Usability Heuristics: Reinventing the Wheel?

Cristian Rusu[1], Virginica Rusu[2], Silvana Roncagliolo[1],
Daniela Quiñones[1](✉), Virginia Zaraza Rusu[1], Habib M. Fardoun[3],
Daniyal M. Alghazzawi[3], and César A. Collazos[4]

[1] Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile
{cristian.rusu,silvana}@ucv.cl, danielacqo@gmail.com,
rvzaraza90@hotmail.com

[2] Universidad de Playa Ancha de Ciencias de la Educación, Valparaíso, Chile
virginica.rusu@upla.cl

[3] King Abdulaziz University, Jeddah, Saudi Arabia
{hfardoun,dghazzawi}@kau.edu.sa

[4] Universidad del Cauca, Popayán, Colombia
ccollazo@unicauca.edu.co

Abstract. Heuristic evaluation is a well-known and widely accepted usability evaluation method. When performing a heuristic evaluation, generic or specific heuristics may be used. But forming heuristic evaluators may be a challenging task. The paper presents a study that evaluates the perception of (novice) evaluators on Nielsen's usability heuristics. A standard survey was applied in five experiments.

Keywords: Usability · User experience · Usability evaluation · Heuristic evaluation · Usability heuristics

1 Introduction

Usability is a well-known basic attribute in software quality. Over the last decades usability was defined and redefined by many authors. Formal usability definitions were also provided by ISO standards. However, there is still no clear and generally accepted usability definition. Usability's complex nature is hard to describe in a unique definition. User eXperience (UX) is usually considered as an extension of the usability concept. Usability evaluation methods may also be applied in order to assess UX.

Heuristic evaluation is one of the most common usability evaluation methods. Usability specialists (evaluators) examine an interactive software system based on a set of established usability design principles, called heuristics. Generic or specific heuristics may be used. We proposed sets of specific usability heuristics (and associated checklists) for several types of applications. We also proposed a methodology to develop usability heuristics.

We systematically conduct studies on the perception of (novice) evaluators over generic (Nielsen's) and specific usability heuristics. We developed a standard survey that assesses evaluators' perception on a set of usability heuristics, concerning four dimensions: D1 – *Utility*, D2 – *Clarity*, D3 – *Ease of use*, D4 – *Necessity of additional*

G. Meiselwitz (Ed.): SCSM 2016, LNCS 9742, pp. 59–70, 2016.
DOI: 10.1007/978-3-319-39910-2_6

checklist. All dimensions are evaluated using a 5 points Likert scale. The studies offer an important feedback for both teaching and research.

The paper presents a study based on five experiments. Section 2 briefly reviews the concepts of usability, UX, and usability heuristics. Section 3 presents the experimental results. Section 4 points out conclusions and future work.

2 Usability and User Experience Evaluation

A well-known usability definition was proposed by the ISO 9241 standard back in 1998 [1]. The ISO 9241 standard was updated in 2010 [2]. Yet a new revision started briefly after, in 2011 [3]. It proves once again the evolving nature of the usability concept. The current ISO 9241 definition of usability refers to "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".

The UX concept gains popularity. To move from usability to UX is a tendency. 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" [2]. Most authors consider UX as an extension of the usability concept; others still use the terms usability and UX indistinctly.

Measuring effectiveness, efficiency and satisfaction does not represent the only way of evaluating usability. Two major conceptions on usability have been pointed out: (1) summative, focused on metrics, "measurement-based usability", and (2) formative, focused on usability problems detection and associated design solutions, "diagnostic usability" [4].

Usability evaluation methods are basically classified as: (1) empirical usability testing, based on users' participation [5], and (2) inspection methods, based on experts' judgment [6]. Usability evaluation methods may also be applied in order to assess UX. A broad collection of UX evaluation methods is provided by Allaboutux.org [7].

Heuristic evaluation is one of the most popular usability inspection methods. Usability specialists (evaluators) analyze every interactive element and dialog following a set of established usability design principles called heuristics [8]. Generic or specific heuristics may be used. Generic heuristics are familiar to evaluators and therefore (relatively) easy to apply, but they can miss specific usability issues. Specific heuristics can detect relevant usability issues related to the application area [9].

We proposed sets of specific usability heuristics (and associated checklists) for transactional web applications [10], touchscreen-based mobile applications [11], smartphones [12], grid computing applications [13], interactive digital television [14], virtual worlds [15], driving simulators, u-Learning applications, and virtual museums. We also developed a set of cultural – oriented usability heuristics [16]. The experience we had gained led to a methodology to develop usability heuristics [9]; the methodology is currently under review.

3 Heuristic Evaluators' Perception

SIGCHI acknowledges the importance of getting down Human-Computer Interaction (HCI) to the practical work [17]. But forming usability/UX evaluators is a challenging task. We believe that a strong relationship between HCI theory, research and practice is particularly important in countries were HCI communities are not yet well established.

Heuristic evaluations and usability tests are compulsory practice for all our students, at undergraduate and graduate level. As standard practice, at least one heuristic evaluation is performed based on Nielsen's set of 10 usability heuristics [8]. Sometimes a heuristic evaluation based on domain-specific usability heuristic is also performed. After each heuristic evaluation a survey is made, using a standard questionnaire. It gives us an interesting and useful feedback, for teaching and research. Some results have been previously published [18].

3.1 The Questionnaire

We systematically conduct studies on the perception of (novice) evaluators over generic and specific usability heuristics. All participants are asked to perform a heuristic evaluation of the same software product (case study). Then, all of them participate in a survey. We developed a standard questionnaire that assesses evaluators' perception over a set of usability heuristics, concerning 4 dimensions:

- D1 – *Utility*,
- D2 – *Clarity*,
- D3 – *Ease of use*,
- D4 – *Necessity of additional checklist.*

All dimensions are evaluated using a 5 points Likert scale. Five experiments are described below. All of them involved graduate/undergraduate Computer Science (CS) students from Pontificia Universidad Católica de Valparaíso, Chile. All heuristic evaluations were performed using Nielsen's usability heuristics. As observations' scale is ordinal, and no assumption of normality could be made, the survey results were analyzed using nonparametric statistics tests.

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

- H_0: there are no significant differences between evaluators with and without previous experience,
- H_1: there are significant differences between evaluators with and without previous experience.

Spearman ρ tests were performed to check the hypothesis:

- H_0: $\rho = 0$, the dimensions D_m and D_n are independent,
- H_1: $\rho \neq 0$, the dimensions D_m and D_n are dependent.

In all Mann-Whitney U and Spearman ρ tests, $p \leq 0.05$ was used as decision rule.

3.2 The Google Cultural Institute Experiment – Undergraduate Students

An experiment was made, involving 33 CS undergraduate students; 21 of them had previous experience in heuristic evaluations, and 12 others were novice evaluators. All participants were asked to perform a heuristic evaluation over the Google Cultural Institute website (www.google.com/culturalinstitute/). Google Cultural Institute is a web portal that provides access to a huge amount of information; it is in fact a "collection" of virtual museums. Later on a survey was conducted based on the standard questionnaire described in Sect. 3.1.

The Mann-Whitney U test results are shown in Table 1:

- There are no significant differences between the two groups of evaluators (with/without previous experience) in the case of dimension D1 – Utility and D4 – Necessity of additional checklist,
- There are significant differences between the two groups of evaluators in the case of dimension D2 – Clarity and D3 – Easy of use.

Table 1. Mann-Whitney U test for the perception of Nielsen's heuristics when evaluating Google Cultural Institute (CS undergraduate students)

	D1: Utility	D2: Clarity	D3: Ease of use	D4: Necessity of additional checklist
p-value	0.792	**0.018**	**0.006**	0.894

The Spearman ρ test results show that:

- In the case of evaluators with previous experience (Table 2), there are strong correlations between dimensions D1 – D2 and D2 – D3. If heuristics are perceived as clear (easy to understand), they are also perceived as useful and easy to use.
- In the case of novice evaluators (Table 3), there is moderate correlation between dimensions D2 – D3.
- When all evaluators are considered (Table 4), there is strong correlation between dimensions D2 – D3. There are moderate correlations between dimensions D1 – D2 and D1 – D4. If heuristics are perceived as useful, the necessity of additional evaluation elements (checklist) is also perceived.

Table 2. Spearman ρ test for evaluators with previous experience, CS undergraduate students (case study: Google Cultural Institute)

	D1: Utility	D2: Clarity	D3: Ease of use	D4: Necessity of additional checklist
D1	1	0.701	Independent	Independent
D2		1	0.672	Independent
D3			1	Independent
D4				1

Table 3. Spearman ρ test for novice evaluators, CS undergraduate students (case study: Google Cultural Institute)

	D1: Utility	D2: Clarity	D3: Ease of use	D4: Necessity of additional checklist
D1	1	Independent	Independent	Independent
D2		1	0.575	Independent
D3			1	Independent
D4				1

Table 4. Spearman ρ test for all evaluators, CS undergraduate students (case study: Google Cultural Institute)

	D1: Utility	D2: Clarity	D3: Ease of use	D4: Necessity of additional checklist
D1	1	0.592	Independent	0.500
D2		1	0.765	Independent
D3			1	Independent
D4				1

3.3 The Google Cultural Institute Experiment – Graduate Students

A similar experiment to the one described in Sect. 3.2 was made, involving 15 CS graduate students; 10 of them had previous experience, and 5 others were novice evaluators.

The Mann-Whitney U test results are shown in Table 5. There are no significant differences between the two groups of evaluators (with/without previous experience), excepting the dimension D4 – Necessity of additional checklist.

Table 5. Mann-Whitney U test for the perception of Nielsen's heuristics when evaluating Google Cultural Institute (CS graduate students)

	D1: Utility	D2: Clarity	D3: Ease of use	D4: Necessity of additional checklist
p-value	0.385	0.788	0.548	**0.022**

The Spearman ρ test results show that:

- In the case of evaluators with previous experience (Table 6), there is very strong correlation between dimensions D1 – D2.
- In the case of novice evaluators (Table 7), there is very strong correlation between dimensions D2 – D3.
- When all evaluators are considered (Table 8), there are strong correlations between dimensions D1 – D2, D2 – D3, and very strong correlation between dimensions D1 – D3. If heuristics are perceived as easy to use, they are also perceived as useful.

Table 6. Spearman ρ test for evaluators with previous experience, CS graduate students (case study: Google Cultural Institute)

	D1: Utility	D2: Clarity	D3: Ease of use	D4: Necessity of additional checklist
D1	1	0.898	Independent	Independent
D2		1	Independent	Independent
D3			1	Independent
D4				1

Table 7. Spearman ρ test for novice evaluators, CS graduate students (case study: Google Cultural Institute)

	D1: Utility	D2: Clarity	D3: Ease of use	D4: Necessity of additional checklist
D1	1	Independent	Independent	Independent
D2		1	0.917	Independent
D3			1	Independent
D4				1

Table 8. Spearman ρ test for all evaluators, CS graduate students (case study: Google Cultural Institute)

	D1: Utility	D2: Clarity	D3: Ease of use	D4: Necessity of additional checklist
D1	1	0.738	0.816	Independent
D2		1	0.753	Independent
D3			1	Independent
D4				1

3.4 The www.tripadvisor.com Experiment

An experiment was made using Tripadvisor as case study. 31 CS undergraduate students participated; 8 of them had previous experience, and 23 others were novice evaluators. All participants performed a heuristic evaluation of www.tripadvisor.com, a popular platform that shares reviews, compares prices, and offers links to several virtual travel agencies. The standard questionnaire was then applied.

The Mann-Whitney U test results are shown in Table 9. For all dimensions, there are no significant differences between the two groups of evaluators (with/without previous experience).

Table 9. Mann-Whitney U test for the perception of Nielsen's heuristics when evaluating www.tripadvisor.com

	D1: Utility	D2: Clarity	D3: Ease of use	D4: Necessity of additional checklist
p-value	0.101	0.803	0.085	0.138

The Spearman ρ test results show that:

- In the case of evaluators with previous experience (Table 10), there are strong correlations between dimensions D2 – D3, D2 – D4, and very strong correlation between dimensions D3 – D4. Even if heuristics are perceived as easy to use, the necessity of additional evaluation elements (checklist) is also perceived.
- In the case of novice evaluators (Table 11), all dimensions are independent.
- When all evaluators are considered (Table 12), there is weak correlation between dimensions D3 – D4.

Table 10. Spearman ρ test for evaluators with previous experience (case study: www.tripadvisor.com)

	D1: Utility	D2: Clarity	D3: Ease of use	D4: Necessity of additional checklist
D1	1	Independent	Independent	Independent
D2		1	0.743	0.798
D3			1	0.858
D4				1

Table 11. Spearman ρ test for novice evaluators (case study: www.tripadvisor.com)

	D1: Utility	D2: Clarity	D3: Ease of use	D4: Necessity of additional checklist
D1	1	Independent	Independent	Independent
D2		1	Independent	Independent
D3			1	Independent
D4				1

Table 12. Spearman ρ test for all evaluators (case study: www.tripadvisor.com)

	D1: Utility	D2: Clarity	D3: Ease of use	D4: Necessity of additional checklist
D1	1	Independent	Independent	Independent
D2		1	Independent	Independent
D3			1	0.380
D4				1

3.5 The www.expedia.com Experiment – Undergraduate Students

An experiment was made, involving 21 CS undergraduate students; 13 of them had previous experience, and 8 others were novice evaluators. All participants performed a heuristic evaluation of the www.expedia.com website. The standard questionnaire was then applied. Expedia is a popular virtual travel agency.

The Mann-Whitney U test results are shown in Table 13. There are no significant differences between the two groups of evaluators (with/without previous experience), excepting the dimension D3 – Ease of use.

Table 13. Mann-Whitney U test for the perception of Nielsen's heuristics when evaluating www.expedia.com (CS undergraduate students)

	D1: Utility	D2: Clarity	D3: Ease of use	D4: Necessity of additional checklist
p-value	0.466	0.743	**0.045**	0.913

The Spearman ρ test results show that:

- In the case of evaluators with previous experience (Table 14), there is strong correlation between dimensions D2 – D3, and moderate correlation between D3 – D4.
- In the case of novice evaluators (Table 15), there is a very strong negative correlation between dimensions D3 – D4. When heuristics are perceived as easy to use, there is no perceived need for additional evaluation elements (checklist).
- When all evaluators are considered (Table 16), all dimensions are independent.

Table 14. Spearman ρ test for evaluators with previous experience, CS undergraduate students (case study: www.expedia.com)

	D1: Utility	D2: Clarity	D3: Ease of use	D4: Necessity of additional checklist
D1	1	Independent	Independent	Independent
D2		1	0.614	Independent
D3			1	0.582
D4				1

Table 15. Spearman ρ test for novice evaluators, CS undergraduate students (case study: www.expedia.com)

	D1: Utility	D2: Clarity	D3: Ease of use	D4: Necessity of additional checklist
D1	1	Independent	Independent	Independent
D2		1	Independent	Independent
D3			1	−0.976
D4				1

3.6 The www.expedia.com Experiment – Graduate Students

A similar experiment to the one described in Sect. 3.5 was made, involving 15 CS graduate students; 10 of them had previous experience, and 5 others were novice evaluators.

Table 16. Spearman ρ test for all evaluators, CS undergraduate students (case study: www.expedia.com)

	D1: Utility	D2: Clarity	D3: Ease of use	D4: Necessity of additional checklist
D1	1	Independent	Independent	Independent
D2		1	Independent	Independent
D3			1	Independent
D4				1

The Mann-Whitney U test results are shown in Table 17. There are no significant differences between the two groups of evaluators (with/without previous experience).

Table 17. Mann-Whitney U test for the perception of Nielsen's heuristics when evaluating www.expedia.com (CS graduate students)

	D1: Utility	D2: Clarity	D3: Ease of use	D4: Necessity of additional checklist
p-value	0.461	0.157	0.356	0.711

The Spearman ρ test results show that:

- In the case of evaluators with previous experience (Table 18), all dimensions are independent.
- In the case of novice evaluators (Table 19), there is a perfect correlation between dimensions D1 – D4.
- When all evaluators are considered (Table 20), there are strong correlations between dimensions D2 – D3, and D1 – D4.

Table 18. Spearman ρ test for evaluators with previous experience, CS graduate students (case study: www.expedia.com)

	D1: Utility	D2: Clarity	D3: Ease of use	D4: Necessity of additional checklist
D1	1	Independent	Independent	Independent
D2		1	Independent	Independent
D3			1	Independent
D4				1

Table 19. Spearman ρ test for novice evaluators, CS graduate students (case study: www.expedia.com)

	D1: Utility	D2: Clarity	D3: Ease of use	D4: Necessity of additional checklist
D1	1	Independent	Independent	1.000
D2		1	Independent	Independent
D3			1	Independent
D4				1

Table 20. Spearman ρ test for all evaluators, CS graduate students (case study: www.expedia.com)

	D1: Utility	D2: Clarity	D3: Ease of use	D4: Necessity of additional checklist
D1	1	Independent	Independent	0.651
D2		1	0.696	Independent
D3			1	Independent
D4				1

3.7 Discussion

The results of The Mann-Whitney U tests indicate that, in general, evaluators' perception (with or without previous experience) over the Nielsen's usability heuristics is quite similar. In two experiments there are no significant differences in none of the four dimensions: D1 – *Utility*, D2 – *Clarity*, D3 – *Ease of use*, and D4 – *Necessity of additional checklist*. In two experiments there are significant differences regarding only one dimension (D3 and D4, respectively). In one experiment there are significant differences regarding two of the four dimensions (D2 and D3). When occur, differences were related to three of the four dimensions: D2, D3, and D4. Differences related to dimension D1 never occurred.

The results of Spearman ρ tests show that most correlations between dimensions occur in the case of evaluators with previous experience (8); only 4 correlations occur in the case of novice evaluators. When all evaluators are considered, 9 correlations occur. When occur, correlations are usually strong or very strong. All correlations are positive, excepting one.

The most recurrent correlation is between dimensions D2 – D3. It occurs in three experiments when considering all evaluators, in three experiments when considering evaluators with previous experience, and in two experiments when considering novice evaluators. When heuristics are perceived as clear, they are also perceived as easy to use.

Correlation between dimensions D1 – D2 occurs twice in the case of evaluators with previous experience, twice when all evaluators are considered, but never occurs for novice evaluators. In some experiments, when heuristics are perceived as clear, they are also perceived as useful.

Correlation between dimensions D1 – D4 occurs twice when considering all evaluators, and once when considering novice evaluators. But in this particular case the correlation is perfect. So, in some experiments, when heuristics are perceived as useful, there is also a perceived necessity for additional evaluation elements (checklist).

Correlation between dimensions D3 – D4 occurs twice in the case of evaluators with previous experience, once when all evaluators are considered, and once for novice evaluators. But in the last case the correlation is negative (and very strong). So when heuristics are perceived as easy to use, the necessity for additional evaluation elements (checklist) is perceived very differently.

4 Conclusions

Heuristic evaluation is a well-known and frequently applied usability evaluation method. Nielsen's generic heuristics are employed for more than two decades. They are familiar to evaluators and therefore (relatively) easy to apply, but they can miss specific usability issues. Many alternative heuristics were proposed, usually specific for a certain type of applications. Specific heuristics may (potentially) detect relevant usability issues related to the application area.

Forming usability/UX evaluators is a challenging task. A study on the perception of (novice) evaluators over Nielsen's usability heuristics was conducted; five experiments were performed. In general, evaluators' perception (with or without previous experience) is quite similar.

When occur, dependencies between the four surveyed dimensions are somehow expected. The most recurrent correlation is between dimensions D2 (*Clarity*) – D3 (*Ease of use*). When heuristics are perceived as clear, they are also perceived as easy to use. The only unexpected correlation is between dimensions D3 (*Ease of use*) – D4 (*Necessity of additional checklist*). Twice the correlation is positive for evaluators with previous experience, but once is negative and very strong for novice evaluators.

The study offered an important feedback for both teaching and research. The number of correlations within the four surveyed dimensions is relatively low for evaluators with previous experience, and very low for novice evaluators. Nielsen's heuristics are not perceived as one would expect, even when evaluators have previous experience in their use. The study offered relevant information particularly for the refinement of the set of usability heuristics for transactional web applications, and for the development of a new set of heuristics for virtual museums.

We examined the results only at dimension level. A more detailed study will be done at heuristic level. Quantitative analyze will be also complemented with qualitative data, collected through surveys and interviews.

References

1. ISO 9241-11: Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: guidance on usability. International Organization for Standardization, Geneva (1998)
2. ISO 9241-210: Ergonomics of human-system interaction – Part 210: human-centred design for interactive systems. International Organization for Standardization, Geneva (2010)
3. Bevan, N., Carter, J., Harker, S.: ISO 9241-11 revised: what have we learnt about usability since 1998? In: Kurosu, M. (ed.) Human-Computer Interaction. LNCS, vol. 9169, pp. 143–151. Springer, Heidelberg (2015)
4. Lewis, J.: Usability: lessons learned… and yet to be learned. Int. J. Hum.-Comput. Interact. **30**(9), 663–684 (2014)
5. Dumas, J., Fox, J.: Usability testing: current practice and future directions. In: Sears, A., Jacko, J. (eds.) The Human – Computer Interaction Handbook: Fundamentals. Evolving Technologies and Emerging Applications, pp. 1129–1149. Taylor & Francis, New York (2008)

6. Cockton, G., Woolrych, A., Lavery, D.: Inspection – based evaluations. In: Sears, A., Jacko, J. (eds.) The Human – Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications, pp. 1171–1189. Taylor & Francis, New York (2008)
7. Allaboutux.org: All About UX. http://www.allaboutux.org/. Accessed 7 Jan 2016
8. Nielsen, J., Mack, R.L.: Usability Inspection Methods. Wiley, New York (1994)
9. Rusu, C., Roncagliolo, S., Rusu, V., Collazos C.: A methodology to establish usability heuristics. In: The Fourth International Conference on Advances in Computer-Human Interactions (ACHI2011), pp. 59–62, IARIA (2011)
10. Quiñones, D., Rusu, C., Roncagliolo, S.: Redefining usability heuristics for transactional web applications. In: 11th International Conference on Information Technology: New Generations (ITNG2014), pp. 260–265. IEEE Computer Society Press (2014)
11. Inostroza, R., Rusu, C., Roncagliolo, S., Rusu, V.: Usability heuristics for touchscreen-based mobile devices: update. In: First Chilean Conference on Human - Computer Interaction (ChileCHI2013), pp. 24–29. ACM International Conference Proceeding Series (2013)
12. Inostroza, R., Rusu, C., Roncagliolo, S., Rusu, V., Collazos, C.: Developing SMASH: a set of SMArtphone's uSability heuristics. Comput. Stan. Interfaces **43**, 40–52 (2016)
13. Roncagliolo, S., Rusu, V., Rusu, C., Tapia, G., Hayvar, D., Gorgan, D.: Grid computing usability heuristics in practice. In: 8th International Conference on Information Technology: New Generations (ITNG2011), pp. 145–150. IEEE Computer Society Press (2011)
14. Solano, A., Rusu, C., Collazos, C., Arciniegas, J.: Evaluating interactive digital television applications through usability heuristics. Ingeniare **21**(1), 16–29 (2013)
15. Rusu, C., Muñoz, R., Roncagliolo, S., Rudloff, S., Rusu, V., Figueroa, A.: Usability heuristics for virtual worlds. In: The Third International Conference on Advances in Future Internet (AFIN2011), pp. 16–19. IARIA (2011)
16. Diaz, J., Rusu, C., Pow-Sang, J., Roncagliolo, S.: A cultural - oriented usability heuristics proposal. In: First Chilean Conference on Human - Computer Interaction (ChileCHI2013), pp. 82–87. ACM International Conference Proceeding Series (2013)
17. ACM SIGCHI: ACM SIGCHI Curricula for Human-Computer Interaction. http://old.sigchi.org/cdg/cdg2.html#2_1. Accessed 7 Jan 2016
18. 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, Heidelberg (2015)