test usability in a web application, two methods were used: heuristic evaluations and usability test users. The objective of this test and the selection of the web application were academics.

3.1 Description of the Web Site

The evaluated web application belongs to a logistics company. This application allows customers to make their schedules of shipments and package pickups, manage contacts and track shipments. As part of the evaluation, only the management of contacts and the scheduling of shipments have been considered.

3.2 Study Design

The purpose of this paper was to compare the results of both, heuristic evaluations and user usability tests based on a web transactional system.

This work was developed in two moments. First a heuristic evaluation was carried out and then a user usability test was developed. Both of them were executed independently.

4 Heuristic Evaluation

4.1 Participants

The heuristic evaluation was performed using the Nielsen's methodology analyzing the ten usability principles "heuristics". The evaluation was performed by three evaluators: one computer engineer, one master of science and one doctoral student.

4.2 Phases

This section describes the steps used to perform the heuristic evaluation. These are described below.

First phase: Each participant carried out the evaluation of the software independently. The results were recorded in their respective reports.

Second phase: A facilitator, arranged a meeting where the evaluators were able to unify the results obtained by briefly explaining the problems they found. A clean and unified listing of the problems encountered was obtained.

Third phase: Each evaluator independently rated the severity and frequency of each problem of the unified listing. With the values of severity and frequency was calculated the criticality: $\text{criticality} = \text{severity} + \text{frequency}$.

Fourth phase: A facilitator, calculated the averages and standard deviations of the three previously calculated values: severity, frequency and criticality of each problem. Based on the results, was established a ranking of the found problems.

The severity was evaluated according to the rating proposed by Nielsen [6], in which 0 means “I don’t agree that this is a usability problem at all” and 4 means “Usability catastrophe: imperative to fix this before product can be released” (see Table 1).

Table 1. Severity ratings

Note	Severity
0	I don’t agree that this is a usability problem at all
1	Cosmetic problem: the problem will be fixed unless there is extra time available in the project
2	Minor problem: fixing the problem is of low priority
3	Major problem: it is important to fix the problem, high priority must be established
4	Usability catastrophe: imperative to fix this before product can be released

The frequency was evaluated according the rating of the Table 2.

Table 2. Frequency ratings

Note	Frequency
0	<1%
1	1–10%
2	11–50%
3	51–90%
4	>90%

4.3 Data Analysis and Results

A total of twenty one usability problems were identified, which were categorized by the participants who performed the heuristic evaluation. The heuristics “Visibility of system status”, “Recognition rather than recall”, and “Help users recognize, diagnose, and recover from errors” were not found non-compliance. In Table 3, it can be seen the times that each unfulfilled heuristic.

Table 3. Unfulfilled heuristics

ID	Heuristic	Problems that non-compliance the heuristics	Number the problems that non-compliance the heuristics
N2	Visibility of system status	P2, P3, P4, P11, P13, P15, P17, P18	8
N3	Match between system and the real world	P10, P14	2
N4	User control and freedom	P6, P7, P8, P18, P19, P20, P21	7
N5	Consistency and standards	P1, P9, P10, P12, P16	5
N7	Flexibility and efficiency of use	P21	1
N8	Aesthetic and minimalist design	P1, P5, P9	3
N10	Help and documentation	P3, P15, P17, P18	4

Table 4. Ranking of the most severe problems

ID	Problem	Average severity
P10	The add button disappears when a piece with weight 0.25 is entered in a specific type of packaging	3.33
P4	The options appear in English	2.67
P21	It is not possible to choose more than one type of packaging in the same shipment	2.67

Of all the problems encountered, approximately 15% of them have a severity value greater than 2.50, which tend to be greater or catastrophic. In Table 4, the problems with greater severity are shown.

The evaluators considered that the most severe problem is that when changing the weight in a specific type of packaging, the Add button disappeared, which causes the user not to know how to continue with the process if he wants to add another piece to submit.

Other of the severe problems is that many of the options appear in English, even when the application is in Spanish. This causes that the users in some moments do not know what option to use if they do not dominate the English language.

Another of the severe problems is that you cannot choose more than one type of packaging in the scheduling of a shipment. What causes the user to have to perform another programming for another type of packaging.

On the other hand, Table 5 shows the problems with the highest criticality value (greater than 5). It is important to mention that in the evaluation the maximum criticality value was 5.34 so it can be said that they are not drastic problems that affect the functionality of the web application.

Table 5. Ranking of the most critical problems

ID	Problem	Average severity	Average frequency	Average criticality
P4	The options appear in English	2.67	2.67	5.34
P10	The Add button disappears when a piece with weight 0.25 is entered in a specific type of packaging	3.33	1.67	5.00

5 Usability Testing

5.1 Test Purpose

The purpose of the usability test is for the user to identify problems that arise when executing routine tasks in the web application. The tasks that have been defined are based on two main functionalities of the application: contact management and shipment scheduling.

5.2 Test Design

The tasks were based on two of the most important processes: contact management and shipment scheduling.

As part of the contact management, the tasks focused on the creation of a new contact. In the scheduling of shipments the tasks were focused from the selection of the origin and destination until the data of the package and type of service as well as the delivery times and additional services.

5.3 Participants

Since this is an academic work and in accordance with previous research [7], many participants are not needed to detect many usability issues [9]. For this reason, three professionals from the area of computer science participate in this evaluation. One has a bachelor’s degree, another master’s degree and the third is a PhD student. Two are male and one female. The age range is from 29 to 35 years.

5.4 Materials

The following materials were developed for the Usability Test:

Confidentiality Agreement: It is a document where the participants show that they agree to participate voluntarily in the usability test.

Previous Indications: Indications are given to participants to be aware of all the stages they will develop as part of the evaluations.

Pre-test Questionnaire: It is a questionnaire to know the demographics of the participants and to classify them, in this way it is possible to contextualize their answers.

Post-test Questionnaire: This questionnaire allows obtaining information complementary to the observations made during the execution of the tests.

Task List: This document details the tasks that participants must perform both for the evaluation of contact management and the scheduling of shipments.

For this evaluation, two scenarios were created: In the first scenario, the user must register a new contact in the application, for which he must record the contact's personal data.

In the second scenario, the user must schedule a shipment for the contact he has created in the previous task. In the scheduling, the user must enter the type of service, shipping information among other necessary information to complete a satisfactory schedule.

Task Compliance Observation Sheet: This document is used by the facilitator to enter the details of the tasks performed by the participants, the time used and any incident that was presented during the evaluation.

5.5 Usability Testing Process

The evaluation is done individually, always with the facilitator attentive to any query of the participant.

At the beginning, each participant is presented with the Confidentiality Agreement and the List of previous indications. After each participant signs the agreement, they initiate the evaluation by answering all the pretest questions [10].

Each participant receives the list of tasks and start the tests with the application. The facilitator records the interaction of the participant with the application and takes note of any incident during the evaluation.

At the end of the tests, each participant fills in the post-test questionnaire and finishes the evaluation.

5.6 Data Analysis and Results

Task 1 Results

Hits presented:

- Users were able to select the options indicated.
- There was not difference in performance executing the task among experimented and not experimented users.
- Users could register a new contact in the website.
- Even when users experience complications, they did not have to seek help from the system.

Inconveniences presented:

- Users have difficulty completing all mandatory fields, as these are not visibly marked as such.

Task 2 Results

Hits presented:

- Users were able to select the options indicated, those who did not have much knowledge of the application were already more familiar.

- There was not difference in performance executing the task among experimented and not experimented users.

Inconveniences presented:

- Users have difficulties placing the weights with two decimals, since the system rounds it to one decimal.
- Users seek help from the system, but it is in English.
- Users can not send more than one piece in the shipment, because the “Add” button disappears after adding the weight of the shipment with two decimals. The users show their discomfort because they have to repeat the process for the second piece.

Data Analysis: Observations. Users use very little help from the system. In Task 1, the main inconvenience that users had was that they could not distinguish whether a field was mandatory. Unlike what was found in the heuristic evaluation, this problem identified as P1, was not evaluated as severe or critical. However, in this usability testing this problem became more relevant. In addition, in Task 2, users do not manage to send two packages in a single schedule, this causes them to do the whole process again. In addition, users can not place the weights of the packages with two decimals.

Data Analysis: Post-test Questionnaire. Table 6 shows that the general appreciation of users with respect to the page evaluated is positive. They emphasize the information is useful, easy to understand and easy to find. Additionally, they agree with using the web application again. The two points that received the lowest rating was the fact that they were not able to complete the tasks because there was poor orientation in the web application.

Table 6. Results of post-test questionnaire

Question		Average result	
1	Fulfill tasks	2.33	Hardly
2	Sufficient and complete information	3.00	Neutral
3	Easy to understand information available	3.67	Easy
4	Required information easy to find	3.67	Easy
5	Information found useful	4.00	Useful
6	Web application easy to navigate	3.33	Neutral
7	Orientation on the web application	2.33	Disagreed
8	Satisfaction with the web application	2.67	Neutral
9	Will use the web application again	3.51	Agreed

6 Result Discussion

In the first part of this work, the results of the heuristic evaluation allowed finding 21 problems. Of these, problems P10, P4 and P21 were the most severe. Additionally, P4 and P10 were the most critical.

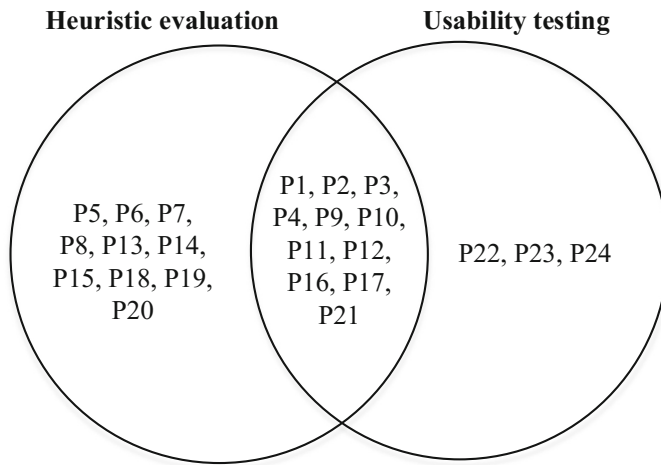


Fig. 1. Identified problems in each stage

On the other hand, in the usability testing some problems obtained in the heuristic evaluation are confirmed. Even some of them in this evaluation are evaluated with greater criticality. Also, in this stage there are new problems that were not identified in the first stage.

In the Fig. 1, the problems found in each evaluation and the problems that were repeated in each stage are summarized.

7 Conclusions

With this work, it is confirmed what was mentioned in a previous one [2], the heuristic evaluation is executed in less time and cost compared to the evaluation of users [8].

In the heuristic evaluation several usability problems were found in the web application. However, there were few problems that really had a high negative impact. Only 15% of problems were considered as very severe and 10% were considered as very critical. In general, under this evaluation it could be said that the web application had a very good result in the evaluation.

However, when the usability testing was performed, a more modest evaluation of the web application was obtained. Additionally, the most critical problems in this evaluation were not considered critical in the heuristic evaluation.

It should be noted that both tests were performed independently, thus there were no biases that could influence the results of the evaluations.

These evaluation methods gave results, since they allowed to make improvements in the web application, achieving an increase in the use of the tool and a better user satisfaction.

In conclusion, it can be affirmed that both tests complement each other, since the results obtained in one are enriched with those of the other test, being able to identify more problems or highlighting problems that an evaluation was not considered critical.

Acknowledgments. This research was carried out thanks to the support of DHL Express Peru, which in its constant search for the excellence of its products and services, is committed to research and innovation of its processes in the IT department. The authors especially thank Hugo Moreno, IT Manager, Adriana Azopardo, Peru Country Manager, Ivan Hay, CIO Central and South America, and Hank Gibson, CIO Americas Region; for their recommendations and their insightful comments to this investigation.

References

1. Nielsen, J.: Usability 101: Introduction to Usability (2012). <https://www.nngroup.com/articles/usability-101-introduction-to-usability/>. Accessed 29 October 2017
2. Murillo, B., Vargas, S., Moquillaza, A., Fernández, L., Paz, F.: Usability Testing as a Complement of Heuristic Evaluation: A Case Study. In: Marcus, A., Wang, W. (eds.) DUXU 2017. LNCS, vol. 10288, pp. 434–444. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58634-2_32
3. Holzinger, A.: Usability engineering methods for software developers. *Commun. ACM* **48**, 71–74 (2005)
4. Paz, F., Asrael Paz, F., Pow-Sang, J.A.: Evaluation of usability heuristics for transactional web sites: a comparative Study. *Adv. Intell. Syst. Comput.* **448**, 1063–1073 (2016)
5. Paz, F., Villanueva, D., Pow-Sang, J.A.: Heuristic evaluation as a complement to usability testing: a case study in web domain. In: 2014 Tenth International Conference on Information Technology: New Generations (ITNG), pp 546–551, April 2015
6. Nielsen, J.: Severity Ratings for Usability Problems (1995). <https://www.nngroup.com/articles/how-to-rate-the-severity-of-usability-problems/>. Accessed 02 November 2017
7. Virzi, R.A.: Refining the test phase of usability evaluation: how many subjects is enough? *Hum. Factors* **34**(4), 457–468 (1992)
8. Jeffries, R., Desurvire, H.: Usability testing vs. heuristic evaluation: was there a contest? *ACM SIGCHI Bull.* **24**(4), 39–41 (1992)
9. Virzi, R.A.: Streamlining the design process: running fewer subjects. In: Proceedings of the Human Factors Society Annual Meeting, vol. 34, no. 4, pp. 291–294. SAGE Publications, Los Angeles, October 1990
10. Rubin, J., Chisnell, D.: Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests. Wiley, Hoboken (2008)