Evaluating Online Travel Agencies' Usability: What Heuristics Should We Use?

Cristian Rusu[1(✉)], Virginica Rusu[2], Daniela Quiñones[1], Silvana Roncagliolo[1], and Virginia Zaraza Rusu[1]

[1] Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile
{cristian.rusu, silvana.roncagliolo}@pucv.cl, danielacqo@gmail.com, rvzaraza90@hotmail.com
[2] Universidad de Playa Ancha, Valparaíso, Chile
virginica.rusu@upla.cl

Abstract. Online travel agencies' customers have nowadays a wide range of alternatives and are more demanding. Usability is a basic attribute in software quality. Heuristic evaluation is arguably the most popular usability inspection method, well-known and widely used. A heuristic evaluation may be performed based on generic or specific heuristics. Many sets of specific (usually domain-related) usability heuristics were published. Heuristic quality scales to validate and/or evaluate new heuristics were proposed. The paper analyzes evaluators' perception on three sets of usability heuristics, when evaluating the same product: Nielsen's generic heuristics, a set of cultural-oriented heuristics for e-Commerce, and a set of heuristics for smartphones applications (SMASH). We made an experiment with 38 Computer Science students, enrolled in a Human-Computer Interaction introductory course, using the online travel agency Expedia.com as case study; the web and mobile versions were evaluated. We assessed students' perception based on a questionnaire that rates each heuristic individually (Utility, Clarity, Ease of use, Necessity of additional checklist), but also the set of heuristics as a whole (Easiness, Intention of future use, Completeness).

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

1 Introduction

Usability is a basic attribute in software quality. The concept is known for decades and is still evolving. The ISO 9241-210 standard defines usability as "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].

Lewis identifies two approaches on usability evaluation: (1) summative, "measurement-based usability", and (2) formative, "diagnostic usability" [2]. Usability evaluation methods are usually classified in two categories: (1) usability testing, based on users' participation, and (2) inspection methods, based on experts' judgment.

G. Meiselwitz (Ed.): SCSM 2018, LNCS 10913, pp. 121–130, 2018.
https://doi.org/10.1007/978-3-319-91521-0_10

For more than two decades heuristic evaluation is arguably the most popular usability inspection method [3]. When performing a heuristic evaluation generic or specific heuristics may be used. Nielsen's ten usability heuristics are well known, but many sets of specific (usually domain-related) usability heuristics were also published [4, 5]. The results of a heuristic evaluation depend on several factors, but at least two of them are critical: (1) evaluators' expertise, and (2) the set of heuristics that are employed. Heuristic quality scales were proposed.

The paper presents a comparative study on the evaluators' perception over three sets of usability heuristics, when evaluating the same product: Nielsen's heuristics [6], a set of cultural-oriented heuristics for e-Commerce [7], and a set of heuristics for smartphones applications - SMASH [8]. The experiment that we made involved 18 graduate and 20 undergraduate Computer Science students, enrolled in an Human-Computer Interaction (Usability and User eXperience oriented) course. We used Expedia.com as case study [9]; web and mobile versions were evaluated.

2 Comparing Three Sets of Usability Heuristics: A Case Study

We conducted studies on the perception of evaluators over generic and specific usability heuristics for several years [10–12]. All participants are asked to perform a heuristic evaluation of the same case study. All of them are then asked to participate in a post-experiment 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?

Each heuristic is rated individually, on 4 dimensions (D1 – Utility, D2 – Clarity, D3 – Ease of use, D4 – Necessity of additional checklist). The set of usability heuristics is also rated globally, through the 3 questions (Q1 – Easiness, Q2 – Intention, Q3 – Completeness). In all cases, we are using a 5 points Likert scale (from 1 – worst, to 5 – best).

We made an experiment with 38 Computer Science graduate and undergraduate students enrolled in a HCI introductory course at Pontificia Universidad Católica de

Valparaíso (Chile). We did not select samples; all students enrolled in the HCI course were also participants in the experiment. Group composition was as follows:

- 20 undergraduate students; 13 students without previous experience in heuristic evaluation, 7 students with some experience based on Nielsen's heuristics.
- 18 undergraduate students; 11 students without previous experience in heuristic evaluation, 7 students with some experience based on Nielsen's heuristics.

All students evaluated the online travel agency Expedia.com, following the same protocol, but based on three different sets of heuristics:

- Nielsen's 10 usability heuristics [6],
- A set of e-Commerce (cultural-oriented) heuristics [7],
- SMASH, a set of usability heuristics for smartphone applications [8].

They evaluated Expedia.com website (based on Nielsen's and e-Commerce heuristics), and also Expedia mobile application (based on Nielsen's, e-Commerce, and SMASH heuristics). After performing the heuristic evaluation all participants were asked to rate their experience, based on the standard questionnaire described above.

Additionally, two open questions were asked:

- 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 the set of usability heuristics does not cover?

3 Results and Discussion

Table 1 presents the average scores for dimensions and questions. Results are presented globally, but also grouped by students' level (undergraduate or graduate), and level of expertise (with or without previous experience in heuristic evaluation).

Heuristics' perceived utility (D1) is high in all cases (average score over 4.00). In the case of students with previous experience, the perceived utility is slightly in favor of e-Commerce and SMASH heuristics, comparing to Nielsen's heuristics. In general, students with previous experience perceive e-Commerce and SMASH heuristics' utility better than their novice colleagues.

Students perceived e-Commerce and SMASH heuristics' clarity (D2) better than Nielsen heuristics' clarity. The perceived clarity is always better among students with previous experience.

Heuristics' perceived ease of use (D3) is lower than their perceived utility and clarity. Ease of use perception is always better in the case of students with previous experience. e-Commerce and SMASH heuristics are perceived as (slightly) easier to use than Nielsen's heuristics.

The perceived necessity for additional checklist (D4) is higher in the case of Nielsen's heuristics; it is still relatively high for e-Commerce and SMASH heuristics. It is generally higher for novices, comparing to their more experienced colleagues.

Table 1. Average scores for dimensions and questions.

Participants	Previous experience	D1 – Utility	D2 – Clarity	D3 – Ease of use	D4 – Necessity of additional checklist	Q1 – Easiness	Q2 – Intention	Q3 – Completeness
Nielsen's heuristics								
All students (38)		**4.25**	**3.97**	**3.67**	**4.24**	**3.05**	**4.05**	**2.76**
24 students	No	4.30	3.81	3.57	4.33	2.79	4.08	2.71
14 students	Yes	4.16	4.24	3.84	4.09	3.50	4.00	2.86
Undergraduate students (20)		**4.19**	**3.99**	**3.71**	**4.16**	**3.10**	**3.80**	**2.60**
13 students	No	4.25	3.88	3.71	4.32	2.85	4.00	2.92
7 students	Yes	4.09	4.19	3.71	3.87	3.57	3.43	2.00
Graduate students (18)		**4.32**	**3.95**	**3.62**	**4.32**	**3.00**	**4.33**	**2.94**
11 students	No	4.36	3.74	3.41	4.34	2.73	4.18	2.45
7 students	Yes	4.24	4.29	3.96	4.30	3.43	4.57	3.71
e-Commerce heuristics								
All students (38)		**4.19**	**4.09**	**3.84**	**4.05**	**3.13**	**4.08**	**3.63**
24 students	No	4.13	3.95	3.68	4.07	3.00	3.96	3.54
14 students	Yes	4.30	4.31	4.12	4.01	3.36	4.29	3.79
Undergraduate students (20)		**4.22**	**4.12**	**3.92**	**4.09**	**3.20**	**4.05**	**3.45**
13 students	No	4.19	3.96	3.77	4.26	3.08	3.92	3.46
7 students	Yes	4.27	4.42	4.20	3.76	3.43	4.29	3.43
Graduate students (18)		**4.16**	**4.05**	**3.75**	**4.00**	**3.06**	**4.11**	**3.83**
11 students	No	4.06	3.95	3.57	3.84	2.91	4.00	3.64
7 students	Yes	4.32	4.20	4.04	4.26	3.29	4.29	4.14
SMASH heuristics								
All students (38)		**4.21**	**4.05**	**3.77**	**4.04**	**3.26**	**4.05**	**3.61**
24 students	No	4.21	3.94	3.60	4.06	3.17	4.04	3.71
14 students	Yes	4.21	4.24	4.07	4.00	3.43	4.07	3.43
Undergraduate students (20)		**4.23**	**4.04**	**3.74**	**4.05**	**3.35**	**4.10**	**3.55**
13 students	No	4.30	3.95	3.56	4.22	3.31	4.15	3.77
7 students	Yes	4.11	4.20	4.06	3.75	3.43	4.00	3.14
Graduate students (18)		**4.19**	**4.07**	**3.81**	**4.02**	**3.17**	**4.00**	**3.67**
11 students	No	4.11	3.94	3.64	3.87	3.00	3.91	3.64
7 students	Yes	4.32	4.29	4.08	4.25	3.43	4.14	3.71

The overall perception on easiness (Q1, how easy was to perform the heuristic evaluation) is lower than heuristics' perceived utility, clarity, and ease of use. It is quite close to the neutral point of the scale (3). As expected, it is lower for novices than for more experienced students.

Even if the heuristic evaluation is not perceived as an easy task, the intention of future use (Q2) is remarkably high for the three sets of heuristics. It is slightly higher among graduate students, comparing to the undergraduate students.

As expected, students consider that e-Commerce and SMASH heuristics covers better than Nielsen's heuristics the usability aspects of online travel agencies (Q3). Their opinion is less favorable to Nielsen's heuristics in roughly 1 point.

The descriptive statistics presented above was complemented with inferential statistics. As mentioned, 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, Friedman, and Spearman ρ).

As samples are independent, Mann-Whitney U tests were performed to check the hypothesis:

- H_0: there are no significant differences between the perceptions of students with different background,
- H_1: there are significant differences between the perceptions of students with different background.

As the same group of students evaluated three different sets of heuristics, Friedman test was performed to check the hypothesis:

- H_0: there are no significant differences between students' perception on Nielsen's, e-Commerce, and SMASH heuristics,
- H_1: there are significant differences between students' perception on Nielsen's, e-Commerce, and SMASH heuristics.

Spearman ρ tests were performed to check the hypothesis:

- H_0: $\rho = 0$, the dimensions/questions D/Qm and D/Qn are independent,
- H_1: $\rho \neq 0$, the dimensions/questions D/Qm and D/Qn are dependent.

In all tests p-value ≤ 0.05 was used as decision rule.

Table 2 shows Mann-Whitney U tests results when comparing students with and without previous experience. Significant differences occur in very few cases:

- In the case of Nielsen' heuristics, regarding Q1 (easiness) for undergraduate students, and regarding Q1 (easiness) and Q3 (completeness) for graduate students.
- In the case of e-Commerce heuristics, regarding none of the dimensions or questions.

Table 2. Mann-Whitney U tests results when comparing students with and without previous experience.

Set of heuristics	Students' level	p-value						
		D1 – Utility	D2 – Clarity	D3 – Ease of use	D4 – Necessity of additional checklist	Q1 – Easiness	Q2 – Intention	Q3 – Completeness
Nielsen	Undergraduate	0.632	0.321	0.936	0.376	0.036	0.243	0.101
	Graduate	0.785	0.186	0.102	0.681	0.015	0.541	0.015
e-Commerce	Undergraduate	0.662	0.110	0.112	0.260	0.237	0.366	0.966
	Graduate	0.524	0.467	0.237	0.187	0.232	0.328	0.130
SMASH	Undergraduate	0.358	0.404	0.032	0.295	0.599	0.354	0.123
	Graduate	0.645	0.340	0.146	0.169	0.235	0.678	0.689

- In the case of SMASH heuristics, regarding D3 (ease of use) for undergraduate students.

As Table 3 shows, there are no significant differences between undergraduate and graduate students, for none of the three sets of heuristics.

Table 3. Mann-Whitney U tests results when comparing undergraduate and graduate students.

Set of heuristics	p-value						
	D1 – Utility	D2 – Clarity	D3 – Ease of use	D4 – Necessity of additional checklist	Q1 – Easiness	Q2 – Intention	Q3 – Completeness
Nielsen	0.557	0.953	0.597	0.636	0.754	0.095	0.284
e-Commerce	0.682	0.849	0.349	0.713	0.481	0.946	0.127
SMASH	1.000	0.872	0.837	0.976	0.413	0.645	0.604

Table 4. Friedman test results when comparing students' perception on Nielsen's, e-Commerce, and SMASH heuristics.

Students' level	p-value						
	D1 – Utility	D2 – Clarity	D3 – Ease of use	D4 – Necessity of additional checklist	Q1 – Easiness	Q2 – Intention	Q3 – Completeness
Undergraduate	0.821	0.338	0.018	0.498	0.368	0.397	0.000
Graduate	0.559	0.808	0.087	0.692	0.607	0.122	0.001

Friedman test results (Table 4) show significant differences between students' perception on Nielsen's, e-Commerce, and SMASH heuristics in only three cases:

- D3 (ease of use) and Q3 (completeness) for undergraduate students.
- Q3 (completeness) for graduate students.

As Mann-Whitney U tests results show no significant differences between undergraduate and graduate students, and very few significant differences when comparing students with and without previous experience, Spearman ρ tests were performed for the whole group of 38 students.

In the case of Nielsen's heuristics, 8 correlations occur between dimensions/questions (Table 5):

- A strong correlation between D2–D3; if heuristics are perceived as clear, they are also perceived as easy to use.
- 4 moderate correlations between D1–D2, D3–Q1, D2–Q2, and Q2–Q3. If heuristics are perceived as clear, they are also perceived as useful, and there is also a declared intention of future use. If heuristics are perceived as easy to use, the whole heuristic evaluation is perceived as easy to perform. If the set of heuristics is perceived as complete, there is a declared intention of future use.

Table 5. Spearman ρ test for Nielsen's heuristics: correlations between dimensions and questions.

	D1 – Utility	D2 – Clarity	D3 – Ease of use	D4 – Necessity of additional checklist	Q1 – Easiness	Q2 – Intention	Q3 – Completeness
D1	1	0.575	0.352	Independent	Independent	0.357	Independent
D2		1	0.650	Independent	Independent	0.594	Independent
D3			1	Independent	0.424	0.348	Independent
D4				1	Independent	Independent	Independent
Q1					1	Independent	Independent
Q2						1	0.429
Q3							1

- 3 weak correlations between D1–D3, D1–Q2, and D3–Q2. If heuristics are perceived as useful, they are also perceived as easy to use and there is also a declared intention of future use. If heuristics are perceived as easy to use, there is a declared intention of future use.
- It worth mentioning that there is no correlation between D4 (necessity of additional checklist) and any other dimension or question.

As Table 6 indicates, 16 correlations occur in the case of e-Commerce heuristics:

Table 6. Spearman ρ test for e-Commerce heuristics: correlations between dimensions and questions.

	D1 – Utility	D2 – Clarity	D3 – Ease of use	D4 – Necessity of additional checklist	Q1 – Easiness	Q2 – Intention	Q3 – Completeness
D1	1	0.731	0.759	0.408	0.428	0.386	0.556
D2		1	0.800	0.377	Independent	0.541	0.413
D3			1	0.415	0.489	0.444	0.359
D4				1	Independent	Independent	Independent
Q1					1	Independent	0.396
Q2						1	0.336
Q3							1

- Dimension D1 is correlated with all others dimensions and questions. When heuristics are perceived as useful, they are also perceived as clear and easy to use (strong correlations D1–D2, D1–D3); however there is also a declared necessity for additional checklist (moderate correlation D1–D4). When heuristics are perceived as useful, the set of heuristics is perceived as complete, the whole heuristic evaluation is perceived as easy to perform (moderate correlations D1–Q1 and D1–Q3), and there is an intention of future use (weak correlation D1–Q2).

- Dimension D3 is also correlated with all others dimensions and questions. When heuristics are perceived as easy to use, they are also perceived as useful (strong correlation D1–D3), clear (very strong correlation D2–D3), and there is a perceived necessity for additional checklist (moderate correlation D3–D4). The whole evaluation is perceived as easy to perform, there is a declared intention of future use of e-Commerce heuristics, and the set of heuristics is perceived as complete (moderate correlations D3–Q1, D3–Q2, and weak correlation D3–Q3).
- Dimension D2 is correlated to all others dimensions and questions, excepting one (Q1). Correlations are very strong (D2–D3), strong (D1–D2), moderate (D2–Q2, D2–Q3), or weak (D2–D4).
- Question Q3 is also correlated to all others dimensions and questions, excepting one (D4). Correlations are moderate (D1–Q3, D2–Q3) or weak (D3–Q3, Q1–Q3, Q2–Q3).
- Question Q2 is correlated to all others dimensions and questions, excepting D4 and Q1. Correlations are moderate (D2–Q2, D3–Q2) or weak (D1–Q2, Q2–Q3).
- Fewer correlations occur for Q1 (moderate correlations with D1 and D3, weak correlation with Q3), and D4 (weak to moderate correlations with other dimensions, but not with questions).

Table 7 highlights 15 correlations in the case of SMASH heuristics:

Table 7. Spearman ρ test for SMASH heuristics: correlations between dimensions and questions.

	D1 – Utility	D2 – Clarity	D3 – Ease of use	D4 – Necessity of additional checklist	Q1 – Easiness	Q2 – Intention	Q3 – Completeness
D1	1	0.737	0.532	0.384	0.338	0.718	0.341
D2		1	0.820	0.486	0.414	0.651	Independent
D3			1	0.467	0.469	0.468	Independent
D4				1	Independent	0.365	Independent
Q1					1	Independent	Independent
Q2						1	0.419
Q3							1

- As in the case of e-Commerce heuristics, dimension D1 is correlated to all others dimensions and questions (strong correlations between D1–D2 and D1–Q2, moderate correlation between D1–D3, and weak correlations between D1–D4, D1–Q1, and D1–Q3).
- Dimensions D2 and D3 are correlated to all others dimensions and questions, excepting Q3. Correlations are strong or moderate. There is a very strong correlation between D2–D3.
- Dimension D4 is correlated to all others dimensions and questions, excepting Q1 and Q3. Correlations are moderate or weak.
- Question Q2 is correlated with all dimensions (weak to strong correlations), and with question Q3 (moderate correlation).

- Question Q1 is correlated only with dimensions D1 (weak correlation), D2 and D3 (moderate correlation).
- Question Q3 is correlated only with dimension D1 (weak correlation) and question Q2 (moderate correlation).

Correlations are fewer in the case of general (Nielsen's) heuristics (7) than in the case of specific heuristics (e-Commerce, 16, and SMASH, 15). When occur, all correlations are positive.

4 Conclusions

Heuristic evaluation is arguably the most popular usability inspection method. We systematically conduct studies on the perception of evaluators over generic and specific usability heuristics. We are using 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).

Performing heuristics evaluation based on Nielsen's heuristics is a standard practice when we are teaching Human-Computer Interaction courses. This time we asked students to evaluate the same product based on three sets of heuristics: Nielsen's heuristics, e-Commerce heuristics, and SMASH heuristics.

The experiment involved graduate and undergraduate students. There were no significant differences between undergraduate and graduate students' perception, for none of the three sets of heuristics. When comparing students with and without previous experience, significant differences occurred in very few cases. Friedman test results showed significant differences between students' perception on Nielsen's, e-Commerce, and SMASH heuristics in only three cases.

Correlations were fewer in the case of general (Nielsen's) heuristics (7) than in the case of specific heuristics (e-Commerce, 16, and SMASH, 15). When occurred, all correlations were positive.

In general, students' perception on specific (e-Commerce and SMASH) heuristics was slightly better than on generic heuristics (Nielsen). As expected, students consider that e-Commerce and SMASH heuristics covers better than Nielsen's heuristics the usability aspects of online travel agencies.

As future work we intend to analyze the perception of each heuristic individually. We will also analyze students' comments to open questions.

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 design 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. Wiley, 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. Nielsen, J.: 10 Usability Heuristics for User Interface Design. http://www.nngroup.com/articles/ten-usability-heuristics/. Accessed 28 Dec 2017
7. Inostroza, R., Rusu, C., Roncagliolo, S., Rusu, V., Collazos, C.: Developing SMASH: a set of SMArtphone's uSability Heuristics. Comput. Stand. Interfaces **50**, 160–178 (2016)
8. Díaz, J., Rusu, C., Collazos, C.: Experimental validation of a set of cultural-oriented usability heuristics: e-commerce websites evaluation. Comput. Stand. Interfaces **53**, 89–122 (2017)
9. Expedia. http://www.expedia.com. Accessed 10 Jan 2018
10. Rusu, C., Rusu, V., Roncagliolo, S., Apablaza, J., Rusu, V.Z.: User experience evaluations: challenges for newcomers. In: Marcus, A. (ed.) DUXU 2015. LNCS, vol. 9186, pp. 237–246. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-20886-2_23
11. 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
12. Rusu, V., Rusu, C., Quiñones, D., Roncagliolo, S., Collazos, C.A.: What happens when evaluating social media's usability? In: Meiselwitz, G. (ed.) SCSM 2017. LNCS, vol. 10282, pp. 117–126. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58559-8_11