An Online Travel Agency Comparative Study: Heuristic Evaluators Perception

Cristian Rusu[1](✉), Federico Botella[2], Virginica Rusu[3],
Silvana Roncagliolo[1], and Daniela Quiñones[1]

[1] Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile
{cristian.rusu,silvana.roncagliolo}@pucv.cl,
danielacqo@gmail.com
[2] Universidad Miguel Hernández de Elche, Elche, Spain
federico@umh.es
[3] Universidad de Playa Ancha, Valparaíso, Chile
virginica.rusu@upla.cl

Abstract. Forming usability professionals, particularly heuristic evaluators, is a challenging task. Heuristic evaluation is a well-known and widely employed usability evaluation method. A heuristic evaluation may be performed based on generic or specific heuristics. A key issue is how new heuristics are validated and/or evaluated; heuristic quality scales were proposed. The paper presents some recurrent problems when teaching the heuristic evaluation method. It also discusses novice evaluators' perception over Nielsen's usability heuristics, based on empirical data. The experiment that we made involved Computer Science graduate and undergraduate students, enrolled in a Human-Computer Interaction introductory course. 50 Chilean students and 18 Spanish students participated. The online travel agency Atrapalo.com was used as case study. We used a questionnaire that assesses evaluators' perception over a set of usability heuristics. It rates each heuristic individually (Utility, Clarity, Ease of use, Necessity of additional checklist), but also the set of heuristics as a whole (Easiness, Intention, Completeness).

Keywords: Usability · Heuristic evaluation · Usability heuristics
Heuristic quality · Online travel agency

1 Introduction

The usability concept is known for decades and is still evolving. As there is still no general agreement on its definition, we prefer the one stated by the ISO 9241-210 standard: "the extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" [1].

User eXperience (UX) extends usability concept beyond its three widely agreed dimensions: effectiveness, efficiency and satisfaction. The ISO 9241-210 standard defines UX as a "person's perceptions and responses resulting from the use and/or anticipated use of a product, system or service" [1].

G. Meiselwitz (Ed.): SCSM 2018, LNCS 10913, pp. 112–120, 2018.
https://doi.org/10.1007/978-3-319-91521-0_9

Several classifications are used for usability evaluation methods. Lewis identifies two approaches on usability evaluation: (1) summative, “measurement-based usability”, and (2) formative, “diagnostic usability” [2].

For more than two decades heuristic evaluation (HE) has been proved that it is one of the most popular usability evaluation methods [3]. Even if HE is a formative or usability-oriented method, it identifies lots of issues that (potentially) affect a satisfactory UX. Therefore, even if HE does not “measure” UX, it may be considered as a UX method.

When performing a HE, generic or specific heuristics may be used. Nielsen’s ten usability heuristics are well known, but many other sets of usability heuristics were proposed [4, 5]. A key issue is how new heuristics are validated and/or evaluated. A heuristic quality scale was proposed [6]. We developed a questionnaire with a similar purpose.

Forming usability/UX professionals is a challenging task [7, 8]. The paper presents some remarks on teaching the HE method. We also discuss the novice evaluators’ perception over Nielsen’s usability heuristics, based on an experiment that we made.

2 Introducing Heuristic Evaluation to Novices

An easy way to raise awareness about usability/UX topics among Computer Science (CS) students is to practice formal evaluations. Our students have to perform each semester at least one HE and one user test.

We are using Nielsen’s protocol and students have to perform their first HE based on Nielsen’s heuristics [9]. But teaching the HE method for more than a decade allowed us to highlight some pitfalls. Some recurrent problems occur and are described below. They express our teaching experience for almost two decades, but are also empirically supported by explicit students’ comments during the experiment described in this paper.

It is quite difficult for CS students to identify usability and UX related issues. They usually focus on technical issues rather than putting themselves in users’ shoes. It is challenging to make them forget about subjective judgment and to be sympathetic with potential users.

When students find usability problems, it is quite hard to relate them to usability heuristics. It is rather common to associate a usability problem to two or even three different heuristics. We have to emphasize that each heuristic has a different purpose and to correctly understand it needs (quite a lot of) practice. Linking usability problems to usability heuristics is particularly challenging when working with generic heuristics, as Nielsen’s.

Heuristics should be used appropriately as usability issues’ detection tool. However, novice evaluators (and sometimes even experienced ones) focus on identifying usability problems instead of heuristics compliance. This somehow explain why is so difficult to determine problem’s nature and to associate it to specific heuristic(s).

Quantifying the severity of a usability problem is perceived as a rather easy task, but rating its frequency is much more difficult. Students tend to overrate the frequency. That means the criticality of the problem will also be overestimated, as criticality is the sum of severity as frequency. Usability problems' ranking will be affected. Evaluation's report may confuse the targeted audience. Subsequently, the effort of solving usability issues may be wrongly focused.

3 The Experiment

We systematically conduct studies on the perception of (novice) evaluators over generic and specific usability heuristics. All participants are asked to perform a HE of the same software product (case study). Then, all of them participate in a survey.

We developed a questionnaire that assesses evaluators' perception over a set of usability heuristics, concerning 4 dimensions and 3 questions:

- D1 – Utility: How useful the usability heuristic is.
- D2 – Clarity: How clear the usability heuristic is.
- D3 – Ease of use: How easy was to associate identified problems to the usability heuristic.
- D4 – Necessity of additional checklist: How necessary would be to complement the usability heuristic with a checklist.
- Q1 – Easiness: How easy was to perform the heuristic evaluation, based on the given set of usability heuristics?
- Q2 – Intention: Would you use the same set of usability heuristics when evaluating similar software product in the future?
- Q3 – Completeness: Do you think the set of usability heuristics covers all usability aspects for this kind of software product?

The set of usability heuristics is rated globally through the 3 questions (Q1 – Easiness, Q2 – Intention, Q3 – Completeness), but each heuristic is rated separately, on each one of the 4 dimensions (D1 – Utility, D2 – Clarity, D3 – Ease of use, D4 – Necessity of additional checklist). After performing the HE, all participants are asked to rate each usability heuristic, concerning each of the 4 dimensions, using a 5 points Likert scale. The 3 questions aim to evaluate the overall evaluators' perception; responses are also based on a 5 points Likert scale.

We made an experiment with CS graduate and undergraduate students enrolled in a Human-Computer Interaction introductory course in Chile (Pontificia Universidad Católica de Valparaíso, Valparaíso) and Spain (Universidad Miguel Hernández de Elche, Elche). 50 Chilean students (33 graduate and 17 undergraduate), and 18 Spanish undergraduate students participated; we did not select samples, all students enrolled in the HCI course were also participants in the experiment. All of them evaluated the online travel agency Atrapalo.com, based on Nielsen's 10 usability heuristics and following the same protocol. Actually, the Chilean students evaluated Atrapalo.cl and the Spanish students evaluated Atrapalo.es.

We chose Atrapalo.com as case study because it is a widely known online travel agency in Latin America and Spain; its Chilean and Spanish versions are very similar. Moreover, usability and UX in online travel agencies is one of the research topics that we have been working on for years.

After performing the HE all participants were asked to rate their experience, based on the questionnaire described above. Additionally, two open questions were included:

- OQ1: What did you perceive as most difficult to perform during the heuristic evaluation?
- OQ2: What domain-related (online travel agencies) aspects do you think Nielsen's heuristics do not cover?

4 Results and Discussion

Most of the answers to the open question OQ1 confirm the problems described in Sect. 2, in all three groups of students. The most recurrent comments were: "*it is difficult to criticize*", "*it is hard to identify usability-related problems, instead of technical problems*", "*it is hard to think as a user, not as a computer scientist*", "*it is challenging to think as a novice/expert user*", "*it is hard to imagine scenarios of use*", "*it is difficult to link problems to appropriate heuristics*", "*I know there is a problem, but I am not sure what kind of problem is", "it is hard to synthetize/specify the problem*", "*I am not sure how to evaluate frequency*". By far, the most recurrent perceived difficulty was to establish the association usability problem – usability heuristic(s). It worth mentioning that all students felt the need to express their thoughts, answering question OQ1.

Answers to open question OQ2 identify several domain-related (online travel agencies) aspects, which students think Nielsen's heuristics do not cover:

- Effective and efficient transactional process,
- Easy to perform transactional process,
- Clear information on how many steps the process includes,
- Slow query, slow response,
- Information of trust,
- Unexpected system behavior,
- Security-related issues,
- Privacy-related issues,
- Accessibility-related issues,
- Responsivity and adaptability.

All questionnaire items are based on a 5 points Likert scale. Observations' scale is ordinal, and no assumption of normality could be made. Therefore the survey results were analyzed using nonparametric statistics tests (Mann-Whitney U and Spearman ρ).

Table 1 presents the average scores for dimensions and questions. Chilean students had a better (rather positive) perception on Nielsen's heuristics. Dimension D3 (ease of use) got the lowest score, in all cases. Even if heuristics are perceived as useful and clear, students think they are not easy to apply in practice. Moreover, they feel the need for a more complete heuristics' specification (necessity of additional checklist).

Heuristics' perceived overall easiness is low, especially in the case of Spanish students. Heuristics' perceived overall completeness is more homogeneous for the three groups of students. In spite of the above, the intention of future use of Nielsen's heuristics,

Table 1. Average scores for dimensions and questions.

	D1 – Utility	D2 – Clarity	D3 – Ease of use	D4 – Necessity of additional checklist	Q1 – Easiness	Q2 – Intention	Q3 – Completeness
Spanish undergraduate students (18 participants)	3.83	3.43	3.30	3.60	2.78	3.89	3.33
Chilean undergraduate students (17 participants)	4.39	4.04	3.73	4.21	3.53	3.89	3.18
Chilean graduate students (33 participants)	4.39	4.19	3.75	4.27	3.12	4.42	3.60

when evaluating similar products, is rather high for undergraduate and remarkably high for graduate students.

Mann-Whitney U tests were performed to check the hypothesis:

- H_0: there are no significant differences between evaluators with different background,
- H_1: there are significant differences between evaluators with different background.

Spearman ρ tests were performed to check the hypothesis:

- H_0: $\rho = 0$, two dimensions/questions are independent,
- H_1: $\rho \neq 0$, two dimensions/questions are dependent.

In all Mann-Whitney U and Spearman ρ tests, p-value ≤ 0.05 was used as decision rule.

As Table 2 shows, there are significant differences between the perception of Spanish and (all) Chilean students in all cases, excepting question Q3 (Nielsen's heuristics completeness).

Table 2. Mann-Whitney U test for Spanish and (all) Chilean students.

	D1 – Utility	D2 – Clarity	D3 – Ease of use	D4 – Necessity of additional checklist	Q1 – Easiness	Q2 – Intention	Q3 – Completeness
p-value	0.000	0.000	0.003	0.001	0.022	0.028	0.551

When comparing only undergraduate (Chilean and Spanish) students (Table 3), there are significant differences in almost all cases, excepting questions Q2 (intention of future use), and Q3 (Nielsen's heuristics completeness).

On the contrary, the perception of Chilean graduate and undergraduate students is quite similar. There are significant differences only in answers to question Q2, regarding the intention of future use of Nielsen's heuristics (Table 4).

Table 3. Mann-Whitney U test for Spanish and Chilean undergraduate students.

	D1 – Utility	D2 – Clarity	D3 – Ease of use	D4 – Necessity of additional checklist	Q1 – Easiness	Q2 – Intention	Q3 – Completeness
p-value	0.005	0.001	0.022	0.021	0.006	0.666	0.902

Table 4. Mann-Whitney U test for Chilean graduate and undergraduate students.

	D1 – Utility	D2 – Clarity	D3 – Ease of use	D4 – Necessity of additional checklist	Q1 – Easiness	Q2 – Intention	Q3 – Completeness
p-value	0.992	0.185	0.735	0.788	0.091	0.045	0.294

Even if there are significant differences between the perception of Spanish and Chilean students in all dimensions and almost all questions, we do not suspect cultural or background-related issues as possible cause. All students analyzed the same product (Atrapalo.com), using the same set of heuristics (Nielsen's), and following the same protocol. However, some of the Spanish students reported difficulties when scheduling and coordinating their tasks. As this was the only observed difference between the two groups of students, it may somehow influence Spanish students' perception not only on how easy was to perform the HE, but also on the set of heuristics they used.

In the case of the Spanish students (Table 5) there are only two significant correlations:

Table 5. Spearman ρ test for Spanish undergraduate students.

	D1 – Utility	D2 – Clarity	D3 – Ease of use	D4 – Necessity of additional checklist	Q1 – Easiness	Q2 – Intention	Q3 – Completeness
D1	1	0.532	Independent	0.664	Independent	Independent	Independent
D2		1	Independent	Independent	Independent	Independent	Independent
D3			1	Independent	Independent	Independent	Independent
D4				1	Independent	Independent	Independent
Q1					1	Independent	Independent
Q2						1	Independent
Q3							1

- A strong one between D1 – D4. If heuristics are perceived as useful, the necessity of additional evaluation elements (checklist) is also perceived.
- A moderate one between D1 – D2. If heuristics are perceived as clear (easy to understand), they are also perceived as useful.

As Table 6 indicates, the same significant correlations identified for Spanish students also occur in the case of the Chilean students (a strong one between D1 – D4, and a moderate one between D1 – D2). But three other significant correlations occur:

- Two moderate correlations between D2 – D3 and Q2 – Q3. If heuristics are perceived as clear, they are also perceived as easy to use; when the set of heuristics is perceived as complete, there is an intention of future use.
- A weak correlation between D2 – D4. Even if heuristics are perceived as clear, evaluators think that additional checklist is necessary.
- Two weak negative correlations between D2 – Q3 and D3 – Q3. Even if heuristics are perceived as clear and easy to use, evaluators feel that Nielsen's set does not cover all usability aspects of an online travel agency.

Table 6. Spearman ρ test for all Chilean students.

	D1 – Utility	D2 – Clarity	D3 – Ease of use	D4 – Necessity of additional checklist	Q1 – Easiness	Q2 – Intention	Q3 – Completeness
D1	1	0.415	Independent	0.623	Independent	Independent	Independent
D2		1	0.479	0.329	Independent	Independent	−0.341
D3			1	Independent	Independent	Independent	−0.286
D4				1	Independent	Independent	Independent
Q1					1	Independent	Independent
Q2						1	0.416
Q3							1

When analyzing Chilean graduate students' perception, most of the significant correlations that occur for the whole group of Chilean students repeat (Table 7). There are four positive and one negative correlations:

- Two moderate positive correlations (D1 – D4, D2 – D3).
- Two weak positive correlations (D1 – D2, Q2 – Q3).
- A moderate negative correlation (D2 – Q3).

Table 7. Spearman ρ test for Chilean graduate students.

	D1 – Utility	D2 – Clarity	D3 – Ease of use	D4 – Necessity of additional checklist	Q1 – Easiness	Q2 – Intention	Q3 – Completeness
D1	1	0.384	Independent	0.500	Independent	Independent	Independent
D2		1	0.565	Independent	Independent	Independent	−0.527
D3			1	Independent	Independent	Independent	Independent
D4				1	Independent	Independent	Independent
Q1					1	Independent	Independent
Q2						1	0.390
Q3							1

Only two positive significant correlations occur when analyzing Chilean undergraduate students' perception (Table 8):

- A very strong positive correlation (D1 – D4).
- A moderate positive correlation (D3 – Q1).

Table 8. Spearman ρ test for Chilean undergraduate students.

	D1 – Utility	D2 – Clarity	D3 – Ease of use	D4 – Necessity of additional checklist	Q1 – Easiness	Q2 – Intention	Q3 – Completeness
D1	1	Independent	Independent	0.827	Independent	Independent	Independent
D2		1	Independent	Independent	Independent	Independent	Independent
D3			1	Independent	0.488	Independent	Independent
D4				1	Independent	Independent	Independent
Q1					1	Independent	Independent
Q2						1	Independent
Q3							1

The last one does not occur for any other group of evaluators, even if one would expect it. When heuristics are perceived as easy to use, the HE as a method is also perceived as easy to perform.

The only correlation that occurs for all groups of evaluators is D1 – D4. Other recurrent correlation is D1 – D2; it is absent only in the case of Chilean undergraduate students. For both Chilean and Spanish undergraduate students correlations are scarce; they occur only twice. On the contrary, they are relatively frequent correlations in the case of graduate (Chilean) students.

A previous study indicates that most correlations between dimensions occur in the case of evaluators with previous experience [10]. As all participants in our experiment were novice, fewer correlations were expected.

5 Conclusions

Heuristic evaluation is a well-known and arguably the most popular usability inspection method. But forming evaluators is not an easy task; some recurrent problems occur. As method's performance depends mostly on evaluators' skills, we are encouraging students to perform as much evaluations as possible.

We systematically conduct studies on the perception of (novice) evaluators over generic and specific usability heuristics. We developed a questionnaire that evaluates each heuristic individually (Utility, Clarity, Ease of use, Necessity of additional checklist), but also the set of heuristics as a whole (Easiness, Intention, Completeness).

In the comparative study that we have done there are significant differences between the perception of Chilean and Spanish CS students in almost all cases. The perception of Chilean graduate and undergraduate students is rather similar. We do not have evidences to suspect cultural or background-related issues as possible cause;

differences are more likely due to difficulties that some Spanish students reported, related to scheduling and coordinating their tasks.

As in previous studies, few correlations occur between dimensions/questions. Even expected correlations are scarce. The rather heterogeneous students' perception shows that usability heuristics are quite difficult to understand by novice.

The study of novice evaluators' perception helps us in at least two aspects. We better understand the challenges that students are facing; it help us to improve the teaching process, focusing on sensitive issues, explicitly stated in students' comments. They also help us to develop new set of heuristics, for specific domains. In this particular study students highlighted domain-related aspects that Nielsen's heuristics do not cover. Their comments are a valuable asset when designing/refining a set of usability heuristics for online travel agencies.

As future work we intend to analyze the perception of each heuristic individually. We will also analyze the usability problems that students identified during the experiment, the usability heuristic(s) they associated, and the way they rated problems' severity and frecuency.

Acknowledgments. We thank all the students involved in the experiment. They provided helpful opinions that allowed us to prepare this and (hopefully) further documents.

References

1. ISO 9241-210: Ergonomics of human-system interaction — Part 210: Human-centered de-sign for interactive systems. International Organization for Standardization, Geneva (2010)
2. Lewis, J.R.: Usability: lessons learned… and yet to be learned. Int. J. Hum-Comput. Interact. **30**(9), 663–684 (2014)
3. Nielsen, J., Mack, R.L.: Usability Inspection Methods. John Wiley & Sons, New York (1994)
4. Hermawati, S., Lawson, G.: Establishing usability heuristics for heuristics evaluation in a specific domain: is there a consensus? Appl. Ergon. **56**, 34–51 (2016)
5. Quiñones, D., Rusu, C.: How to develop usability heuristics: a systematic literature review. Comput. Stand. Interfaces **53**, 89–122 (2017)
6. Anganes, A., Pfaff, M.S., Drury, J.L., O'Toole, C.M.: The heuristic quality scale. Interact. Comput. **28**(5), 584–597 (2016)
7. Rusu, C., Rusu, V., Roncagliolo, S.: Usability practice: the appealing way to HCI. In: The First International Conference on Advances in Computer-Human Interactions (ACHI 2008) Proceedings, pp. 265–270. IEEE Computer Society Press (2008)
8. Rusu, C., Rusu, V., Roncagliolo, S., González, C.: Usability and user experience: what should we care about? Int. J. Inf. Technol. Syst. Approach **8**(2), 1–12 (2015)
9. Nielsen, J.: 10 Usability Heuristics for User Interface Design. http://www.nngroup.com/articles/ten-usability-heuristics/. Accessed 28 Dec 2017
10. Rusu, C., Rusu, V., Roncagliolo, S., Quiñones, D., Rusu, V.Z., Fardoun, H.M., Alghazzawi, D.M., Collazos, C.A.: Usability heuristics: reinventing the wheel? In: Meiselwitz, G. (ed.) SCSM 2016. LNCS, vol. 9742, pp. 59–70. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-39910-2_6