

ANDRÉ DE LIMA SALGADO

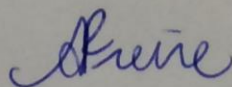
**INVESTIGATING PERFORMANCE
CHARACTERISTICS OF NOVICE EVALUATORS
IN A HEURISTIC USABILITY EVALUATION**

Trabalho de Conclusão de Curso de
Graduação apresentado ao Colegiado do
Curso de Bacharelado em Sistemas de
Informação, para obtenção do título de
Bacharel.

APROVADA em 19 de novembro de 2014.

Dr. Antônio Maria Pereira de Resende

Dr. José Monserrat Neto



Dr. André Pimenta Freire (Orientador)

**LAVRAS-MG
Novembro/2014**

Investigating performance characteristics of novice evaluators in heuristic usability evaluations

André de Lima Salgado

¹Department of Computer Science – Federal University of Lavras (UFLA)
Postal Code 3037 – 37200-000 – Lavras – MG – Brazil

andre@bsi.ufla.br

***Abstract.** Heuristic Evaluation (HE) of usability of web systems has received special attention in the literature. Traditional heuristics are not sufficient to evaluate such systems. In HEs, it is recommended that expert evaluators be employed in HEs. An evaluator became expert after several years of job in usability area. However, many companies resort to novice evaluators in HEs. The effect these evaluators cause in HE is not well known, and deeper research studies about this still remain as a gap in the literature. This study aims to investigate if a set of heuristic focused on usability of web systems can help novice evaluators to find more usability problems than traditional heuristics in a HE of a web system. Results from the HE were compared to results of a test with real users. Obtained results showed the effect of using different sets of heuristics and the quality of reports of novice evaluators. The conclusions of the study showed the importance of investigating the effects of novice evaluators in HE, as well as the extent to which using different sets of heuristics can help improve the overlap between HE results performed by novice evaluators and problems encountered by real users.*

1. Introduction

Usability is the attribute that refers to ensuring that a product is easy to learn, has efficiency and efficacy in use, and is enjoyable from the perspective of users [ISO 9241-11 1998, Rogers et al. 2007]. It is a system quality attribute capable of increasing achievements and results, and leading companies to reduce monetary losses and increase profits [Barua and Mukherjee 2012]. Methods that evaluate this attribute are called Usability Evaluation Methods (UEM).

Web systems are important tools for business activities. The challenge of developing more usable web systems has led to emergence of UEM focused on web usability [Torrente et al. 2013, Fernandez et al. 2013]. UEM underwent various and frequent modifications to support usability on the web to consider specific characteristics of this platform, such as being distributed client-server systems based on hypermedia infrastructure [Fernandez et al. 2011]. Among the numerous methods of UEM, the heuristic evaluation [Nielsen and Molich 1990] has received special attention due to its facility of execution, lower costs in comparison to other inspection methods, and to the quality of its results.

Heuristic Evaluations are widely applied to evaluate usability in desktop applications. It is a type of UEM classified as an inspection method, and its results are strongly determined by the participation of evaluators [Nielsen and Molich 1990, Ling

and Salvendy 2009, de Lima Salgado and Freire 2014]. According to Nielsen and Molich (1990), the execution of heuristic evaluation has both advantages and disadvantages, as listed in Table 1. For web systems, Petrie and Power (2012) state that the traditional set of ten heuristics of Nielsen and Molich [Nielsen 1994a] are not sufficient, as they do not cover specific features of such systems.

Table 1 Advantages and Disadvantages of Heuristic Evaluation according to Nielsen and Molich [1990].

Advantages
Low cost of execution.
Heuristic evaluation is intuitive and it is easy to motivate people to undertake it.
Does not require advance planning.
Heuristic evaluation can be used early in the development process.
Disadvantages
It can identify usability issues without providing ways to correct it.
Are not capable of finding all usability issues.
Depend on usability experts, what is still expensive, to produce better results.

Not just the kind of a system can impact on the results of a HE. The scale of a system, its platform and the expertise of evaluators also impact on the results of HE [Nielsen and Molich 1990, Ling and Salvendy 2009, Botella et al. 2013]. Regarding the effect that the expertise of evaluators can cause on HE results, evidence from previous studies shows that it happens because of judgment bias and individual preferences [Lanzilotti et al. 2011].

According to Nielsen (1992), an expert evaluator has “graduate degrees and/or *several years of job experience in usability area*”. These evaluators are not always easy to be recruited by companies, particularly for Small and Medium-sized Enterprises (SME) and start-ups. Evaluators with such expertise demand more investment, especially if they need to have expertise in a specific domain of application [Nielsen 1992]. For this reason, many organizations resort to novice evaluators in HEs.

Some of the aspects that affect the effectiveness of UEM are the participation of novice usability practitioners and the lack of appropriate approaches. Appropriate approaches can help novice usability practitioners to develop strategies to overcome common issues and improve their experience. This leads to the comprehension of novice evaluator effect in HE knowledge of important interest [Howarth et al. 2009].

The literature in this theme shows that only a few works deal with the effect from novice evaluators in UEM. The effect these evaluators cause in HE is not well known, and deeper research studies about this still remain as a gap in the literature.

This study aims to investigate if a set of heuristic focused on usability of web systems can help novice evaluators to find more usability problems than traditional heuristics in a HE of a web system. The results of this work will contribute for the development of less expensive HEs, for upgrading of UEM training sessions for beginners and for a better understanding of UEM focused on web systems undertaken by novice evaluators.

2. Literature Review

In order to provide a sustainable background on the main terms of this work, this Section explains and shows the literature about *Participation of novice evaluators in UEM*, *usability on web*, the *usability evaluation methods*, including *usability tests* and *heuristic evaluations*, and also describes the *assessment of usability evaluation methods*.

2.2. Usability on the web

Usability is considered a factor that provides easy of use, efficiency and pleasure in use [Mayhew 1999, Rogers et al. 2011]. It is an inherent component of software quality [ISO/IEC 25010 2011]. According to ISO/IEC 25010 (2011), usability is a set of attributes that supports the user to easily understand the logic and applicability of the software product. In addition, usability is the attribute that can be used by users to reach specific goals with efficacy, efficiency and satisfaction in a software product [ISO/IEC 9241-11 1998, Fernandez et al. 2011].

Achieving usability still remains a challenge to be tackled by development teams. The involvement of real users during the development process can incur in high costs. Because of this, it is difficult for developers to understand and implement software to work in the way users think, which can lead to software releases with a number of usability issues. The occurrence of diverse usability problems is a common reality in this context of production, and the development of web systems deals with the barrier.

2.3. Usability Evaluation Methods (UEM)

UEM evaluate the interaction between human and computer for identifying aspects that can be improved and, then, improve the usability [Gray and Salzman 1998]. Different methods for evaluating usability are reported in the literature. Rogers et al. [2011] show that the use of each one UEM “*depends on the goal of the evaluation*”. According to Bastien (2010), the goal of a UEM is to measure usability in terms of efficiency and efficacy.

UEM show problems that impact the usability of a product. In the best scenario, organizations should apply tests with real users to identify usability issues. This is the best method to identify usability issues that real users really care about. However, this kind of UEM is expensive.

HE costs less than test with real users, for this reason it is widely applied by organizations. However, it does not find all usability issues that tests with real users would find. Because of this, the performance of HE is commonly compared to results of tests with real users.

2.3.1. Tests with users

Test with users are classified as a usability test method within UEM. In these tests, a sample of real users of the system is invited to execute some pre-defined tasks. During the period of test, a moderator observes the users' behaviour. Later, a specialist analyzes it. All these steps can occur beneath controlled conditions, or can be performed in more relaxed conditions, depending on the goals of a given test [Bastien 2010, Rogers et al. 2011].

Test with users has some limitations. For tests of web systems, all users must have the same conditions of the Internet. Furthermore, each user has different learning and verbal capabilities, cultures and other specificities, which result in different feedbacks [Bastien 2010].

For success in test with users, the definition of the sample size is an important step, as recruiting all users has high costs. Evaluators have to

To organize tests with users, one of the most important steps of preparation is to define the sample size. Only in few cases is it possible to gather every user of the system. Studies in the literature show that the sample size must be between 7 and 12 users [Dumas and Redish 1999, Sauro and Lewis 2012].

2.3.2. Think Aloud technique

Think Aloud technique incorporates the use of solutions to capture what users are thinking during the use of a system. It is widely applied in combination with test with users [Nielsen et al. 2002]. This technique consists of asking that the user speak out loud his/her thoughts and feelings during the interaction with the software. In test with users, it happens during the performance of the tasks.

The environment for a Think Aloud test is normally pleasurable and capable to generate spontaneous suggestions. This characteristic indicates that the technique is pleasant to users [Plaisant and Shneiderman 2010]. However, some users may be inhibited to talk and may need training to perform better at think-aloud sessions.

2.3.3. Heuristic Evaluation (HE)

Heuristic Evaluation is classified as a usability inspection method. It consists of having a team of evaluators to audit an interface based on a list of heuristics [Nielsen 1994b].

Nielsen and Molich (1990) created the HE method. The purpose was to define a concise set of heuristics that could be used to inspect user interfaces. The authors

proposed an initial set of nine heuristics. Later, Nielsen (1994a) added a tenth heuristic to complete the group.

HEs are performed by the observation of an interface. In HE, evaluators list issues that count as not observing the set of heuristics being considered [Nielsen and Molich 1990]. The strategy of this evaluation is to employ less rigid procedures (when compared to stringent and detailed guideline reviews). It involves the possibility of having specialists judge whether each interface element follows or not usability principals [Nielsen 1994c].

Compared to other usability inspection methods, heuristic evaluations have a faster execution and a lower cost for organizations. Guidelines review and Pluralistic Walkthrough are others popular methods of usability inspection. Guidelines review requires checking of lists with hundreds and even thousand of guidelines, which makes its application more difficult and less frequent. Pluralistic walkthrough demands the presence of users, developers and specialists in usability to be executed [Baranauskas and Rocha 2003].

The execution of a HE comprehends three sessions. In the first session, the brief and preliminary session, specialists must be informed about what to do and receive a general description of the HE goals. In this session, it is recommended to use a pre-defined script that can be given to evaluators. This can ensure that every evaluator will receive the same set of instructions [Rogers et al. 2011].

The second session is the evaluation period. In this session, each evaluator will check the interface at least twice times, inspecting its different interface elements. Evaluators are responsible for reporting all usability issues associate them to one or more related heuristic, and report a severity degree for them. Severity degrees are derived from the impact made by the identified usability issue, from the frequency that the issue occurs and from the persistence of the occurrence of the problem [Rogers et al. 2011, Baranauskas and Rocha 2003].

Finally, the third session is the results session. At this time, evaluators discuss what they have just discovered, to agree on a unified list of problems and severity ratings and to prioritize the issues and to suggest solutions [Rogers et al. 2011].

2.4. Participation of novice evaluators in UEM

Limited evidence exists in the literature about the exploration of characteristics of novice evaluators applying UEMs. Previous work shows that UEM can be compromised with diverse evaluators evaluating through the same UEM and producing different results, with different sets of problems [Hertzum and Jacobsen 2003, Nielsen 1992]. Lanzilotti et al. (2011) investigated the use of patterns, based on expert evaluators' experiences, to help novice evaluators during usability evaluations. Lanzilotti et al. (2011) found that, regarding problems detection, a pattern-based evaluation provided better results than traditional heuristic evaluations. Ling and Salvendy (2009) demonstrated the effect of evaluators' cognitive style during heuristic evaluations and showed that the average of severity was not strongly affected by it. Botella et al. (2013)

prepared a framework, based on a collection of designs and good practices from other works, to help novice evaluators in reporting usability problems.

2.5. Empirical Assessment of Usability Evaluation Methods

The assessment of usability evaluation methods has received important attention in the literature. The literature shows that studies of comparison between different HE frequently use results of test with real users as base for comparison. Hartson et al. (2001) show that the ultimate criterion for UEM effectiveness is finding real usability problems. According to them, a “*usability problem (e.g., found by a UEM) is real if it is a predictor of a problem that users will encounter in real work-context usage and that will have an impact on usability (user performance, productivity, and/or satisfaction)*”.

Hartson et al. (2001) show the three measures for examining a UEM. The comprehension of these three measures depends on the comprehension of *hits*, *misses* and *false alarms*. *Hits* are the problems found by a HE that are found by test with real users, *misses* are problems found by test with real users and not by HE and *false alarms* are problems found by HE and not by test with real users. The measures for examining a UEM are:

- Validity = $Hits / (Hits + False\ Alarms)$.
- Thoroughness = $Hits / (Hits + Misses)$.
- Effectiveness = Validity * Thoroughness.

According to the mapping study of Fernandez et al. (2011), the majority of studies on UEM use the *Thoroughness* measure in evaluation of these methods. This mapping study covered several publications between the years of 1996 and 2009.

The study of Hvannberg et al. (2007) assessed different heuristics sets for conducting HE. In this study, novice evaluators conducted the HE having a highly structured training material. Five evaluators used the heuristic set of Nielsen and Molich, and five used the set of Gerhardt-Powals (1996). Its results showed that the *validity*, *thoroughness* and *effectiveness* of both sets were the same.

Other studies compare different UEM in order to validate a new one. Fernandez et al. (2013) proposed the Web Usability Evaluation Process (WUEP). This study conducted experiments to validate this new method. The validation was aimed to compare evaluators' *Effectiveness*, *Efficiency*, *Perceived Ease of Use* and *Satisfaction* when using WUEP and traditional heuristic evaluation. *Effectiveness* was calculated as the number of reported problems divided by the number of possible problems, what Hartson et al. (2001) call *Thoroughness*. *Efficiency* was calculated as the ratio between the numbers of usability problems detected by the time of evaluation. The results of Fernandez et al. (2013) study show that the WUEP is more effective (that Hartson et al. (2001) call *Thoroughness*) and more efficient than HE. The qualitative analysis showed that WUEP achieved better results in *Perceived Ease of Use* and *Satisfaction* as well. However, the results of this study are limited by some characteristics. First, the analysis of its results was done using the baseline of two expert evaluators for comparison. One of these two experts is one of the authors of Fernandez et al. [2013]. The presence of

one of the authors as baseline for comparisons can influence the comparison because the same author is proposing the new method that is under comparison. In addition, its results are not compared with results from tests with real users, as recommended by previous studies.

Masip (2011) compared two different heuristic sets, aiming to validate a new one. One of the two heuristic sets was the set of Nielsen and Molich. For this set, Masip [2011] considered what González et al. (2009) show, 82 sub-heuristics grouped inside the ten heuristics of Nielsen and Molich and four complementary heuristics. The new group of heuristics was a new set of 16 heuristics and 250 sub-heuristics, composed by Masip (2011). The results showed a comparison between specific aspects of each evaluation: total number of reported problems, heuristics *hits*, evaluators understanding of the heuristics and; heuristics suitability with answers of severity factors. However, the work of Masip (2011) considered a heuristics set as a group of sub-heuristics. Considering sub-heuristics in a heuristics set can transform the characteristics of a heuristics set in characteristics of set of guidelines. This can impact on the advantages of performing a heuristic evaluation and require more planning, more motivation for its execution, and the complexity of undertaking evaluations.

3. Methods

3.1. Characteristics of the Research

The present study is a technological research with an exploratory goal. Every result in this study was tested in an empirical manner, observing quantitative and qualitative aspects. The quantitative dimension of this study studies involved the collection of usability problems by means of a heuristic evaluation of a specific website. The qualitative dimension refers to observations and understanding of users' perspective of the usability of websites. This was collected by means of tests with users and the "think-aloud" method. The results were collected from a case study, performing data collection through observation [Wainer 2007, Jung 2004].

All UEMs applied in this study, user tests and heuristic evaluations were conducted in one website that was chosen by the author according to the convenience of performing both UEMs inside an intranet with fast connection to the server. To mitigate the existence of bias in the results of this work, the author chose a website with close characteristics to the websites used by Petrie and Power (2012) - a website of public and governmental domain. Choosing same websites that Petrie and Power (2012) was not possible since the availability of English speakers users was out of reach.

3.2. Methodological Procedures

The aim of this study was to investigate if a set of usability heuristics focused on usability of web systems could help novice evaluators to find more usability problems than traditional usability heuristics in a HE of a web system. To compare the use of each set of heuristic, this study conducted test with real users. The results of each HE were compared to the results of tests with real users. The number of problems *hits* that the use

of each set of heuristic result was the main factor of comparison. Problems *hits* are problems found by a HE and by test with real users. This study compared the set of heuristics of Petrie and Power (2012), focused on usability of web system, with the traditional set of Nielsen and Molich [Nielsen 1994a].

To realize this comparison, eight (8) volunteers took part in the HEs. These volunteers were undergraduate students of bachelor courses in the Computer Science area with only basic experience in Human-Computer Interaction. They have no more than one year of job experience in usability area and can be considered novice evaluators. The HEs evaluated the website of the Federal University of Lavras and followed three tasks pre-defined by two usability researchers. Test with real users were conducted and its results were used to compare both HEs.

The choice of use results from tests with real users as baseline for comparison is in accordance with the literature. Eight (8) volunteers took part in test with real users. They are real users of the website of the Federal University of Lavras. All tests were performed following the same three pre-defined tasks, as in HEs.

All data from each participant was kept in confidentiality and all names kept in anonymity according to agreements done prior to any evaluation. The project was submitted for evaluation and approved by the Local Research Ethics Committee of the Federal University of Lavras, with code CAAE 17638113.4.0000.5148.

3.3. Procedures for test with real users

8 volunteers took part in the tests with users. They were real users of the Federal University of Lavras website. A room containing one computer with intranet connection to the website domain and a web-browser accessing the website was prepared for the tests. One user did the test per time. Only the respective user and a moderator remained inside the evaluation room during the tests. The moderator was responsible for conducting the *Think Aloud* technique.

The users were asked to perform the three pre-defined tasks. All their interactions with the system were recorded with specific usability software. The usability software recorded the computer monitor screen and mouse movements and clicks. In addition, with a webcam, the software recorded users video and audio. Later, two usability researchers collected usability problems analyzed all software recordings. The researchers prepared a list of all usability problems from the test with real users.

The researchers identified degrees of severity to each usability problem showed in the list from test with real users. This severity classification was done in accordance with the degrees of severity showed by Nielsen Norman Group [Nielsen 2014]:

- 1 = Cosmetic problem: just fix it if extra time is available - does not prevent users from completing their tasks in any way.
- 2 = Minor problem: low priority in a fixing list - causes minor nuisances to users, but they can easily recover to undertake their tasks.

- 3 = Major problem: important to fix and high priority in a fixing list - causes serious trouble to users, who are only able to complete their tasks after spending considerable effort.
- 4 = Catastrophic problem: must fix before the release of the product - completely prevents users from completing their tasks.

After, the researchers prepared a spreadsheet with the list of usability problems from test with users. In this spreadsheet, the severity for each usability problem was identified.

3.4. Procedures for Heuristic Evaluation (HE)

The heuristic evaluations were conducted with the participation of 8 volunteers. All these volunteers are considered novice in usability because they have neither graduate degrees nor several years of job experience in usability area, as defined by Nielsen [1992]. These volunteers have some experience in Human-Computer Interaction after having done an undergraduate course in the area.

Before the conduction of the HEs, all volunteers of HE took the same training session. The training session was about "what is" and "how to conduct" a HE. All these volunteers performed a trial HE during the training session. Thus, their performance was recorded.

Two usability researches divided the 8 HE volunteers in two groups of 4 evaluators. This division was made based on the recordings of their performance on the trial HE during the training session, to create counter-balanced groups. This counter-balance regarded the number of reported issues and previous experience in usability evaluation. With this division, one group was asked to use the traditional heuristics of Nielsen and Molich; and the other to use the heuristics of Petrie and Power (2012), especially defined based on problems encountered by users on websites. The author chose Nielsen and Molich's set because it is widely used in a number of studies reported in literature and in industry to evaluate software usability. The choice for Petrie and Power's (2012) heuristics was done because this set of heuristics covers specifically usability of websites, defined based on problems actually encountered by real users attempting to perform tasks on websites.

The heuristics of Nielsen and Molich [Nielsen 1994a] are: Nielsen and Molich are: 1) Visibility of system status; 2) Match between system and the real world; 3) User control and freedom; 4) Consistency and standards; 5) Error prevention; 6) Recognition rather than recall; 7) Flexibility and efficiency of use; 8) Aesthetic and minimalist design; 9) Help users recognize, diagnose, and recover from errors and; 10) Help and documentation.

The heuristics of Petrie and Power (2012) are: 1) Make text and interactive elements large and clear enough; 2) Make page layout clear; 3) Avoid short time-outs and display times; 4) Make key content and elements and changes to them salient; 5) Provide relevant and appropriate content; 6) Provide sufficient but not excessive content; 7) Provide clear terms, abbreviations, avoid jargon; 8) Provide clear, well-organized information structures; 9) How and why; 10) Clear labels and instructions; 11) Avoid duplication/excessive effort by users; 12) Make input formats clear and easy;

13) Provide feedback on user actions and system progress; 14) Make the sequence of interaction logical; 15) Provide a logical and complete set of options; 16) Follow conventions for interaction; 17) Provide the interactive functionality users will need and expect; 18) Indicate if links go to an external site or to another webpage; 19) Interactive and non-interactive elements should be clearly distinguished; 20) Group interactive elements clearly and logically and; 21) Provide informative error messages and error recovery.

After this division, both groups of HE volunteers conducted the HEs. All HEs happened at the same laboratory and at the same time. This laboratory had computers with intranet connection to the website domain and a web browser accessing the website. All computers had the same hardware configuration and the same operating system. For the HEs, evaluators were asked to perform the same three tasks that users were asked to perform in the tests.

The results session of the HEs provided a list of usability problems and their severities for each evaluator. The severities attributions of the HEs followed the same list of degrees used in the test with real users. All these HE lists of usability problems were gathered in one final HE spreadsheet.

3.5. Participants in the user test

Volunteers of user tests were Brazilians with mean age of 20 years and standard deviation 2. Three females and five males composed the group. All participants were undergraduate students at the same university of the website under evaluations. None of the participants of tests had participated in a usability test before. Regarding their preference of web browser, one prefers Firefox and seven prefer Google Chrome.

Users classified themselves regarding their experience in using computers. This classification was from medium to high experience in a seven-degree scale of experience, with 1 being “No experienced at all” and 7 being “Very experienced”. The median value of experience with computers was 4.5. Only one user described him/herself as not being expert using computers, choosing degree 2 in a maximum of 7.

Volunteers of the HE were all Brazilians. Their mean age was of 24 and standard deviation 1.51. Seven males and one female composed the group of 8. The group had six undergrad students and two with undergraduate degrees – one system analyst and one project coordinator. Both volunteers with undergraduate degrees had just a few years of job experience at the time. Regarding their preference of web browser, one prefers Firefox and seven prefer Google Chrome.

3.6. Website under evaluation

All tests with real users and HEs evaluated the same website. The Federal university of Lavras website was chosen for this purpose. Two usability researches choose this because its characteristics are close to the websites that were evaluated by the study of Petrie and Power [2012]. This website has numerous features and different styles. Figure 1 shows a screenshot of the website’s home page.



Figure 1. Website of Federal University of Lavras.

Only features of the website could be used during tests and HE. All volunteers were students of the Federal university of Lavras and had previous contact with the website domain.

3.7. Tasks

All tests with users and the HEs were conducted through the same three tasks. All tasks were pre-defined by two usability researchers. The researchers considered the possibility of using every heuristic from both sets to define the tasks. The first task was to find the list of all courses of Bachelor of Science offered by the university. The second was to find the document of an institutional development plan, which must be downloaded in PDF format. Finally, the third task was to find information about the location of a specific building inside the university campus.

3.8. Data analysis

This study analyzed 8 video recordings from test with real users and 8 lists of usability problems from HEs.

For the test with real users, two usability researchers coded the usability problems. This coding was done by independent analysis of video recordings. At this phase, each researcher rated the severity for each usability problem. After these first analyses, both researchers met to compare their findings and merge a result list.

Each researcher compared the result list independently, to map the occurrence of distinct problems. After this, they met to compare each list of distinct problem. They calculated a mean severity for problems that happened for more than one user. In these cases, the natural rounding was used to round the mean severity.

After these analyses, they organized the final list of usability problems from test with real users. This list was saved in a spreadsheet of results from test with users.

The same two usability researchers analyzed the results from the HEs. First, each one mapped independently the problem *hits* of the HE results spreadsheet in the spreadsheet of results from test with users. The criteria to identify whether a usability problem from a HE *hit* a usability problem in the results of test with real users were the effect that it causes in the user and the interface component where the problem is.

After this phase, both met to compare their mappings and decide the final analysis of problems *hits*. With the number of problems *hits*, the researchers could calculate the numbers of problems *misses* and *false alarms*. In consequence, they calculated the *Validity*, the *Thoroughness* and the *Effectiveness* of the use of each heuristic set.

4. Results

This section contains results from the user test sessions and heuristic evaluations. The analyzed data are described in this section. This analysis contemplated descriptive and inferential statistics and interpretation of tables. The organization of this section was outlined according to the kind of evaluation (test with users and HE). It is followed by the analysis of the usability problems *hits* among results from each HE and test with users, and the analysis of the severity ratings of problems reported by test with users and HEs.

4.1. Results from test with real users

A total of 212 usability problems was found by applying the test with real users. Two usability researchers did analysis of multiple occurrences and a total of 126 instances of distinct usability problems were found.

4.2. Results from Heuristic Evaluations

The HEs led to a total of 60 identified usability problems. Among this number, novice evaluators that used the heuristics of Nielsen and Molich detected 33 problems, and novice evaluators that used the heuristics of Petrie and Power found 27 problems.

Table 2 shows the number of usability problems detected by each evaluator. Column "Evaluator" has all novice evaluators that took part in HEs and column "Number of reported problems" has the number of reported problems according to each evaluator and to the use of each different set of heuristic. Evaluators 01, 02, 03 and 04 used Nielsen and Molich's heuristics. The row "TOTAL" shows the number of usability problems identified by all evaluators together.

Table 2. Number of usability problems detected by each evaluator.

Evaluator	Number of reported problems
Using heuristics of Nielsen and Molich	
01	6
02	10
03	7
04	10
Using heuristics of Petrie and Power	
05	5
06	9
07	7
08	6
TOTAL	60

This study checked the number of citations of each heuristic in the different groups. At least one heuristic had to be cited by evaluators to identify a usability problem in the HE report. Various usability problems could be identified using the same heuristic.

Novice evaluators who used the heuristics of Nielsen and Molich cited 9 of the 10 heuristic in the HE report to identify the usability problems. Novice evaluators who used the heuristic of Petrie and Power cited 12 of the 21 heuristics to identify the usability problems in the HE report. Figure 2 shows the total number of citations for each one of the heuristics of Nielsen and Molich, according to the report from HEs that novice evaluators used Nielsen and Molich's heuristic set. Figure 3 shows the total number of citations for each one of the heuristics of Petrie and Power, according to the report from HEs that novice evaluators used Petrie and Power's heuristic set.

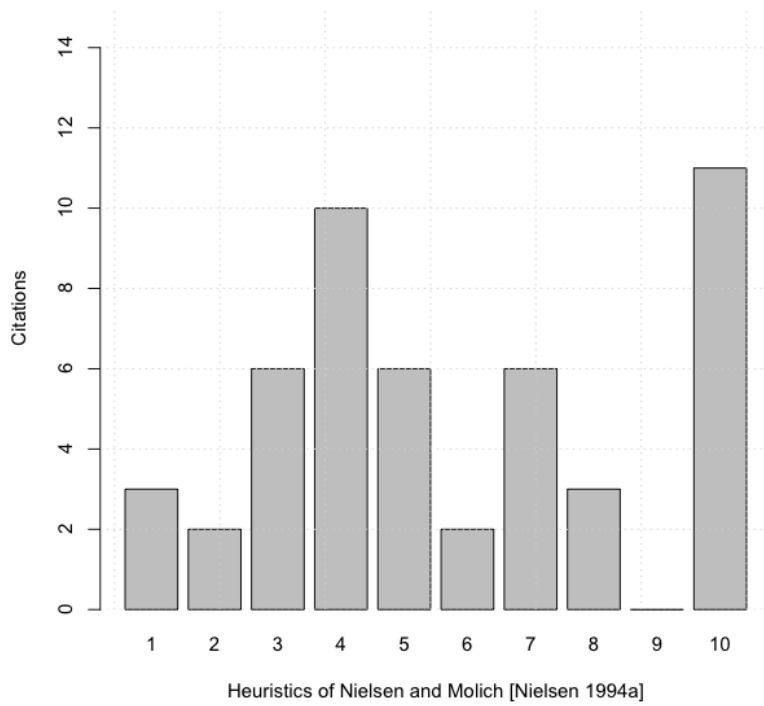


Figure 2. Number of citations of each heuristic of Nielsen and Molich among novice evaluators that used it in HE.

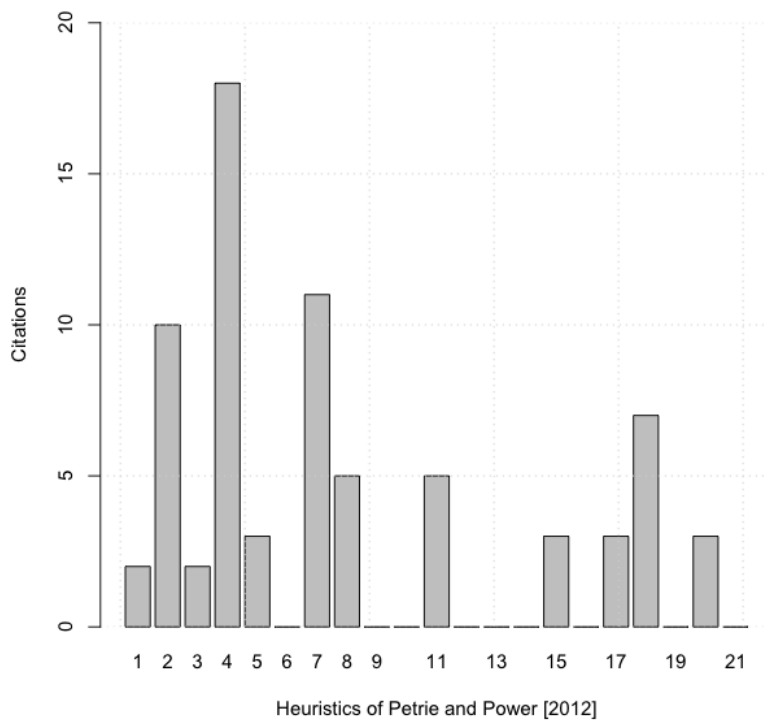


Figure 3. Number of citations of each heuristic of Petrie and Power among novice evaluators that used it in HE.

4.3. Hits of usability problems

This section shows results of problems *hits* analysis. Initial analysis of data showed that novice evaluators had difficulties in describing distinct usability problems. Usability problems that novice evaluators describe inside the HE reports *hit* more than one usability problem in the test with real users results.

As an example, one of the novice evaluators described the following problem:

“A not intuitive way to users that do not know the university; redirection to another website; it is necessary to download a document in another website to find information that is important to new students (the directions); users will have difficulty to remember where he/she found the information; user must browse through much information that is not related to what he/she is looking for; there is no help option to the user.”

This example shows that novice evaluators were including more than one usability problem inside the same the description of only one usability problem.

The analysis of problems *hits* showed that, in this specific case, the usability problem described hits 3 of the usability problems found by test with real users. For this reason, this kind of problem was named a “swollen problem”. The usability problems from tests with users related to this problem are: “*the website does not provide help to users*”, “*user tells that the way to get the information is difficult to discover*”, and “*user showed had difficulty with the large amount of other information that are not related to what he/she is looking for*”.

The researchers analyzed the occurrence of swollen problems and identified 102 usability problems among all the 60 usability problems in HEs’ reports. This analysis considers the real number of problems that exist inside each one of the 60 usability problems listed in the HE report. Results show that only 30 of the initial 60 usability problems were not swollen problems.

Sixteen (16) of the 33 usability problems reported by novice evaluators using heuristics of Nielsen and Molich were not swollen problems. The 17 swollen problems of this group represented 43 usability problems.

Fourteen (14) of the 27 usability problems that novice evaluators using heuristics of Petrie and Power reported were not swollen problems. The 15 swollen problems of this group represented 29 usability problems.

4.3.1. HE using the heuristics of Nielsen and Molich

The results show that the use of the heuristics of Nielsen and Molich during HEs led to 27 *hits*. Regarding the definitions of Hartson et al. (2001), the Validity, Thoroughness and Effectiveness of the performance of HE conducted by novice evaluators using the heuristics of Nielsen and Molich were:

$$\text{Validity} = 27 / (27 + 32) = \mathbf{0.458}$$

$$\text{Thoroughness} = 27 / (27 + 99) = \mathbf{0.214}$$

$$\text{Effectiveness} = 0.458 * 0.214 = \mathbf{0.098}$$

Figure 4 shows a Venn diagram containing the number of hits, false alarms and misses of usability problems considering the HE that novice evaluators used the heuristic set of Nielsen and Molich.

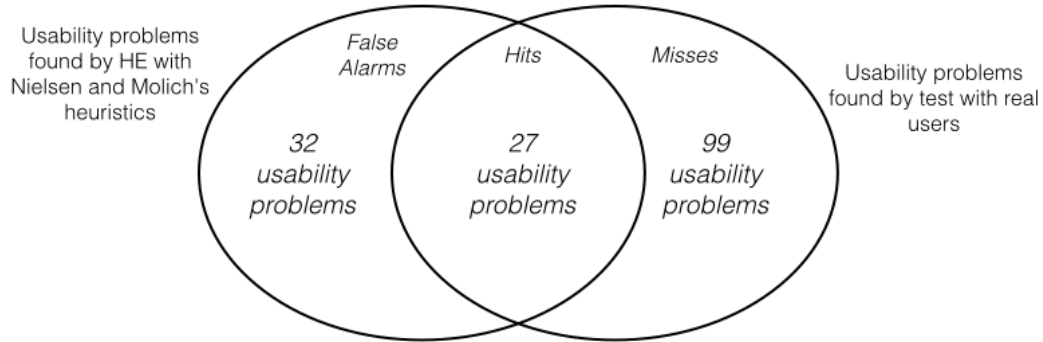


Figure 4. Venn diagram of problems hits, false alarms and misses of HE using heuristics of Nielsen and Molich.

4.3.2. HE using heuristics of Petrie and Power

The results show that the use of the heuristics of Petrie and Power during HEs led to 11 *hits*. Regarding the definitions of Hartson et al. (2001), the Validity, Thoroughness and Effectiveness of the performance of HE conducted by novice evaluators using the heuristics of Petrie and Power were:

$$\text{Validity} = 11 / (11 + 32) = \mathbf{0.256}$$

$$\text{Thoroughness} = 11 / (11 + 115) = \mathbf{0.087}$$

$$\text{Effectiveness} = 0.256 * 0.087 = \mathbf{0.022}$$

Figure 5 shows a Venn diagram containing the number of hits, false alarms and misses of usability problems considering the HE that novice evaluators used the heuristic set of Petrie and Power.

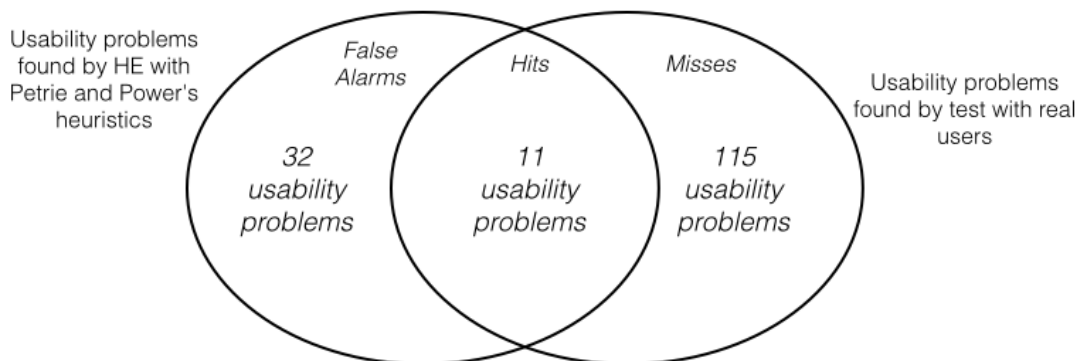


Figure 5. Venn diagram of problems hits, false alarms and misses of HE using heuristics of Petrie and Power.

4.3.3. Common Hits for HEs with both heuristic sets

Results showed that 4 distinct usability problems were found by novice evaluators using heuristics of Nielsen and Molich and by novice evaluators using heuristics of Petrie and Power.

4.4. Severity ratings of problems reported by users and novice evaluators

Comparison between severity ratings from HEs and from tests with real users needed to be done regarding the occurrence of swollen problems. As in swollen problems novice evaluators could not distinguish the distinct usability problems, a unique severity rating was assigned for each swollen problem. Thus, the severity of a swollen problem was compared to the severity of more than one usability problem in test with real users.

As an example, one novice evaluator reported a swollen problem and rated severity 2 to it. This swollen problem hits 2 usability problems in test with real users results. One of the usability problems that this swollen problem hits has severity 2 in test with real users, and the other has severity 3. In the first case of hit, no difference was observed between both severity ratings (the severity in HE report and the severity from test with real users). In the second case, a difference of 1 degree of severity was observed.

The results of comparisons of difference between severity ratings are shown in Table 4. A difference in severity rating is calculated as severity rating of a problem reported in HE reports minus the severity rating of its hits in test with real users. Table 4 shows the number of occurrence of the 7 possible differences, from -3 to +3. The row “Nielsen and Molich” shows only data of evaluators that used the heuristic set of Nielsen and Molich, and the row “Petrie and Power” shows only data of evaluators that used the heuristic set of Petrie and Power.

Table 3. Difference of severities from swollen problems to severity of problems in test with real users.

Heuristic set	Differences						
	-3	-2	-1	0	+1	+2	+3
Nielsen and Molich	0	4	7	8	4	4	0
Petrie and Power	0	0	3	5	1	2	0

5. Discussions

The aim of this study was to investigate if a set of heuristic focused on usability of web systems can help novice evaluators to find more usability problems than traditional heuristics in a HE of a web system. This section discusses the main results of this study considering the performance of novice evaluators during all HEs.

5.1. Number of reported problems

Considering the reports of novice evaluators using both different kinds of heuristics, novice evaluators using the heuristics of Nielsen and Molich only reported 5 usability

problems more than novice evaluators using the heuristics of Petrie and Power. Considering the occurrence of swollen problems and the real number of usability problems, novice evaluators using the heuristics of Nielsen and Molich only found 16 usability problems more than novice evaluators using the heuristics of Petrie and Power. As both groups of novice evaluators had the same time to conduct HE, the use of Nielsen and Molich's heuristics helped novice evaluators to identify more usability problems.

5.2. Citation of heuristics

Each one of the different heuristic sets has a different number of heuristics. The set of Nielsen and Molich has 10 heuristics and the set of Petrie and Power has 21. The results of this study show evidences that novice evaluators have more facilities to consider a higher percentage of heuristics of Nielsen and Molich than the heuristics of Petrie and Power. Future studies can compare both sets of heuristic in order to find equivalence between heuristics. Not only the number of heuristics can be the cause of this difference, evaluators' previous experience with Nielsen and Molich's heuristics can be another reason for this difference in percentage of used heuristics. As they were novice at HE, the few experience they had with HE can be determinant in results. The evaluators had previous experience with Nielsen and Molich heuristics and not with Petrie and Power heuristics. Future studies can replicate the comparisons of this study considering novice evaluators with previous experience with Petrie and Power's heuristics.

5.3. Occurrence of Swollen problems

Fifty percent (50%) of the usability problems reported in all HEs were swollen problems. Seventeen (17) of the 33 usability problems in HE with the heuristics of Nielsen and Molich were swollen problems, and 14 of the 27 usability problems in HE with the heuristics of Petrie and Power were swollen problems. Wilcoxon signed rank tests showed that there was a significant occurrence of swollen problems in the HEs with heuristics of Nielsen and Molich ($V = 153$, $p\text{-value} = 0.0001952$) and with heuristic of Petrie and Power ($V = 91$, $p\text{-value} = 0.0005371$). The Wilcoxon signed rank test was chosen because the distribution of the sample was non-parametric. This evidence shows that novice evaluators have difficulties in describing distinct usability problems, which can cause serious problems when the results of such evaluations are used by development teams. The traceability of such problems can be seriously compromised, given that developers can report having corrected an entire problem, but may actually have only done so for one of the problems contained in the "swollen problem".

This finding should be confirmed by further studies, and the training session must occur under controlled conditions to verify if improving training sessions could help alleviate this problem with novice evaluators. However, the findings in this study point to important implications that should be further investigated and that can have important impact on how evaluators are trained and how their results should be taken into consideration.

5.4. Validity, Thoroughness and Effectiveness

The results of this study show that novice evaluators have a significantly higher number of problems *hits* (Chi-square = 7.94, df = 1, p-value < 0.01) and unique *hits* (Chi-square = 9.69, df = 1, p-value < 0.01) using the heuristics of Nielsen and Molich. The number of false alarms was the same for HE with Nielsen and Molich heuristics and HE with Petrie and Power heuristics.

Novice evaluators that used the heuristics of Nielsen and Molich achieved higher levels of Validity, Thoroughness and Effectiveness as well. This evidence suggests that novice evaluators may have better performance in HE of web systems using the traditional set of heuristics. Future studies should investigate the factors that influenced these results of performance, such as familiarity and lack of experience with more specific issues related to specific domains and technologies.

Novice evaluators that took part in this study achieved a higher Validity in performance using heuristics of Nielsen and Molich than the evaluators of Hvannberg et al. (2007) using the same set of heuristics. However, novice evaluators had higher Thoroughness and Effectiveness in Hvannberg et al. (2007) with the use of Nielsen and Molich heuristics. In the study conducted by Hvannberg et al. (2007), novice evaluators used paper tools and a specific software tool during HE, whilst in the present study they only used spreadsheet software. Another difference between both studies is that the numbers of Hvannberg et al. (2007) refer to HE conducted by 10 novice evaluators, whilst in this study only 4 novice evaluators took part in HE with heuristics of Nielsen and Molich.

5.5. Difference in severity ratings

The results of this study calculated the difference between evaluator's severity ratings and severity ratings from test with real users. Either for novice evaluators using heuristics of Nielsen and Molich or for novice evaluators using the heuristics of Petrie and Power, no difference was observed in the major part of problems. Difference of -1 was the second more common difference attested for both cases. The evidences from these results are not sufficient to ensure that novice evaluators have different severity ratings than those of problems encountered by users. If the use of novice evaluators are in discussion, this probability of different severities must be take into account within cost/benefits analysis.

6. Conclusion

The results of this study showed evidences that novice evaluators have difficulties in reporting distinct usability problems inside HE reports. Novice evaluators find more usability problems; and perform more problems *hits* and unique *hits* using the heuristic set of Nielsen and Molich than using the heuristic set of Petrie and Power. The Validity, Thoroughness and Effectiveness of HE conducted by novice evaluators using the heuristics of Nielsen and Molich were better than the same measures from HE conducted by novice evaluators using the heuristics of Petrie and Power.

The findings of this study show that novice evaluators initially can have better performance using traditional heuristics sets such as those proposed by Nielsen and Molich, when conducting HEs in web systems.

Development teams must consider that novice evaluators tend to report swollen problems. Besides the difficulties with identifying and tracing problems with new requirements in development cycles, having several problems reported as one also makes it difficult to interpret the severity ratings attributed to problems, since each of the problems reported as one could have a different severity rating. This is an important implication for the training of novice evaluators. Training sessions must address this fact and make it clearer to evaluators how to report problems individually. However, further work is needed to confirm this finding, since this kind of problem can persist and occur even with improved training sessions.

In conclusion, to consider using novice evaluators in HE of web systems, a structured analysis must be done. Novice evaluators are less expensive to recruit, but they produce reports with less quality. Recruiting novice evaluators mean that a HE can result in many differences on severity ratings compared to the severity of problems users may actually encounter. Besides, it will require more training sessions, and training costs must be included in the cost/benefit analysis of having or not novice evaluators in a HE. Another important factor to consider is the number of evaluators that will participate on the HE and which heuristics set to use.

The goal of this study was to help further the knowledge in literature about employment of novice evaluators in HE, and the study reported interesting characteristics of this employment, that can help deeper the understanding of the issues encountered and foster new research in the area. However, more studies must be done to confirm the findings. Such studies should include comparisons of this one and investigate the same characteristics with expert evaluators. In addition, more studies can verify the same results among other varieties of ubiquitous technologies as mobile devices, smart watches and smart televisions. Future studies should consider investigating the novice evaluator performance in different heuristic evaluation methods, such as the investigate the difference between traditional heuristic evaluation and collaborative heuristic evaluation [Petrie and Buykx 2010].

7. Acknowledges

The author thanks all the participants who took part in this study for their valuable contribution, the ALCANCE research group and his supervisor for the great help with this study.

8. References

- Baranauskas, M. C. C. and Rocha, H. d. (2003) "Design e avaliação de interfaces humano-computador." *Campinas-SP: Nied/Unicamp*, pages: 244.
- Barua, A. and Mukherjee, R. (2012) "Measuring the business impacts of effective data." Retrieved on September 15, 2014, from: [http://www.sybase.com/files/White Papers](http://www.sybase.com/files/White_Papers)

- Bastien, J. (2010) "Usability testing: a review of some methodological and technical aspects of the method." *International Journal of Medical Informatics*, 79(4) pages e18–e23.
- Botella, F., Alarcon, E., and Peñalver, A. (2013) "A new proposal for improving heuristic evaluation reports performed by novice evaluators." In *Proceedings of the 2013 Chilean Conference on Human-Computer Interaction*, pages 72–75. ACM.
- de Lima Salgado, A. and Freire, A. P. (2014) "Heuristic evaluation of mobile usability: A mapping study." In *Proceedings of the 16th International Conference on Human-Computer Interaction, Part III, Lecture Notes in Computer Science*, pages 178–188. Springer.
- Dumas, J. S. and Redish, J. (1999) "*A practical guide to usability testing.*" Intellect Books, Portland OR, USA, pages: 404.
- Fernandez, A., Abrahão, S., and Insfran, E. (2013) "Empirical validation of a usability inspection method for model-driven Web development." *Journal of Systems and Software* 86.1, pages 161-186.
- Fernandez, A., Insfran, E., and Abrahão, S. (2011) "Usability evaluation methods for the web: A systematic mapping study." *Information and Software Technology*, 53(8), pages 789–817.
- Gerhardt-Powals, J. (1996) "Cognitive engineering principles for enhancing human-computer performance", *International Journal of Human-Computer Interaction* 8, 189–211.
- Gray, W. D. and Salzman, M. C. (1998) "Damaged merchandise? a review of experiments that compare usability evaluation methods." *Human-Computer Interaction*, 13(3), pages 203–261.
- Harrison, C. and Petrie, H. (2007) "Severity of usability and accessibility problems in ecommerce and egovernment websites." In *In People and Computers XX—Engage*, pages 255–262. Springer.
- Hartson, H. Rex, Andre, T. S., and Williges, R. C. (2001) "Criteria for evaluating usability evaluation methods." *International journal of human-computer interaction* 13.4, pages 373-410.
- Hertzum, M. and Jacobsen, N. E. (2003) "The evaluator effect: A chilling fact about usability evaluation methods." *International journal of human-computer interaction*, 15(1), pages 183–204.
- Hertzum, M., Molich, R., and Jacobsen, N. E. (2014) "What you get is what you see: revisiting the evaluator effect in usability tests." *Behaviour & Information Technology*, 33(2), pages 144–162.
- Howarth, J., Smith-Jackson, T., and Hartson, R. (2009) "Supporting novice usability practitioners with usability engineering tools." *International Journal of Human-Computer Studies*, 67(6), pages 533–549.

- Hvannberg, Thora, E., Law, E. L., and Lérusdóttir, M. K. (2007) "Heuristic evaluation: Comparing ways of finding and reporting usability problems." *Interacting with computers* 19.2, pages 225-240.
- ISO, I. (2011) "Iec 25010: 2011: Systems and software engineering—systems and software quality requirements and evaluation (square)—system and software quality models." *International Organization for Standardization*. Pages: 34
- ISO, WD. (1998) "9241-11. Ergonomic requirements for office work with visual display terminals (VDTs)." *The international organization for standardization*. Pages: 22
- Jacobsen, N. E., Hertzum, M., and John, B. E. (1998) "The evaluator effect in usability tests." In *CHI 98 Conference Summary on Human Factors in Computing Systems*, pages 255–256. ACM.
- Jung, C. F. (2004) "Metodologia para pesquisa e desenvolvimento: aplicada a novas tecnologias, produtos e processos." Axcel Books, Rio de Janeiro, pages: 312.
- Lanzilotti, R., Ardito, C., Costabile, M. F., and De Angeli, A. (2011) "Do patterns help novice evaluators? a comparative study." *International journal of human-computer studies*, 69(1), pages 52–69.
- Ling, C. and Salvendy, G. (2009) "Effect of evaluators' cognitive style on heuristic evaluation: Field dependent and field independent evaluators." *International journal of human-computer studies*, 67(4), pages 382–393.
- Mayhew, D. J. (1999) "The usability engineering lifecycle." In *CHI'99 Extended Abstracts on Human Factors in Computing Systems*, pages 147–148. ACM.
- Nielsen, J. (1992) "Finding usability problems through heuristic evaluation." In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 373–380. ACM.
- Nielsen, J. (1994a) "Enhancing the explanatory power of usability heuristics." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 152–158. ACM.
- Nielsen, J. (1994b) "Usability inspection methods." In *Conference companion on Human factors in computing systems*, pages 413–414. ACM.
- Nielsen, J. (1994c) "Usability inspection methods." In *Conference companion on Human factors in computing systems*, pages 413–414. ACM.
- Nielsen, J., Clemmensen, T., and Yssing, C. (2002) "Getting access to what goes on in people's heads?: reflections on the think-aloud technique." In *Proceedings of the second Nordic conference on Human-computer interaction*, pages 101–110. ACM.
- Nielsen, J. and Molich, R. (1990) "Heuristic evaluation of user interfaces." In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 249–256. ACM.
- Petrie, H. and Buykx, L. (2010) "Collaborative heuristic evaluation: improving the effectiveness of heuristic evaluation." In *Proceedings of UPA 2010 International Conference*. Omnipress.

- Petrie, H. and Kheir, O. (2007) “The relationship between accessibility and usability of websites.” In Proceedings of the SIGCHI conference on Human factors in computing systems, pages 397–406. ACM.
- Petrie, H. and Power, C. (2012) “What do users really care about?: a comparison of usability problems found by users and experts on highly interactive websites.” In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2107–2116. ACM.
- Plaisant, C. and Shneiderman, B. (2010) “*Designing the user interface: strategies for effective human-computer interaction.*” Addison-Wesley Publ. Co., Reading, pages: 624.
- Rogers, Y., Sharp, H., and Preece, J. (2011) “*Interaction Design: Beyond Human - Computer Interaction.*” Wiley, pages: 602.
- Sauro, J. and Lewis, J. R. (2012) “*Quantifying the user experience: Practical statistics for user research.*” Elsevier, pages: 312.
- Torrente, M., Prieto, A., Gutiérrez, D. A., and De Sagastegui, M. (2013) “Sirius: A heuristic-based framework for measuring web usability adapted to the type of website.” *Journal of Systems and Software*, 86(3), pages 649–663.
- Wainer, J. (2007) “*Atualização em informática.*” Chapter Métodos de pesquisa quantitativa e qualitativa para a Ciência da Computação, pages 221–262. PUC-Rio.