


Enhancing Usability Assessment with a Novel Heuristics-Based Approach Validated in an Actual Business Setting

Afra Pascual Almenara   [Universitat de Lleida | afra.pascual@udl.cat]

Toni Granollers Saltiveri  [Universitat de Lleida | toni.granollers@udl.cat]

Juan Enrique G. Navarro  [Universidad de Castilla-La Mancha | juanenrique.garrido@uclm.es]

Marta Albets Mitjaneta  [Universitat de Lleida | marta.albets@udl.cat]

 *Departamento de Informática y Diseño Digital. Universitat de Lleida, Spain*

Received: 30 January 2024 • **Accepted:** 22 June 2024 • **Published:** 01 July 2024

Abstract

We have long been committed to improving Usability Evaluation and one of the proposals we have been working on the most is the use and improvement of the Heuristic Evaluation (HE) technique. With this in mind, we proposed an improvement which was tested in an experiment. This article describes an experiment carried out in a real business professional context. Fifteen usability experts from a reputable company evaluated eight websites (four supermarket and four bank platforms) using our HE proposal for the first time in real-world scenarios. This experimentation analyzed two main aspects: firstly, whether individual or group evaluations affect the final result, and secondly, whether the heuristic evaluation technique is effective in a real business and professional context. Regarding the Usability Percentage (UP) event, the results indicate that there was little difference between group and individual evaluations. The mean UP for the group was 57.88%, while the mean UP for individuals was 56.66%. The experiment provided sufficient information to suggest a new version of our HE methodology, specifically designed to improve results in real-life contexts. Furthermore, the experiment's findings support the proposal of this new methodology, which is better suited to the business environment.

Keywords: Heuristic evaluation, usability evaluation, experts of usability, usability percentage.

1 Introduction

Studying everything related to the Human-Computer Interaction discipline provides significant benefits to the design of any interactive system. This is because everything that is built must ultimately interact with a human at some point Dix *et al.* [2003] Gerlach and Kuo [1991]. Concepts such as Usability or User eXperience (UX), have become as essential as the functionality of any technology-mediated system. Usability is an internal quality of interactive systems defined as “the ability of software to be understood, learned, used and appealing to the user, under specific conditions of use” Bevan *et al.* [1991]. User eXperience (UX), refers to the overall perception of a user when interacting with a product, system or service, and how that interaction impacts on their emotions, attitudes and satisfaction Allam *et al.* [2013]. In this context, companies are increasingly adopting User-Centered Design (UCD) methodologies to improve their products through Quality of Use ISO/IEC [2020] ISO/IEC [2017]. In addition, new UX consultancy firms are emerging worldwide to assist in these tasks and guide online businesses.

Since many years, usability evaluation techniques emerged, and still are evolving, to find mechanisms for measuring what usability or UX mean (in terms of real quality of use of interactive systems) Fernández and Macias [2021] Goldbort [2016]. The justification for conducting the research outlined in the provided text stems from the fundamental importance of Human-Computer Interaction (HCI) in the design of interactive systems. Given the increasing adoption of User-Centered Design (UCD) methodologies by companies to enhance their products' Quality of Use,

there is a growing need for effective usability evaluation techniques. These techniques aim to measure the real quality of use of interactive systems and are continuously evolving to meet the demands of modern digital environments.

In this work, we put our focus in one of the most known and used techniques, that is Heuristic Evaluation (HE) Nielsen [1994] Nielsen and Molich [1990]. We will take one of these proposals (developed by Granollers Granollers [2018b] Granollers [2018a] and mentioned in the newest version of *Interaction Design* book Preece *et al.* [2015] as a highly interesting proposal) to carry out a set of experiments in a business context in order to validate and improve the above mentioned method.

The article is organised as follows: firstly, background and the context of the study is presented followed by the methodology, results obtained in the evaluation and, a brief discussion. The final part presents potential threats to validity of research, conclusions, and future work.

2 Background and related work

Technological advances, increased market competition and more sophisticated customer expectations have made usability, once a luxury, a necessity. In this context, HE is a very useful method because its effectiveness and large body of studies, methodological variations and new proposals appeared to reach more and more effective usability evaluations in several contexts: web pages Bonastre and Granollers [2014], Charts of LIS Journals Alcaraz *et al.* [2021b], Augmented Reality (AR) systems Derby and Chaparro [2021] even, virtual reality (VR) interfaces Patnaik and Adrian

[2022], Cheiran et al. [2021],.

However, several author explore about how usable are they to carry on a evaluation between group of UX experts Mutlu [2023] or even how the heuristics could be for them easy of understand to judge the system accurately Hvannberg et al. [2006]

In this sense, we have been working for a long time on related aspects to improve usability evaluations. In particular, HE is one of the techniques in which we have been most involved. Members of our research group have created the chapter "Heuristic evaluation" in the book "Human-Computer Interaction" [González, 2001], as well as several research works in which we have studied on improving the heuristic evaluation technique. First, we used classical heuristics to perform evaluations of Spanish university web sites Lorés J [2005] Gonzalez MP [2008] and e-commerce environments Pascual-Almenara and Granollers-Saltiveri [2021]. Subsequently, we created lists of heuristic principles to analyze static graphs Alcaraz et al. [2021a]. Then, we analyzed and unified in a new proposal the heuristics of Nielsen and Tognazzini Granollers [2018b], this allowed us to obtain a list of heuristics more generic Granollers [2010b] or more directed to e-commerce environments Granollers [2010a].

2.1 Methodology of Heuristic Evaluation

In this work, we analyse whether this new proposal heuristic evaluation technique (developed by Granollers Granollers [2018b]) works well in a real business and professional context. The methodology proposed by Granollers have this characteristics: a complete list of 15 principles, as commented before, resulting from analyzing and synthesizing the Usability Heuristic Principles for the Design of User Interfaces by J. Nielsen and the Interface Design Principles by B. Tognazzini (see Table 1). The selection of these heuristics stems from a strategic combination of Nielsen’s focus on system usability and Tognazzini’s emphasis on interaction design. By integrating both perspectives, a more comprehensive and optimized list of criteria for evaluating the usability and interaction of digital systems can be achieved. This approach ensures a holistic assessment that accounts for both usability principles and design elements, thereby enhancing the effectiveness of heuristic evaluation in assessing digital system performance.

In the end of the analysis the list of Nielse and Tognazzini’s, a total of 60 specific questions were obtained.

Each of them have a scored with only 4 answers ("Yes" with value 1, "Yes, but some cases missing" with value 0,66, "Not always" with value 0,33, "No" with value 0).

In the case of questions where the answer is "Not applicable" or "Not a problem", these will not be computed, i.e. it will be as if that question did not exist. Questions where the answer is "Warning (impossible to check)", will be counted, but no marks will be assigned to them (see Table 2).

Finally, with all the values of each answer we could get a value called Usability Percentage (UP) that gives a global idea of the usability level of the analyzed interface.

The resulting heuristic has never been tested before and this document shows how it was used in a real business envi-

ronment.

Table 1. List of 15 heuristic principles evaluated in the proposed methodology.

Id	List of heuristic principles
1	Visibility and system state
2	Connection between the system and the real world, metaphor usage and human objects
3	User control and freedom
4	Consistency and standards
5	Recognition rather than memory, learning and anticipation
6	Flexibility and efficiency of use
7	Help users recognize, diagnose and recover from errors
8	Preventing errors
9	Aesthetic and minimalist designs
10	Help and documentation
11	Save the state and protect the work
12	Colour and readability
13	Autonomy
14	Defaults
15	Latency reduction

Table 2. Score associated with each answer.

Answers	Score
YES, in all cases	1
Yes, but some cases missing	0,66
Not always	0,33
NO	0
Not applicable	N/A
It is NOT a problem	N/P
WARNING (Impossible to check)	WR

3 Study context

Sperientia [studio-lab]¹, a research laboratory in User Experience with its headquarters in Mexico, specialized in evaluating and researching the user experience of digital products and services, they agree to participate in this study as it could be useful for improving their job. To enable Sperientia participation in the experiment, take part in the study. Fifteen usability experts were organized in three teams or "labs"², (LabX, LabY, LabZ) consisted of between 3 and 6 experts, only senior usability experts with 3-5 years of experience were selected. The variation of experts in each lab depended on the number of people working in each of the company’s sites. They participate in the study, being responsible for the different evaluations, and the discussion of the results and the final surveys. The experiment consisted of evaluating 8 websites, four supermarket platforms (HEB, LaComer, Walmart and Chedraui) and four bank platforms (Santander, BBVA, Banorte and HSBC) considering two relevant sectors

¹Sperientia [studio-lab]: <https://www.overleaf.com/project/6592f856a028f41b1bcabfec>
²They prefer the term "lab" more than "team".

such as food and banking, where millions of users make transactions and purchases online. Each website was evaluated by three *Sperientia labs* and following with the same instrument, the heuristic methodology proposed in by Granollers. See section 2 for more details.

In order to evaluate whether it is more effective to conduct evaluations individually, and then share the results with the other experts in the laboratory or, on the contrary, it is better to perform them in a group from the beginning, the type of evaluation of each website was alternated between individual and group evaluations. Out of the 8 evaluations carried out, 4 were done in groups and 4 individually (see Table 4).

3.1 Launching the Study

The evaluation of supermarkets and banks websites was carried out from February to May 2021 with 15 expert evaluators from *Sperientia [studio-lab]*. To facilitate the knowledge of sites to be assessed by the evaluators before starting the heuristic evaluation, a set of different tasks was proposed for each site: 4 tasks for the supermarkets and 3 tasks for the banks (see Table 3). All the tasks were directly related with habitual uses.

To carry out the study, the usability evaluators first performed the heuristic evaluations, either as a group or individually. Subsequently, they answered a survey in Google Forms to obtain proposals for improving the methodology.

3.2 Usability evaluation

To carry on the usability evaluation, all evaluators used the heuristic methodology proposed by Granollers [2018a] and using a MS© Excel template³ specifically created for this purpose. The template has the 60 questions organized into the 15 heuristic principles as part of the proposal (see Table 1). The template should be used for each individual assessment.

To answer each of the questions in the list of heuristic principles provided by the methodology, different answers were proposed. As can be seen in (see Table 2), the first four answers have an associated score, which depends on the degree to which the question is fulfilled. Each answer follows the colour metaphor of a traffic light, so a reddish colour indicates low compliance, and a greenish colour indicates high compliance of the criterion on the website. Regarding the last three answers (not applicable, not a problem or warning) they do not intervene in the total score since the fact that a question is not fulfilled is not considered a negative aspect. In addition, it is possible that the evaluator may not be able to check the entire system, which should not be scored negatively either. In next section, we can see different tables on Figures (1 to 8). It is important highlight that if the score of a principle is of medium value, it is represented in yellow colour; if the cell is green, it indicates that in general good answers have been obtained; if the colour of the cell is orange or red, it means that most of the answers were negative. Finally, if the cell shows a 0 without colour, it indicates that none of

the questions in that principle have been considered a problem. Another relevant factor to consider in the methodology is that each question has a cell for the evaluator's comments, something really important to understand the values given by every answer.

The heuristic evaluation results in a percentage according to the evaluated principles, the *Usability Percentage (UP)*, which corresponds to the level of usability that the website has. UP is the result of adding the positive values evaluated in each heuristic question and transformed to percentage. The more green colour in the cells, the higher the UP value.

3.3 Survey of evaluation the Heuristic Methodology

One of the aims of this research was to observe whether heuristic principles and methodology were correct for carry out the heuristic evaluation proposal in a business context. Concerning the above, in order to obtain proposals for improving the methodology, each evaluator answered a survey at the end of every heuristic evaluation what he/she did. The survey was launched in google forms format seeking the as honest as possible opinion of the usability experts about the methodology and which aspects should be improved. (See all form questions in the Annex 9).

4 Results

The results are organized according to quantitative data, and they are presented in individual, group and survey results. The UP, which stands for Usability Percentage, is a quantitative measure obtained from synthesising the detailed data of evaluations conducted using the heuristic evaluation methodology.

4.1 Individual results

Supermarket webpages like HEB, Walkmart, and partially Chadraui and bank webpages like Santander, Banorte, and partially BBVA were evaluated individually

4.1.1 Supermarket WebPages – HEB

Several aspects will be discussed on the table shown in Fig 1.

- **LabX:** As we can see, apart from (eval 3), which has a higher percentage, the rest of the evaluations obtained a Usability Percentage around 50%. And in most of the principles, similar results had been obtained. The heuristic principles #15 (Latency reduction) had the lowest score and the heuristic principles #3 (User control and freedom) and #4 (Consistency and standards) were the best rated. The mean of Usability Percentage of Lab X was of 53,04%.
- **LabY:** In general, the results of each principle of HE was very different among them. The two evaluations that obtained the highest results was those with several questions that could not be evaluated (not applicable or not a problem). And although we found that most of the results were different, they coincide in the result

³Heuristic evaluation MS© Excel template: <http://mpiua.invid.udl.cat/wp-content/uploads/2018/04/Evaluaci%C3%B3n-Heuristica-v2018-OK.xlsx>

Table 3. List of tasks by the type of website.

Type of website	List of Tasks
Supermarket	<ul style="list-style-type: none"> 1. Create a shopping list and add products (a litre of Apurna lactose-free milk and 18 rolls of Cottonelle toilet paper); 2. Search for products, identify their nutritional information and add them to the shopping cart (Selecta wheat flour and Gloria unsalted butter 90 grams); 3. Search for a discounted TV screen and add it to the cart (Samsung 65” UltraHD Smart LED TV); 4. Review the products added to the cart and check the shipping price and available delivery times.
Bank	<ul style="list-style-type: none"> 1. Explore the website to search for a credit card with particular characteristics (minimum income of pesos to apply for it, earning points in supermarkets and cost of card cancellation); 2. Search for a nearby ATM and find the route and distance to the ATM, hours and services available; 3. Search for information to apply for a personal loan with the following characteristics (minimum amount of \$100,000.00 Mexican pesos and a maximum repayment period of 36 months)

Table 4. List of websites evaluated.

Id	CompaniesURL	LabX	LabY	LabZ
1	HEB https://www.heb.com.mx/	5 Individual	6 Individual	5 Individual
2	LaComer https://www.lacomer.com.mx/	Group	Group	Group
3	Walmart https://www.walmart.com.mx/	4 Individual	4 Individual	5 Individual
4	Chedraui https://www.chedraui.com.mx/	Group	4 Individual	Group
5	Santander https://www.santander.com.mx/	5 Individual	4 Individual	6 Individual
6	BBVA https://www.bbva.mx/	3 Individual	Group	Group
7	Banorte https://www.banorte.com/	3 Individual	4 Individual	3 Individual
8	HSBC https://www.hsbc.com.mx/	Group	Group	Group

HEB	RESULTS LabX					RESULTS LabY						RESULTS LabZ					
	Eval 1	Eval 2	Eval 3	Eval 4	Eval 5	General	Eval 1	Eval 2	Eval 3	Eval 4	Eval 5	Eval 6	Eval 1	Eval 2	Eval 3	Eval 4	Eval 5
1.- Visibility and system state	1.32	1.65	2.33	2.32	1.38	2.65	1.99	2.65	3.66	1.32	1.65	1.99	1,65	3,3	1,98	0,66	1,32
2.- Connection in the real world	0.99	1.32	3.33	2.64	0.99	2.32	1.99	2.65	2.65	2.65	1.32	2.99	2,66	2,98	1,32	2,65	2,65
3.- User control	2.33	2.33	2.66	3	2.66	0.99	3	1.98	2.33	1.33	0.66	1.99	2,33	2,33	2,33	1,33	1,98
4.- Consistency and standards	3.65	2.99	6	4.33	3.65	1.66	5.33	4.64	3.98	5.66	4.32	4.66	4,65	4,64	2,33	2,98	3,32
5.- Recognition	1.32	1.32	3.99	2.64	0.99	2.33	3.99	2.98	3.31	2.33	1.65	1.66	1,66	2,32	0,33	0,33	1,65
6.- Flexibility	1.99	1.99	1.32	1.33	1.99	2.31	1.65	1.65	2.99	1.33	0	2	0,66	1	0	0,33	1,32
7.- Recover from errors	2.32	1.99	1.99	0.33	1.33	2.65	0.33	1.65	0.33	2	1.99	2.33	0,99	3	1,66	0,33	0
8.- Preventing errors	1.33	1.66	2.66	0.66	1.66	1.32	1.66	0.33	1.33	1.66	0	1	0,33	2	0,66	0,33	0,66
9.- Minimalist design	1.65	1.65	3.66	2.31	1.32	1.98	3.32	2.66	3.32	2.65	0.66	3.32	0,66	0,99	0	0,66	0,66
10.- Help and documentation	2.98	2.98	4.33	2.32	2.98	2.99	3.32	0.99	2.65	3.99	0	0	3,65	2,32	0,99	1,98	0
11.- Save the state	0.99	0.99	2	0	0.66	1	0	0.66	3	0	2	0	0	1	0,66	0,33	0,66
12.- Color and readability	2.99	2.66	1.33	1.65	2.66	1.98	1.99	0.66	2.99	1.32	1.98	3.66	0,66	1,98	1,33	0,66	0,66
13.- Autonomy	0.99	1.32	2.66	1.32	0.99	1.99	2.33	0.99	2.66	1.33	1.99	3	0,66	0,66	0	0,99	0,33
14.- Defaults	0.33	0.33	2	0	0.33	0	0	1	2	0	0	0	0	0	0	0	0,66
15.- Latency reduction	0	0	1.66	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	25.18	25.18	41.92	24.85	23.53	29.17	30.9	25.49	37.2	27.57	18.22	28.6	20,56	28,52	13,59	13,56	15,87
Completed Test	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100,00%	100,00%	100,00%	100,00%	100,00%
MISSING questions	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
# Countable questions	57	56	53	45	53	50	42	45	54	47	49	38	45	47	42	52	49
# NON countable questions	3	4	6	14	7	9	17	4	3	10	8	21	8	13	14	8	6
# WARNINGS	0	0	1	1	0	1	0	11	3	3	3	0	7	0	4	0	5
Usability Percentage (UP)	44.2%	45.0%	77.6%	54.0%	44.4%	57.2%	71.9%	45.5%	65.3%	55.1%	35.0%	73.3%	39,50%	60,70%	29,50%	26,10%	29,40%
MEAN (UP)	53,04%					57.2%	57,68%						37,04%				

Figure 1. Results of individual HE of H-E-B: LabX, LabY, LabZ , with mean of Usability Percent-age.

of worst evaluated, principles #1 (Visibility and system state), and #8 (Preventing errors) and best evaluated, principles #4 (Consistency and standards) and #11 (Save the state and protect the work). The mean individual of Usability Percentage of Lab Y was 57.68% and the group of Usability Percentage was 57.2%. Both values were very similar, which may suggest that group evaluations can be equally or more effective than individual evaluations.

- **LabY:** Most of the results obtained were around the same range of values. The highest rated heuristic principles were 2 (Connection between the system and the real world, metaphor usage and human objects), 3 (User control and freedom) and 4 (Consistency and standards). The worst rated heuristic principles were 9 (Aesthetic and minimalist design) and 13 (Autonomy). The mean of Usability Percentage of Lab Z was 37.04%, a very low value.

4.1.2 Supermarket WebPages – Walmart

Several aspects will be discussed on the table shown in Fig 2.

- **LabX:** In this evaluation, the results obtained were very similar. Only the third evaluation stands out slightly with respect to the others because a higher score was obtained. The highest rated heuristic principles were 4 (Consistency and standards) and 11 (Save the state and protect the work). There is no principle with an outstanding low score, but the principle with the lowest score was 7 (Help users recognize, diagnose, and recover from errors). A mean of Usability Percentage of Lab X score of 73.8% was obtained.
- **LabY:** The results were similar and a Usability Percentage of around 75-85% had been obtained, which is a very good result for a first evaluation of the website. We can highlight that all the evaluators had not considered applicable any heuristic question of principle 11 (Save the state and protect the work) and there was no principle that indicates a serious problem in the evaluated website. A mean of Usability Percentage of Lab Y of 80.98% was obtained.
- **LabZ:** There were two clearly differentiated ranges of values, since while three of the five evaluators who participated in the heuristic evaluation obtained a Usability Percentage of between 50-60%, the remaining two evaluators obtained a Usability Percentage of around 70%. The best rated heuristic principles were 4 (Consistency and standards) and 8 (Preventing errors), and the worst rated heuristic principles were 1 (Visibility and system state) and 2 (Connection between the system and the real world, metaphor usage and human objects). A mean of Usability Percentage of 62.24% was obtained.

4.1.3 Bank WebPages - Santander

Several aspects will be discussed on the table shown in Fig 3.

- **LabX:** The highest scoring heuristic principles were 4 (Consistency and standards) and 10 (Help and documentation). The lowest scoring heuristic principles were 6

(Flexibility and efficiency of use) and 15 (Latency reduction). The mean of Usability Percentage was 53.90

- **LabY:** The highest scoring heuristic principles were 1 (Visibility and system state) and 4 (Consistency and standards). The lowest scoring heuristic principles were 7 (Help users recognize, diagnose, and recover from error) and 9 (Aesthetic and minimalist design). The mean of Usability Percentage of LabY was 60.25%.
- **LabZ:** In these individual evaluations, it could be observed that the first three evaluators marked many answers as "warning". Specifically, the first two evaluators left more than half of the questions unchecked, and the third one left a third unanswered. For this reason, the written comments on each question had been taken into consideration and the quantitative results will not be taken into account, as they were not comparable to the rest. Regarding the three remaining evaluators, two of them obtained very similar Usability Percentage, around 40%, and the last one obtained a much higher Usability Percentage, around 80%. The heuristic principles with the highest scores were 3 (User control and freedom) and 6 (Flexibility and efficiency of use). The lowest scoring heuristic principles were 7 (Help users recognize, diagnose, and recover from errors), 11 (Save the state and protect the work) and 15 (Latency reduction). The mean of Usability Percentage was 58.27%.

4.1.4 Bank WebPages - Banorte

Several aspects will be discussed on the table shown in Fig 4.

- **LabX:** The highest scoring heuristic principles were 1 (Visibility and system state) and 12 (Color and readability). The lowest scoring heuristic principles were 10 (Help and documentation) and 11 (Save the state and protect the work). The mean of Usability Percentage was 43.87
- **LabY:** The highest scoring heuristic principles were 12 (Color and readability) and 13 (Autonomy). The lowest scoring heuristic principles were 5 (Recognition rather than memory, learning and anticipation) and 9 (Aesthetic and minimalist design). The mean of Usability Percentage was 57.03
- **LabZ:** The highest scoring heuristic principles were 3 (User control and freedom) and 4 (Consistency and standards). The lowest scoring heuristic principles were 5 (Recognition rather than memory, learning and anticipation) and 9 (Aesthetic and minimalist design), the same as LabY. The mean of Usability Percentage was 41.77%.

4.2 Group results

Supermarket webpages like LaComer and partially Chadraui and bank webpages like HSBC and partially BBVA were evaluated on group way.

4.2.1 Supermarket WebPages – LaComer

Several aspects will be discussed on the table shown in Fig 5.

Walmart	RESULTS LabX				RESULTS LabY				RESULTS LabZ				
	Eval 1	Eval 2	Eval 3	Eval 4	Eval 1	Eval 2	Eval 3	Eval 4	Eval 1	Eval 2	Eval 3	Eval 4	Eval 5
1.- Visibility and system state	3.31	5	2	3.98	4.66	4.33	3.98	3.65	3.33	4.32	0.99	2.99	3.66
2.- Connection in the real world	2.98	2.32	3.66	2.65	4	4	3.32	2.99	2.66	2.66	0.99	2.32	1.98
3.- User control	1.66	1.66	2.66	2.32	2.66	2	2.66	2.32	3	1.98	2	2.32	1
4.- Consistency and standards	5.66	5.33	5.66	4.98	4.98	6	5.32	4.99	6	3.66	3.33	5	4.31
5.- Recognition	3.31	1.66	4.66	3.31	3.64	4.66	4.32	2.99	4.33	3.65	1.32	1.99	1.99
6.- Flexibility	1.99	3.33	2	2.98	2	1.66	3.32	3	2	0.66	1	0	1.33
7.- Recover from errors	2.66	2.99	0.66	0	0.66	3	2.66	3	1	4	0	4	2
8.- Preventing errors	1.99	2.33	3	0.66	1.65	3	3	2.66	2.32	0.66	1.66	1.66	1.99
9.- Minimalist design	2.98	1.98	3.66	3.32	3.32	4	3.32	2.32	2.66	1.32	2.66	3	2.31
10.- Help and documentation	3.65	3.66	4	0.66	3.32	4.33	3	3.66	0	3.98	0	1.66	2.33
11.- Save the state	2.66	2	2	1	1	1	1	2	1	0.33	1	1	0.66
12.- Color and readability	3	2.31	2.66	1.98	2.65	2.98	4	2.33	2.32	2.98	2	3	1.65
13.- Autonomy	1.98	3	3	2.66	2.32	1.99	2.66	3	3	2.32	0	1.66	1.65
14.- Defaults	0.66	0	0	0	0	0	0	0	1	0	0	0	0
15.- Latency reduction	0.66	1	2	2	0	0	1.66	0	0	0	0	0	0.33
0	39.15	38.57	41.62	32.5	36.86	42.95	44.22	38.91	34.62	32.52	16.95	30.6	27.19
Completed Test	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
MISSING questions	0	0	0	0	0	0	0	0	0	0	0	0	0
# Countable questions	58	51	46	4	45	49	52	50	44	41	23	38	50
# NON countable questions	2	7	13	11	11	10	8	10	2	14	26	20	10
# WARNINGS	0	2	1	3	4	1	0	0	14	5	11	2	0
Usability Percentage (UP)	67.5%	72.8%	88.6%	66.3%	75.2%	85.9%	85.0%	77.8%	59.7%	70.7%	49.9%	76.5%	54.4%
MEAN (UP)	73.8%				80.98%				62.24%				

Figure 2. Results of individual HE of Walmart: LabX, LabY, LabZ, with mean of Usability Percentage.

Santander	RESULTS LabX					RESULTS LabY				RESULTS LabZ					
	Eval 1	Eval 2	Eval 3	Eval 4	Eval 5	Eval 1	Eval 2	Eval 3	Eval 4	Eval 1	Eval 2	Eval 3	Eval 4	Eval 5	Eval 6
1.- Visibility and system state	2.98	2.33	2.99	2.99	3.32	3.32	5	4.32	3.32	2.65	5	1.33	4	2.65	4.33
2.- Connection in the real world	1.98	2.32	2.65	2.65	3.33	2.66	3.32	3.66	2.31	0	4	2.65	0.66	2.32	3.33
3.- User control	1.99	1.32	1.66	2.66	1	2	1	0.66	1	1	0.66	0.66	1.32	1	2
4.- Consistency and standards	3.64	3.99	6	6	3.98	4.66	6	1.65	4.32	1.66	2.64	3.65	3.32	3.99	4.99
5.- Recognition	1.98	1.65	2.66	2.97	1.65	3.98	4.66	2.64	1.32	1.32	4.66	0.99	1.98	1	2.32
6.- Flexibility	2.32	1.65	2.32	1.32	0.33	1.66	2	1.98	1	0.66	2	0.66	1.66	1.66	2
7.- Recover from errors	2.65	2	0	0	0	0	0	0	0	0	0	0	1	0	1
8.- Preventing errors	1.32	3	2.66	2.66	0	1.32	1	1	1	0	1	0	0	0.66	0
9.- Minimalist design	1.32	1.98	3.66	3.66	1.32	4	1.32	2.31	1.65	0.33	1.65	2.65	3.33	1.32	4
10.- Help and documentation	2.32	4.66	4	4	0	1.33	2.99	4.66	3	0	0	0	1.66	1	2
11.- Save the state	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
12.- Color and readability	2.66	1.99	1.33	2	2.66	2.32	2.66	2.98	2.33	2.66	2.65	2	0.66	1.32	2.65
13.- Autonomy	2.32	2	2.66	2.32	1	1.98	2.33	0.66	0	2	0	0	0.33	1.33	2
14.- Defaults	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15.- Latency reduction	0.66	0.66	0	0	0	0	0	1	0	0	0	0	0	0	0
0	28.14	29.55	34.59	33.23	18.59	32.23	32.28	30.52	21.25	12.28	24.26	14.59	19.92	18.25	30.62
Completed Test	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
MISSING questions	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
# Countable questions	49	41	48	42	35	41	42	39	31	30	31	22	41	40	37
# NON countable questions	1	13	6	5	6	7	9	14	18	2	0	18	19	20	22
# WARNINGS	10	6	6	13	19	12	9	7	11	28	29	20	0	0	1
Usability Percentage (UP)	47.7%	62.9%	64.1%	60.4%	34.4%	60.8%	63.3%	66.3%	50.6%	21.2%	40.4%	34.7%	48.6%	45.6%	80.6%
MEAN (UP)	53,90%					60,25%				58,27%					

Figure 3. Results of individual HE of Santander: LabX, LabY, LabZ, with mean of Usability Percentage

Banorte	RESULTS LabX			RESULTS LabY				RESULTS LabZ		
	Eval 1	Eval 2	Eval 3	Eval 1	Eval 2	Eval 3	Eval 4	Eval 1	Eval 2	Eval 3
Heuristic Principles										
1.- Visibility and system state	4.33	3.99	2.98	4.66	3.99	3.31	3.32	0.33	1.99	3.99
2.- Connection in the real world	2.98	2.99	0.99	2.32	2.98	2.99	2.98	0.66	2.32	3.32
3.- User control	1.99	1.33	1.33	1	0	0.66	1	1	1.33	1
4.- Consistency and standards	4.99	3.66	2.98	5.32	5.66	3.65	2.99	2.65	3.65	6
5.- Recognition	2.98	1.99	1.98	1.98	1.65	1.65	3.3	0	0.99	0.66
6.- Flexibility	1.32	0.33	0.99	2.98	4.32	2.32	1	0.66	0	1
7.- Recover from errors	0	0	0.33	0	0	0	0	0	0	0
8.- Preventing errors	1.33	1.32	0	1	1	1	0	0.33	0	0
9.- Minimalist design	1.98	2.32	0	1.32	2.31	1.65	0.33	0	0.33	0.66
10.- Help and documentation	0	0	0.33	1.33	3.64	3.66	3.99	0	1.33	3
11.- Save the state	0	0	0	0	0	0	0	0	0	0
12.- Color and readability	3.33	3	2.66	3.66	2.33	3.66	2.66	1.65	1.65	3.66
13.- Autonomy	1.99	2.32	0.99	2	3	0	2	2.33	0	2
14.- Defaults	0	0	0	0	0	0	0	0	0	0
15.- Latency reduction	0	0	0	0	0	0	1	0	0	0
0	27.22	23.25	15.56	27.57	30.88	24.55	24.57	9.61	13.59	25.29
Completed Test	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
MISSING questions	0	0	0	0	0	0	0	0	0	0
# Countable questions	42	41	38	40	41	36	35	33	32	33
# NON countable questions	4	6	21	13	13	10	15	11	28	20
# WARNINGS	14	13	1	7	6	14	10	16	0	7
Usability Percentage (UP)	48.6%	43.1%	39.9%	58.7%	65.7%	49.1%	54.6%	19.6%	42.5%	63.2%
MEAN (UP)	43,87%			57,03%				41,77%		

Figure 4. Results of individual HE of Banorte: LabX, LabY, LabZ, with mean of Usability Percentage

LaComer	LabX	LabY	LabZ
Heuristic Principles	Group	Group	Group
1.- Visibility and system state	2.99	4.32	2.65
2.- Connection in the real world	3.66	3.66	2.32
3.- User control	1.99	1.66	1.33
4.- Consistency and standards	5	4.66	3.65
5.- Recognition	0.66	0.66	3.64
6.- Flexibility	3.66	4	3.66
7.- Recover from errors	1.99	2.66	3.32
8.- Preventing errors	0.99	2.32	2
9.- Minimalist design	0.99	2.99	1.65
10.- Help and documentation	0	1	1
11.- Save the state	1.33	1	0
12.- Color and readability	1.65	1.65	1.66
13.- Autonomy	1.32	1.66	1.99
14.- Defaults	0	0	0
15.- Latency reduction	0	0	0
0	26.23	32.24	28.87
Completed Test	100.0%	100.0%	100.0%
MISSING questions	0	0	0
# Countable questions	44	43	48
# NON countable questions	16	16	11
# WARNINGS	0	1	1
Usability Percentage (UP)	59.6%	73.3%	58.9%

Figure 5. Results of group HE of LaComer: LabX, LabY, LabZ.

- **LabX:** The highest scoring heuristic principles were 2 (Connection between the system and the real world, metaphor usage and human objects), 4 (Consistency and standards) and 6 (Flexibility and efficiency of use). The lowest scoring heuristic principles were 5 (Recognition rather than memory, learning and anticipation) and 10 (Help and documentation). The Usability Percentage of the group was 59.6%.
- **LabY:** The highest scoring heuristic principles were 2 (Connection between the system and the real world, metaphor usage and human objects), 6 (Flexibility and efficiency of use) and 11 (Save the state and protect the work). The lowest scoring heuristic principles were 5 (Recognition rather than memory, learning and anticipation), 10 (Help and documentation) and 12 (Color and readability). The Usability Percentage of the group was 73.3
- **LabZ:** The highest scoring heuristic principles were 6 (Flexibility and efficiency of use) and 7 (Help users recognize, diagnose, and recover from errors). The lowest scoring heuristic principle was 11 (Save the state and protect the work). The Usability Percentage of the group was 58.9%.

4.2.2 Supermarket WebPages – Chedraui

Several aspects will be discussed on the table shown in Fig 6.

- **LabX:** The highest scoring heuristic principles were 7 (Help users recognize, diagnose, and recover from errors) and 11 (Save the state and protect the work). The lowest scoring heuristic principles were 1 (Visibility and system state), 12 (Color and readability) and 13 (Autonomy). The Usability Percentage of the group was 41.4%.
- **LabY:** This HE was individual. There were principles where most evaluators had agreed on a more positive or negative assessment. The heuristic principles with the highest scores were 4 (Consistency and standards) and 9 (Aesthetic and minimalist design). The heuristic principle with the lowest score was 7 (Help users recognize, diagnose, and recover from errors), despite not having a very low value. The mean of Usability Percentage of the individual HE was 72.73%.
- **LabZ:** The highest scoring heuristic principles were 7 (Help users recognize, diagnose, and recover from errors) and 11 (Save the state and protect the work). The lowest scoring heuristic principles were 1 (Visibility and system state), 12 (Color and readability) and 13 (Autonomy). In both cases these are the same heuristic principles as LabY. The Usability Percentage of the group was 38.0%.

As we can see in the results, the individual mean of Usability Percentage was 72.73% (LabY) and the group result was 41.4% and 38% respectively (LabX, LabZ). There was no correlation between individual and group evaluation

4.2.3 Bank WebPages - BBVA

On the table in Fig 7, we will comment on several aspects.

- **LabX:** This HE was individual. The highest scoring heuristic principles were 1 (Visibility and system state) and 4 (Consistency and standards). The lowest scoring heuristic principle was 11 (Save the state and protect the work). The mean of Usability Percentage was 52.83
- **LabY:** The highest scoring heuristic principles were 4 (Consistency and standards) and 6 (Flexibility and efficiency of use). The lowest scoring heuristic principle was 13 (Autonomy). The mean of Usability Percentage of the group was 59.3
- **LabZ:** The highest scoring heuristic principles were 4 (Consistency and standards) and 12 (Color and readability). The lowest scoring heuristic principle was 15 (Latency reduction). The mean of Usability Percentage of the group was 55,5

As we can see in the results, the mean of Usability Percentage was 52.83% (LabZ) in individual evaluation and the group evaluation had obtained 59.3% and 55.5% respectively (LabY, LabZ). A very similar value was observed between them.

4.2.4 Bank WebPages - HSBC

On the table in Fig 8, we will comment on several aspects.

- **LabX:** The highest scoring heuristic principles were 1 (Visibility and system state), 6 (Flexibility and efficiency of use) and 13 (Autonomy). The lowest scoring heuristic principles were 10 (Help and documentation) and 11 (Save the state and protect the work). The mean of Usability Percentage of the group was 33.3
- **LabY:** The highest scoring heuristic principles were 3 (User control and freedom), 5 (Recognition rather than memory, learning and anticipation), 6 (Flexibility and efficiency of use), 7 (Help users recognize, diagnose and recover from errors) and 13 (Autonomy). And no heuristic principle could be highlighted with a low score. The mean of Usability Percentage of the group was 83.2
- **LabZ:** The heuristic principles with the best scores were 6 (Flexibility and efficiency of use) and 7 (Help users recognize, diagnose and recover from errors). The lowest scoring heuristic principle was (Save the state and protect the work). The mean of Usability Percentage of the group was 66.5

As we can see in the results, the mean of Usability Percentage was 52.83% (LabZ) in individual evaluation and the group evaluation had obtained 59.3% and 55.5% respectively (LabY, LabZ). A very similar value was observed between them.

4.3 Survey results

A total of fifteen evaluators participated on the survey (See Annex 9). The quantitative and qualitative data obtained from the satisfaction surveys completed by the participants at the end of all the heuristic evaluations are presented below.

Chedraui	LabX	INDIVIDUAL RESULTS LabY				LabZ
Heuristic Principles	Group	Eval 1	Eval 2	Eval 3	Eval 4	Group
1.- Visibility and system state	1.98	2.66	4.32	4.32	4.32	1.99
2.- Connection in the real world	1.66	2.65	2.66	2.99	3.32	1.66
3.- User control	1.98	2.32	2.66	1.66	3	1.66
4.- Consistency and standards	3.66	5.33	5.66	4.98	4.99	2.32
5.- Recognition	1.33	2.98	4.32	4.66	2.98	2.66
6.- Flexibility	1.32	1	3.32	1.32	2.99	0.66
7.- Recover from errors	1	1.33	2	1.99	0.33	1
8.- Preventing errors	1.99	2	1.33	2.66	1.66	1.66
9.- Minimalist design	1.65	2.99	3.66	3.66	3.32	2.65
10.- Help and documentation	2.32	4.32	3.66	4.66	2.33	0
11.- Save the state	1.66	0.33	0	2	0	1
12.- Color and readability	0.33	1.66	3.66	4	3	0.66
13.- Autonomy	0.66	1.32	3	1.66	0.66	0.33
14. - Defaults	0	0	0.66	0	0	0
15.- Latency reduction	0	0	0	1	0.66	0
0	21.53	30.89	40.91	41.56	33.56	18.25
Completed Test	100.0%	100.0%	100.0%	100.0%	1	100.0%
MISSING questions	0	0	0	0	0	0
# Countable questions	44	49	51	49	45	36
# NON countable questions	8	10	6	11	11	12
# WARNINGS	8	1	3	0	4	12
Usability Percentage (UP)	41.4%	61.8%	75.8%	84.8%	68.5%	38.0%
MEAN (UP)	41,40%	72,73%				38,00%

Figure 6. Results of group and individual HE of Chedraui: LabX, LabY, LabZ, with mean of Usability Percentage.

BBVA	INDIVIDUAL RESULTS LabX			LabY	LabZ
Heuristic Principles	Eval 1	Eval 2	Eval 3	Group	Group
1.- Visibility and system state	4.32	4.32	3.99	2.99	3.32
2.- Connection in the real world	4	3.66	1.99	2.31	3.66
3.- User control	3	2.66	1	0	1
4.- Consistency and standards	6	5.66	3.33	5	6
5.- Recognition	3.31	3.31	1.99	2.64	2.65
6.- Flexibility	0.66	0.66	0.33	2	1
7.- Recover from errors	0	0	0.33	0	0
8.- Preventing errors	1	1	1.32	1	0
9.- Minimalist design	4	3.66	2.32	3.32	3.66
10.- Help and documentation	3.33	3.33	0	2.99	0
11.- Save the state	0.33	0.33	0	0	0
12.- Color and readability	3	3	2.66	2.66	3
13.- Autonomy	2.32	1.99	2.32	0	2
14. - Defaults	0	0	0	0	0
15.- Latency reduction	0.66	1	0	0	0.33
0	35.93	35.58	21.58	24.91	26.62
Completed Test	100.0%	100.0%	100.0%	100.0%	100.0%
MISSING questions	0	0	0	0	0
# Countable questions	44	42	42	35	30
# NON countable questions	0	1	6	18	12
# WARNINGS	16	17	12	7	18
Usability Percentage (UP)	59.9%	58.6%	40.0%	59.3%	55.5%
MEAN (UP)	52.83%			59.3%	55.5%

Figure 7. Results of group HE of BBVA: LabX, LabY, LabZ, with mean of Usability Percentage.

HSBC	LabX	LabY	LabZ
Heuristic Principles	GROUP	GROUP	GROUP
1.- Visibility and system state	4	3.66	3.98
2.- Connection in the real world	1.66	2.66	4
3.- User control	1.66	1	2
4.- Consistency and standards	2.98	4.99	5.32
5.- Recognition	1.33	5	3.66
6.- Flexibility	1	3	0
7.- Recover from errors	0	2	0
8.- Preventing errors	0.33	0	0
9.- Minimalist design	0.66	3.66	3.32
10.- Help and documentation	0	3.98	0
11.- Save the state	0	0	0
12.- Color and readability	2	3	3
13.- Autonomy	3	2	0
14. - Defaults	0	0	0
15.- Latency reduction	0	0	0
0	18.62	34.95	25.28
Completed Test	100.0%	100.0%	100.0%
MISSING questions	0	0	0
# Countable questions	38	39	28
# NON countable questions	4	18	22
# WARNINGS	18	3	10
Usability Percentage (UP)	33.3%	83.2%	66.5%

Figure 8. Results of group HE of HSBC: LabX, LabY, LabZ.

4.3.1 Quantitative data

The results of the survey allowed see where we could improve the heuristic evaluation methodology in order to be easier to understand. See the charts on Annex 10: Graphic Results of final survey. Regarding Question 1 (Difficulty in understanding the functioning of the heuristic evaluation methodology) (See Fig 11), one third of the evaluators responded that it was of medium difficulty, and the rest of the responses were positive (a score of 6 or more). Regarding Question 2 (Adequacy of the value scale of the heuristic evaluation) (See Fig 12), the results showed that some evaluators were confused about the meaning of the answers, to the point of not knowing which answer to select to answer the evaluation (see Table 4). The results obtained in Question 3 (Are the heuristic principles evaluated sufficient?) (See Fig 13) showed that more than half of the evaluators (66.7%) considered that they were sufficient and adequate to make a complete usability evaluation of a web site. The results obtained in Question 4 (Are the questions asked for each principle sufficient and adequate?) (See Fig 14) more than half of the evaluators (53.3%) considered that the questions included in each principle were not sufficient or adequate for a complete evaluation of a website.

According to Question 5 (Comments are important and add value to the final result) (See Fig 15), 60% of the evaluators considered comments to be essential in the evaluation. The remaining experts also considered the comments important, but not essential. Only 6.7% of the experts did not consider the comments to be important and that the same result could be reached without them. Furthermore, this question shows that, if there were no group discussion after an individual heuristic evaluation, it would be very difficult to understand the responses of the other evaluators. Finally, the evaluators believed that the comments help to identify an error in case you want to resolve it. The results of Question 6 (How much better is this methodology than the one used so far) (See Fig 16) were mixed. According to the comments of evaluators, they indicated that one of the positive points of the methodology is the numerical result (Usability Percentage) as it could be interesting for customers. According to the results obtained in question 7 (Would I use the heuristic evaluation methodology in future evaluations) (See Fig 17), there is a diversity of opinions (33% consider it better, 33% find no difference between the methodology they usually use and 33% do not prefer to use it). About Question 8 (Leave a comment or opinion on aspects to improve the methodology), in general, all the opinions received were positive and indicated that the methodology can be very useful in the business world, although it needs some improvements: for example, adding some principles of the Gestalt law [Graham, 2007] or principles of psychology [Yablonski, 2020] that help to be clearer about the aspects to be improved in each web. It would also be interesting to improve the approach of some questions and clarify the meaning of the answers at the beginning of the evaluation.

4.3.2 Qualitative data

Although the evaluators rated the methodology positively, and that it can be useful in the business environment due to its ease and speed in obtaining a quantitative result of the usability of a website (the Usability Percentage UP), there are several aspects of improvement that can be applied to the methodology to contribute to a better system. The results have been subdivided into three groups: 1. Improvements related to general aspects of the method; 2. Improvement about specific questions; 3. Improvements about specific answers. Each of the problems observed is explained below, together with the best solution considered in each case.

1. Improvements related to general aspects of the method.

- It is necessary to have a section that explains in more detail what each possible question/answer consists of. Improvement: add a comment to the question
- Indicate the evaluation process to observe progress in the total evaluation. Improvement: add a progress bar.

2. Improvement about specific questions

- There are spelling mistakes in some questions. Improvement: correct them.
- Some questions may confuse the evaluator depending on the way they are worded: a) some questions are worded in a very similar way and the evaluator does not know what to answer; b) some questions are write in positive and others in negative and make the evaluator doubt which answer to apply; c) some questions are complex to answer because the evaluator does not understand what is to be evaluated. Improvement: in all these cases the question should be worded more appropriately, or comments should be added to help the evaluator in the evaluation so that he/she does not get confused about the answer.
- There is difficulty in evaluating principle 12 (Color and legibility) because some evaluators do not know how to evaluate it. Improvement: indicate specific instructions or tools in the comments of the question so that the evaluators are clear about the evaluation process.

3. Improvements about specific answers

- The answers are only in one language. Improvement: include both languages.
- The answer "NO" contrasts with "Yes, but some cases are missing". Improvement: for consistency in the answers, change "NO" to "No, in any case" (see Table 2).
- The evaluators had doubts when choosing between the answers "Not applicable" and "No problem", because they were not sure if they had correctly evaluated the guideline. Improvement: It is suggested to use an ASQ (After-Scenario Questionnaire) response system [Lewis, 1991] or a Likert scale [Clark and Watson, 2019], as they are standard ratings, or to use the current scale but supplemented with a glossary specifying what each of the answers means.

5 Discussion

The results of the study were summarised and the highest and lowest rated principles were analysed. The highest rated heuristic principles were: 4. Consistency and standards (7 times), 11. Save the state and protect the work (5 times), 7. Help users recognize, diagnose, and recover from errors (3 times), 2. Connection in the real world (3 times) and 6. Flexibility and efficiency of use (3 times). The lowest rated principles were: 1. Visibility and system state, 12. Color and readability and 13. Autonomy (all of them 3 times). The data indicate that the websites of supermarket analyzed are consistent and store user workspace to ensure better service; the user perceives errors adequately and understands the actions performed by icons and graphic elements; the website interface adapts to different screen resolutions. On the contrary, and according to the data collected in the research carried out, there are interactive elements that users do not perceive (links without underlining, for example); the size and color of the website text makes it difficult to read properly (for example: small and gray text in white font); in some cases the system status is not visible or updated and the user don't know what have to do. About banks evaluation, the highest rated heuristic principles were: 4. Consistency and standards (6 times), 6. Flexibility and efficiency the of use (5 times), 1. Visibility and system state (4 times), 3. User control and freedom (3 times), 12. Color and readability (3 times) and 13. Autonomy (3 times). The lowest rated principles were: 11. Save the state and protect the work (5 times), 15. Latency reduction (3 times), 9. Aesthetic and minimalist design (3 times). Like the supermarket websites, the bank websites analyzed are consistent and the interface is well adapted to small screens. In contrast to the supermarket websites, the websites of banks have interactive elements that the user can easily navigate, the size and color of the text is optimal and the user understands the status of the system.

As shown in the table in Fig 9, the best rated supermarket was Walmart with a mean of Usability Percentage (UP) of 72.34% in individual evaluations and the worst rated was HEB with a mean of UP of 49.25% in group evaluations. The best rated bank was HSBC with a group mean of UP of 61.00% and the worst rated was Banorte with an individual mean of UP of 47.56%. The mean value of the evaluations carried out on the 4 supermarket websites was 59.06% of UP, and on the 4 bank websites was 55.48% of UP. This indicates that the supermarket websites have a better usability than the bank websites. Regarding the analysis of the data obtained individually vs. group, it can be seen that the results have not varied excessively. We obtained 57.88% of mean of UP from group evaluations: HEB: 49.25%, Walmart: 72.34%, BBVA: 55.88% and HSBC: 61.00%. We obtained 56.66% of mean of UP from individual evaluations: LaComer: 63.93%, Chedraui: 50.71%, Santander: 57.47%, Banorte: 47.56%. According to these results, it is considered that evaluate in group or in individual way is adequate. However, it is important to keep in mind that evaluating individually may reveal more usability problems, but it also requires a larger budget due to the additional time needed.

The survey results enabled a qualitative assessment of the methodology. The evaluators analysed all proposals and

Websites		LabX	LabY	LabZ	Mean
HEB	Ind	53,04%	57,68%	37,04%	49,25%
LaComer	Group	59,60%	73,30%	58,90%	63,93%
Walmart	Ind	73,80%	80,98%	62,24%	72,34%
Chedraui	Group	41,40%	72,73%	38,00%	50,71%
Supermarket	Mean	56,96%	71,17%	49,05%	59,06%
Santander	Ind	53,90%	60,25%	58,27%	57,47%
BBVA	Group	52,83%	59,30%	55,50%	55,88%
Banorte	Ind	43,87%	57,03%	41,77%	47,56%
HSBC	Group	33,30%	83,20%	66,50%	61,00%
Banks	Mean	45,98%	64,95%	55,51%	55,48%

Figure 9. Results of questions 1 (Difficulty in understanding the functioning of the heuristic evaluation methodology).

comments to identify those that could substantially improve the heuristic evaluation methodology. Solutions were proposed for all identified problems, resulting in a new version of the methodology (version 2021⁴).

One conclusion drawn from the survey results is that there are both strengths and areas for improvement in the heuristic evaluation methodology. While a significant portion of evaluators found the heuristic principles and questions to be sufficient for conducting a usability evaluation, there were notable challenges identified regarding the understanding and adequacy of certain aspects of the methodology. For instance, the survey revealed that some evaluators struggled with understanding the functioning of the methodology and the adequacy of the value scale used. Additionally, there were mixed opinions regarding the sufficiency and adequacy of the questions asked for each principle, with a majority indicating that they were not entirely sufficient for a complete evaluation. Furthermore, the importance of comments in the evaluation process was highlighted, with a majority of evaluators considering them essential for understanding and adding value to the final result. This suggests that group discussions after individual evaluations play a crucial role in clarifying responses and identifying errors. In conclusion, the survey results underscore the importance of continuous refinement and adaptation of heuristic evaluation methodologies to address usability challenges effectively in the dynamic digital landscape. Incorporating feedback from evaluators and integrating additional principles from related fields could contribute to enhancing the clarity, effectiveness, and applicability of the methodology in real-world contexts.

6 Potential threats to validity of research

Potential threats or limitations in this research could include:

- Limited Generalizability: The experiment was conducted with a specific group of usability experts from a single reputable company. This may limit the generalizability of the findings to broader contexts or different types of evaluators.
- Bias in Evaluators: The expertise and background of the evaluators could introduce bias into the evaluation pro-

⁴Template of HE methodology: <http://mpiua.invid.udl.cat/wp-content/uploads/2023/06/Evaluaci%C3%B3n-Heuristica%20NUEVA%20PROPUESTA%20v2021.xlsx>

cess, potentially impacting the reliability and validity of the results.

- **Influence of Familiarity:** Since the evaluators were using the HE proposal for the first time in real-world scenarios, their lack of familiarity with the technique may have influenced their assessments and the outcomes of the experiment.
- **Scope of Evaluation:** The evaluation focused solely on supermarket and bank websites, which may not fully represent the diversity of digital systems or the range of usability challenges encountered in other contexts.
- **Short-term Effects:** The experiment may not capture the long-term effectiveness or practical implications of the proposed HE methodology in real-world business environments.
- **External Factors:** External factors such as time constraints, organizational dynamics, or technological limitations within the company may have influenced the execution and outcomes of the experiment.
- **Subjective Nature of Usability Evaluation:** Usability evaluation, including heuristic evaluation, inherently involves subjective judgments and interpretations by evaluators, which could introduce variability in the results.

Addressing these potential threats through the methodology, transparency in reporting, and careful interpretation of findings can help strengthen the validity and reliability of the research outcomes.

7 Conclusions and future work

The aim of this research was to analyse the effectiveness of the heuristic evaluation technique in a professional business context. The development and refinement of the New Proposal Heuristic Evaluation methodology mark a significant advancement in usability evaluation within professional business contexts. The experiment carry out provided sufficient information to propose an improved version of the HE methodology, particularly for real-life business environments. After conducting heuristic evaluations of eight websites, including supermarket and banking websites, in a business context, comments were collected from a total of fifteen expert evaluators who used the methodology for a few days. The results showed that the methodology is useful due to its well defined heuristic principles and corresponding questions, as well as its ease of use. The Usability Percentage (UP) study found little difference between group and individual evaluations (group mean of UP is 57.88% and individual mean of UP is 56.66%). Both evaluation options were considered adequate, but conducting the HE individually and then sharing the results with the group of evaluators may help to identify more usability issues. However more budget has to be dedicated to the evaluation as it takes more time. Conducting the evaluation directly in a group can be faster and more cost-effective. All recommendations obtained from the user surveys have led to a new version of the methodology, known as the New Proposal Heuristic Evaluation⁵, more fo-

cused on the business context.

In conclusion, the New Proposal Heuristic Evaluation methodology represents a dynamic and adaptable approach to heuristic evaluation that holds promise for enhancing usability assessment in professional business contexts. By combining individual expertise with collaborative insights, and by embracing iterative refinement and innovation, the New Proposal Heuristic Evaluation methodology framework stands poised to contribute to the ongoing pursuit of user-centered design excellence in the digital age.

Looking ahead, the New Proposal Heuristic Evaluation methodology opens avenues for further refinement and innovation in usability evaluation methodologies. Future research endeavors could explore the integration of complementary evaluation techniques, such as user testing and cognitive walkthroughs, to provide comprehensive insights into website usability. Additionally, the adaptation of the New Proposal Heuristic Evaluation methodology to emerging technologies and digital platforms beyond traditional websites could extend its applicability to diverse contexts, such as mobile applications and e-commerce platforms.

8 Acknowledgements

We would like to sincerely thank the company Speriencia: [studio+lab], for their collaboration, the experience and knowledge provided, as well as all the hours invested in this project. Without them, it would not have been possible to carry out this project as it has been done.

References

- Alcaraz, R., Ribera Turró, M., and Granollers, T. (2021a). Methodology for heuristic evaluation of the accessibility of statistical charts for people with low vision and color vision deficiency. *Universal Access in the Information Society*, 21. DOI: <https://doi.org/10.1007/s10209-021-00816-0>.
- Alcaraz, R., Ribera Turró, M., Marcelino, J., Pascual-Almenara, A., and Granollers, T. (2021b). Accessible charts are part of the equation of accessible papers: a heuristic evaluation of the highest impact journals. *Library Hi Tech*, ahead-of-print. DOI: <https://doi.org/10.1108/LHT-08-2020-0188>.
- Allam, A. H., Hussin, A. R. C., and Dahlan, H. M. (2013). User experience: challenges and opportunities.
- Bevan, N., Kirakowski, J., and Maissel, J. (1991). What is usability.
- Bonastre, L. and Granollers, T. (2014). A set of heuristics for user experience evaluation in e-commerce websites. In *International Conference on Advances in Computer-Human Interaction*.
- Cheiran, J. F. P., Rodrigues, A., and Pimenta, M. S. (2021). Virtual look around: comparing presence, cybersickness and usability for virtual tours across different devices. *Journal on Interactive Systems*, 12(1):191–205. DOI: <https://doi.org/10.5753/jis.2021.2063>.

⁵Speriencia [Template of HE methodology: <http://mpiua.invid.udl.cat/wp-content/uploads/2023/06/Evaluaci%C3%B3n-Heuristica%20NUEVA%20PROPUESTA%20v2021.xlsx>

- Clark, L. A. and Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological assessment*.
- Derby, J. L. and Chaparro, B. S. (2021). The challenges of evaluating the usability of augmented reality (ar). *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 65(1):994–998. DOI: <https://doi.org/10.1177/1071181321651315>.
- Dix, A., Finlay, J., Abowd, G. D., and Beale, R. (2003). *Human Computer Interaction*. Pearson Prentice Hall, Harlow, England, 3 edition.
- Fernández, J. and Macías, J. A. (2021). Heuristic-based usability evaluation support: A systematic literature review and comparative study. In *Proceedings of the XXI International Conference on Human Computer Interaction*, Interacción '21, New York, NY, USA. Association for Computing Machinery. DOI: <https://doi.org/10.1145/3471391.3471395>.
- Gerlach, J. H. and Kuo, F.-Y. (1991). Understanding human-computer interaction for information systems design. *MIS Q.*, 15(4):527–549.
- Goldbort, R. (2016). *Usability evaluation, Chap. 15*. Interaction Design Foundation.
- Gonzalez MP, Granollers T, P. A. L. J. (2008). Testing website usability in spanish-speaking academia through heuristic evaluation and cognitive walkthroughs. In *Journal Universal Comput Sci (JUICS)*, page 14(9):1513–29. Special Issue “Designing the Human–Computer Interaction: Trends and Challenges”. In: Bravo C, Redondo MA, Ortega M, editors. Graz University of Technology, ISSN: 0948-69.
- González, M. P.; Pascual, A. L. J. (2001). *Evaluación Heurística*. Introducción a la Interacción Persona-Ordenador. AIPO: Asociación Interacción Persona-Ordenador.
- Graham, L. M. (2007). Gestalt theory in interactive media design.
- Granollers, T. (2010a). Template heuristic evaluation ecommerce environment. Available: <https://mpiaa.invid.udl.cat/evaluacion-heuristica-para-e-commerce/>. Access: 01 July 2024.
- Granollers, T. (2010b). Template heuristic evaluation generic environment. Available: <https://mpiaa.invid.udl.cat/evaluacion-heuristica-una-nueva-propuesta/>. Access: 01 July 2024.
- Granollers, T. (2018a). Usability evaluation with heuristics , beyond nielsen ’ s list. Available: <https://api.semanticscholar.org/CorpusID:198928969>. Access on 01 July 2024.
- Granollers, T. (2018b). Usability evaluation with heuristics. new proposal from integrating two trusted sources. In *Design, User Experience, and Usability: Theory and Practice: 7th International Conference, DUXU 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15-20, 2018, Proceedings, Part I*, page 396–405, Berlin, Heidelberg. Springer-Verlag. DOI: https://doi.org/10.1007/978-3-319-91797-9_28.
- Hvannberg, E. T., Law, E. L.-C., and Lérusdóttir, M. K. (2006). Heuristic evaluation: Comparing ways of finding and reporting usability problems. *Interacting with Computers*, 19(2):225–240. DOI: <https://doi.org/10.1016/j.intcom.2006.10.001>.
- ISO/IEC (2017). *ISO/IEC TS 25011:2017(en). Information technology — Systems and software Quality Requirements and Evaluation (SQuaRE) — Service quality models*.
- ISO/IEC (2020). *ISO/CD 9241-11. Ergonomics of human-system interaction—Part11: Ergonomics of human-system interaction — Part 110: Interaction principles*.
- Lewis, J. R. (1991). Psychometric evaluation of an after-scenario questionnaire for computer usability studies: The asq. *SIGCHI Bull.*, 23(1):78–81. DOI: <https://doi.org/10.1145/122672.122692>.
- Lorés J, Gonzalez MP, P. A. (2005). Primera etapa de la iniciativa usaipo: usabilidad de páginas de inicio de las universidades españolas. *VI Congreso de Interacción persona ordenador*, 1(1):217–221.
- Mutlu, T. (2023). *Usability Evaluation Methods: How Usable Are They?* Global Studies on Management Information Systems. DOI: <https://doi.org/10.26650/B/SS28ET06.2023.006.16>.
- Nielsen, J. (1994). *Usability Engineering. Capcher 5. Usability Heuristics*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
- Nielsen, J. and Molich, R. (1990). Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '90*, page 249–256, New York, NY, USA. Association for Computing Machinery. DOI: <https://doi.org/10.1145/97243.97281>.
- Pascual-Almenara, A. and Granollers-Saltiveri, T. (2021). Combining two inspection methods: Usability heuristic evaluation and wcag guidelines to assess e-commerce websites. In Ruiz, P. H., Agredo-Delgado, V., and Kawamoto, A. L. S., editors, *Human-Computer Interaction*, pages 1–16, Cham. Springer International Publishing.
- Patnaik, M. and Adrian, A. M. (2022). Chapter 4 - a perspective depiction of heuristics in virtual reality. In Mishra, S., Tripathy, H. K., Mallick, P. K., Sangaiah, A. K., and Chae, G.-S., editors, *Cognitive Big Data Intelligence with a Metaheuristic Approach*, Cognitive Data Science in Sustainable Computing, pages 101–116. Academic Press. DOI: <https://doi.org/10.1016/B978-0-323-85117-6.00006-6>.
- Preece, J., Rogers, Y., and Sharp, H. (2015). *Interaction Design: Beyond Human-Computer Interaction*. Wiley, Hoboken, NJ, 4 edition.
- Yablonski, J. (2020). *Laws of UX: Using psychology to design better products services*. O'Reilly Media.

9 Annex 1: List of questions of survey

These questions were answers for the evaluators on a final survey. The document of the final survey can be found at the URL below: <https://forms.gle/YT5vttQPdFiLPBwd6>

1. Rate the difficulty in understanding the functioning of the methodology provided for carrying out the heuristic evaluations (1. Very difficult - 10. Very easy).
2. Rate the adequacy of the scoring scale for each question (1. Very inadequate - 10. Very adequate). In case you do not find it adequate, please indicate why. What system/scale would you propose (text)?
3. Do you think that the evaluated principles are sufficient/adequate for a complete usability evaluation of a user interface? (Yes - No) If the question was NO, what would you change? (text)
4. Do you think that the questions asked in each principle are sufficient/adequate for a complete evaluation of that principle? (Yes - No). If the question was NO, what would you change (text).
5. Do you consider that the comments are an important part and contribute something positive to the result of the evaluation? (Yes, they contribute a lot; Yes, although not always; They are good, but I think they do not contribute much to the final result of the evaluation - No, I think they are not important and you could reach the same result without them) Why? (text)
6. Based on your first experience, how much better do you consider this methodology than the one used so far (1. very little - 10. a lot) Why? (text)
7. Do you consider using this methodology in future evaluations? (Yes, I find it better than the one used so far; I don't find much difference between this methodology and the one used so far; No, I prefer to use the usual one).
8. Optionally, you can leave a comment about your opinion and/or aspects that you think should be improved about the methodology. (text)

10 Annex 2: Graphic Results of final survey

Due to lack of space, the image of graphic results of the survey are showed in this section.

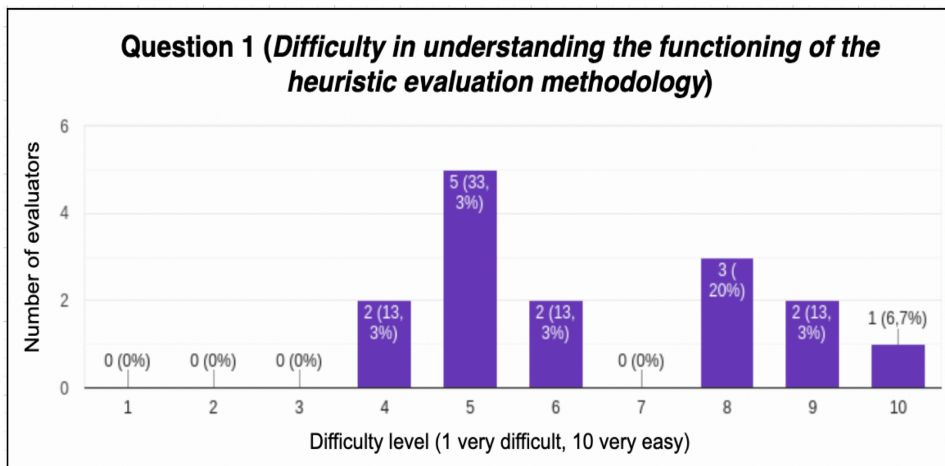


Figure 10. Results of questions 1 (Difficulty in understanding the functioning of the heuristic evaluation methodology).

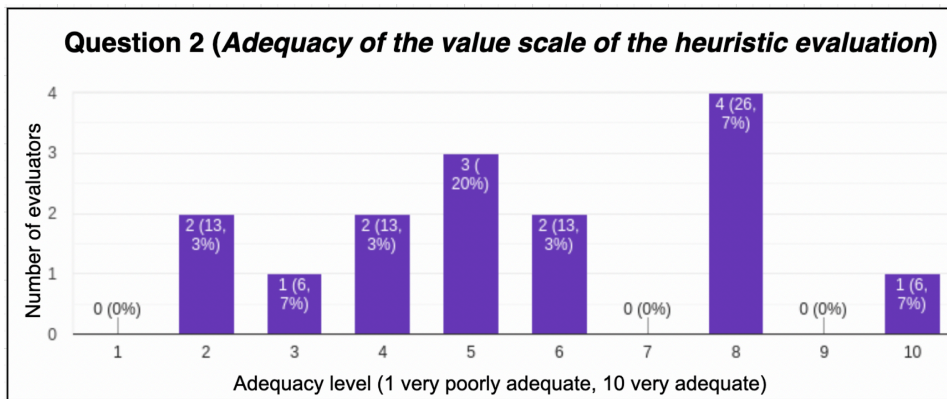


Figure 11. Results of questions 2 (Adequacy of the value scale of the heuristic evaluation)).

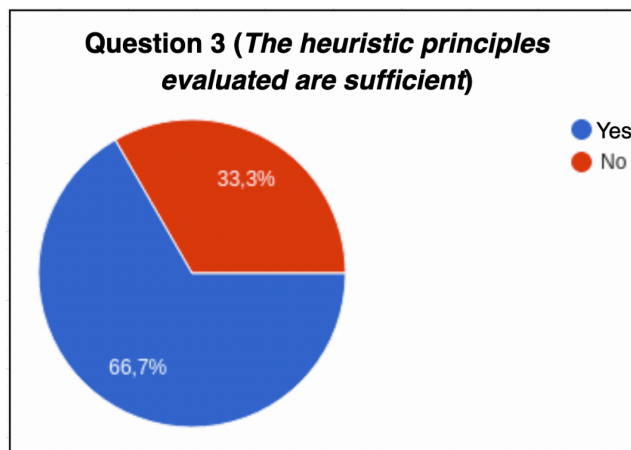


Figure 12. Results of questions 3 (Are the heuristic principles evaluated sufficient?).

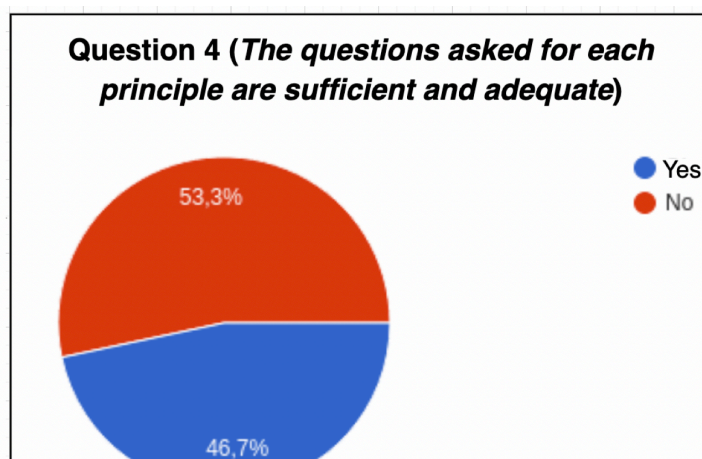


Figure 13. Results of questions 4 (Are the questions asked for each principle sufficient and adequate?).

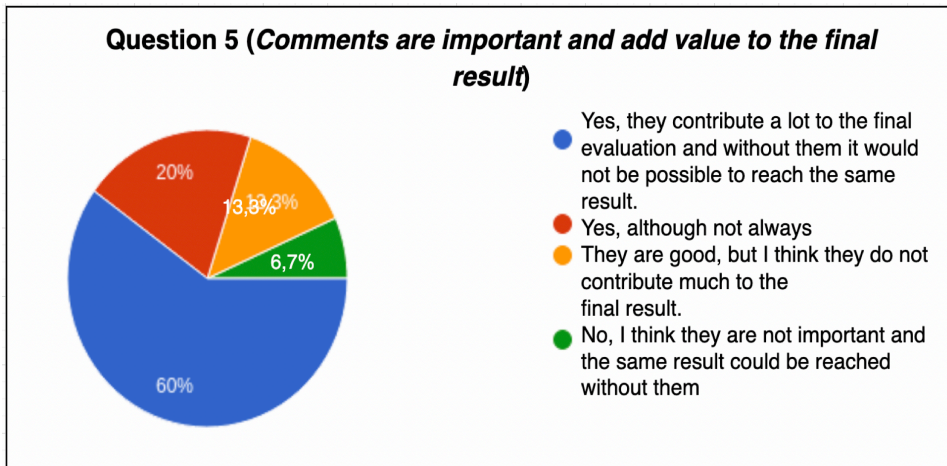


Figure 14. Results of questions 5 (Comments are important and add value to the final result).

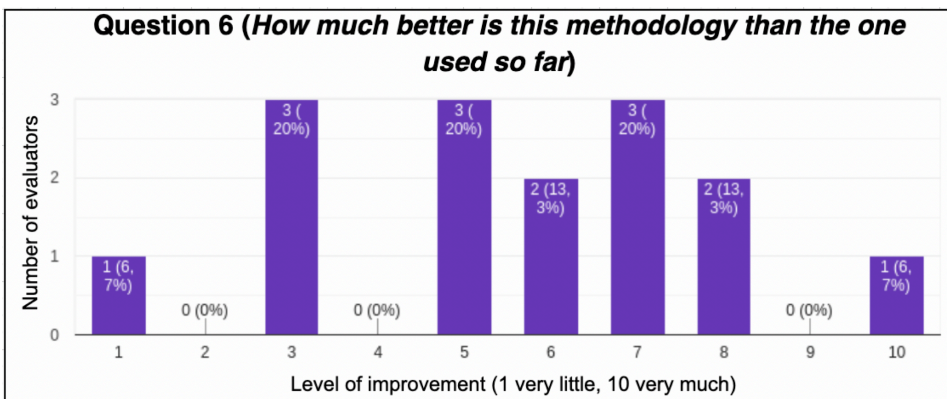


Figure 15. Results of questions 6 (How much better is this methodology than the one used so far).

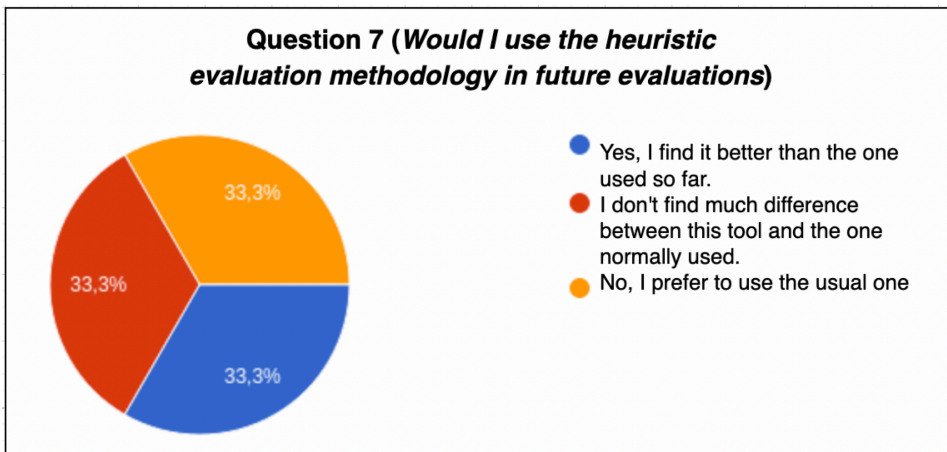


Figure 16. Results of questions 7 (Would I use the heuristic evaluation methodology in future evaluations).