A Method for Classifying Usability Findings to Enhance Validation of New Heuristics

André de Lima Salgado[1,2], Renata Pontin de Mattos Fortes[1], Patrick C. K. Hung[2] and Dilvan de Abreu Moreira[1]

[1]ICMC, University of São Paulo, São Carlos, SP, Brazil

[2]University of Ontario Institute of Technology, Oshawa, ON, Canada

alsalgado@usp.br, renata@icmc.usp.br, Patrick.Hung@uoit.ca, dilvan@gmail.com

Abstract. *Different usability heuristics have been proposed as new application domains arise. Such proposals usually depend on the validation of the new heuristics. However, current validation methods are still biased by subjective comparisons of usability findings. In this paper, we aimed to enhance the process of matching usability finding descriptions and mitigate the bias of such process. To reach our goal, we adopted ontology techniques to extend the User Action Framework for the context of validating new usability heuristics. We tested three hypotheses about the feasibility of our new framework based on a case study with 173 usability findings. These usability findings were retrieved from an online project of a private mobile browser. Our data analysis of supported merging three classification schemes for our framework: User Action Framework, Typical Usability Defects (from ISO) and the heuristics of Nielsen. Finally, we describe a logical process for our method, because some of the contents from the classification schemes are not disjoint.*

Keywords: *Usability. Heuristic Evaluation. Validation. Classification. Heuristic Evaluation.*

1. Introduction

In formative usability field, Usability Evaluation Methods (UEMs) are methods for diagnostic of usability findings. Usability findings are "*identified usability defect and/or usability problem or positive usability-related attribute*" [ISO/IEC 25066 2016], and such a diagnostic is necessary to identify ways to enhance the usability of an interface [Lewis 2014]. Heuristic Evaluation (HE), Cognitive Walkthrough and Usability Testing are popular among formative UEMs [Preece *et al.* 2015].

Constantly, new computing technologies are proposed and influence new paradigms of user interfaces, to be employed among different domains. In consequence, the literature on formative usability has to keep UEMs up to date with such new paradigms and domains [Hermawati and Lawson 2016]. For this reason, adapting UEMs, or proposing new methods, are important tasks to move forward in the field. However, since the important discussions of Hornbæk (2010) reinforcing the need for more appropriate methods of assessment of UEMs, a little has been done in the topic

(such fact could be verified by checking the citations of Hornbæk's article through Google Scholar search engine).

One of the main challenges about assessing different UEMs regards to the process of matching usability finding descriptions from different UEM reports [Hornbæk 2010; Yusop 2017]. UEM reports are the main outcomes of a UEM, they are often composed by usability finding descriptions. Matching usability findings is usually based on comparison of usability finding descriptions. In this context, individual differences among usability professionals have its impacts on the descriptions of usability findings, which become dynamic and cannot be known a priori, a characteristic of open world [Bendale and Boult 2015; Hertzum *et al.* 2014 ; Araujo 2017]. Therefore, it is plausible to understand that usability finding classifications may enhance the process of matching. The User Action Framework (UAF), the Classification of Usability Problems (CUP), the Root Cause Defect Analysis (RCA), the Orthogonal Defect Classification (ODC) and the Usability-Error Ontology (UEO) are examples of usability finding classifications [Vilbergsdottir *et al.* 2014; Elkin *et al.* 2013]. Among such classifications, the UEO is the only domain specific classification, focused on health informatics.

Although different classifications have been proposed in the literature, there is no widely adopted classification to describe usability findings [Hornbæk 2010; Yusop *et al.* 2017]. Yusop *et al.* (2017) argued that the Human-Computer Interaction (HCI) community should develop a more comprehensive and agreed classification, which remains as a gap in the literature. In this regard, we understand that UEM focused classifications (domain focused classifications) are appropriate to fill out this gap through a divide and conquer approach. For example, describing a usability finding classification to support comparisons of HE approaches (e.g. employment of new sets of heuristics) can help to fill out such a gap for such a popular domain. Nevertheless, the following question remains to be answered:

> ***Research Question****: How to classify usability findings to enhance the matching process for validation of new heuristics?*

The goal of our study was to answer this question. The goal of this study can also be understood as to create a usability finding classification to enhance the matching process for future researches on HE. We decided to achieve this goal by creating an extension for the UAF. We adopted the UAF as basis for our UAF-HE classification, because it is domain-free, and it is the classification with most recent and relevant case studies [Vilbergsdottir *et al.* 2014]. In addition, we increased the expressiveness of the UAF with the Typical Usability Defects (TUD) from ISO/IEC 25066 (2016) and the standard usability heuristics of Nielsen (1994). We named our proposal as the *User Action Framework for the Heuristic Evaluation Domain* (UAF-HE). Because we were not sure about the contribution of additional classificatory schemes to the UAF, we elaborated the following hypothesis for this study:

> **H0:** *(Null) Employing the UAF-HE cannot enhance the matching process in comparison to employing only the UAF.*

H1: *Adding the TUD classification to the UAF can enhance the matching process in comparison to employing only the UAF.*

H2: *Employing the UAF-HE can enhance the matching process in comparison to employing only the UAF.*

We aimed to reject the null hypothesis (H0) and accept the others (H1 and H2). To test the hypothesis, we conducted a case study classifying 173 usability findings, retrieving issues from the online project (GitHub) of Firefox Focus browser (a private browser). We chose the usable privacy context as the scope of this study because the interest in such a field had a rapid development during the past two decades [Kawakani *et al.* 2017; Still 2016; De and Zezschwitz 2016; Garfinkel and Lipford 2014; Cranor and Buchler 2014]. In addition, the online project of Firefox Focus gathers contributors from all over the world, which reinforce the open world characteristic of their project.

The most part of the 173 usability findings was classified in the translation content of the UAF (T-C2), in the "Insufficient and/or poor information on the user interface" type of usability defect (TUD3) and in the heuristic "Aesthetic and minimalist design" (H8). In addition, the data analysis of our study supported the acceptation of H1 and H2, and the rejection of H0. This article presents discussions for each of these findings.

The remaining of this article is organized as follows: the *Background* for our study (Section 2), the description of the *User Action Framework for the Heuristic Evaluation Domain (UAF-HE)* (Section 3), the *Methods and Materials* (Section 4), the *Results and Discussion* (Section 5), the *Implications for Design* (Section 6) and the *Conclusions* (Section 7).

2. Background

This section presents the literature review of our study. We reviewed the main topics related to our research. Therefore, we present a review about *Formative Usability* (Section 2.1), *Comparing Usability Evaluation Methods* (Section 2.2) and *Matching Usability Finding Descriptions* (Section 2.3).

2.1 Formative Usability

Usability is one of the aspects of software quality and ergonomics [ISO/IEC 25066 2016; ISO/TR 9241-100 2010]. It can be understood 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*" [ISO/IEC 25066 2016]. However, usability can also be understood in two main concepts: formative and summative usability [Lewis 2014]. Formative usability focus on diagnostic of usability findings and how to mitigate the impact of such findings, while summative usability focuses on measuring usability through metrics towards a defined goal [Lewis 2014]. For each concept, the literature shows evaluation methods that can be classified as inspection-based or user-based evaluations. Inspection methods usually do not require the participation of end users, while user-based evaluation does [ISO/IEC 25066 2016].

Usability Evaluation Methods (UEMs) are methods to evaluate the usability of an interface. Such UEMs can be grouped according to formative or summative usability. Because this study is about formative usability, we focus on formative UEMs. In this context, Heuristic Evaluation (HE), Cognitive Walkthrough and Usability Testing are popular among formative UEMs [Preece *et al.* 2015].

UEMs usually produce usability reports, which are a collection of usability findings. Usability findings are "*identified usability defect and/or usability problem or positive usability-related attribute*". For instance, usability problems are situations that result in decrease of any usability-related attribute, while usability defects are product attributes that leads to "*a mismatch between user intentions and/or user actions and the system attributes and behaviour*". Therefore, usability defects are often a diagnostic of inspection-based methods, while usability problems are often a diagnostic of user-based evaluations [ISO/IEC 25066 2016].

The differences between usability problems and usability defects are often not a relevant issue in the literature about usability finding classifications. Usually, usability-finding classifications are proposed for the broad variety of UEMs [Yusop 2017]. For this reason, we assume usability finding, usability problems and usability defect as synonym during the literature review of this article. Nevertheless, the contributions are focused on reports from HEs (an inspection-based UEM). Therefore, our conclusions aimed reports of usability defect.

2.2. Comparing Usability Evaluation Methods

Hartson et al. (2001) reviewed different measures adopted in the literature in assessment of UEMs. According to them, the ultimate criteria for assessing the effectiveness of different UEMs should be comparing a set of real usability findings with the set of usability findings found by the UEM being assessed. Hartson et al. (2001) showed that such comparison could be performed through distinct means, as:

- **Comparison against a standard list of usability findings.** This method adopts the premise that a list of all usability findings of an interface exists and is known. Thus, the outcomes from the UEM assessed can be compared to such a list. Hartson et al. (2001) show that traditional user-based evaluations conducted in laboratory are usually accepted as a gold standard.
- **Determining the realness of usability findings by expert review and judgment.** For this method, usability experts review the list of usability findings originated from the assessed UEM in order to judge the realness of each finding.
- **Determining the realness of usability findings by end-users review and judgment.** For this approach, a sample of end-users review and judge the realness of each usability finding reported from a UEM.

According to Hartson et al. (2001), the literature commonly compares different UEMs based on the measures: *Reliability*, *Thoroughness*, and *Validity*. *Reliability* is the consistency of UEM outcomes, independent of evaluator or expertise effect. *Thoroughness* is how close outcomes of a UEM are to a standard set of usability

findings. Finally, *Validity* is how correct are the outcomes of a UEM, also evaluating its realness.

In addition to such measures, Hartson *et al.* (2001) proposed the *Effectiveness*: a combination of *Thoroughness* and *Validity*. Such measures are based on the following metrics:

- **Hits:** usability findings reported by the assessed UEM that exist in the standard set of usability findings.
- **Misses:** usability findings not reported by the UEM that exist in the standard set of usability findings.
- **False alarms:** usability findings reported by the UEM that do not exist in the standard set of usability findings.

Considering these metrics, the formulas for *Thoroughness*, *Validity* and *Effectiveness* [Hartson et al. 2001, p. 390-394] are as follows:

$$Validity = hits / (hits + false\ alarms)$$

$$Thoroughness = hits / benchmark\ set$$

$$Effectiveness = Validity * Thoroughness$$

Hartson et al. (2001, p. 394) argued that *Validity* and *Thoroughness* have a preference over other measures. For this reason, they described a weighted combination of *Validity* and *Thoroughness*, called *F-measure*, adapting it from Manning *et al.* (1999). The formula for the *F-measure* is based on a α value, as follows:

$$F\text{-}measure = 1 / (\alpha * (1 / Validity) + (1 - \alpha) * (1 / Thoroughness))$$

Considering the *F-measure* formula, and for a $\alpha = 0.5$, both *Validity* and *Thoroughness* receive the same weight. Thus, the *F-measure* formula could be described as the following [Hartson et al. 2001, p. 394]:

$$F\text{-}measure = 2 * Validity * Thoroughness / (Validity + Thoroughness)$$

As described previously, *hits* and *misses* are basis to calculate the *Validity*, the *Thoroughness*, the *Effectiveness* and the *F-measure*. These metrics indicate similarity between usability findings resulted by a specified UEM and a standard UEM method. Therefore, identifying *hits* and *misses* is a valuable task that requires attention. The following section shows methods for matching similar usability findings in order to identify finding *hits* and *misses*.

2.3. Matching Usability Finding Descriptions

To calculate the number of finding *hits* or *misses*, practitioners must identify which usability findings can be considered similar or not. The process of identifying such similarity is called matching usability findings. In such a topic, Hornbæk and Frøkjær (2008) reviewed four (4) popular methods, as described following:

1. **Similar changes:** findings that implicate in similar changes of the interface should be considered as similar.

2. **Practical prioritization:** practitioners are asked to prepare a prioritized list of usability findings. For such list, findings with similar prioritization can be considered as similar.

3. **The model of Lavery *et al.* (1997):** usability finding descriptions are organized in four (4) categories: cause, breakdown, outcome and change. Such categories can be used to compare similarity of usability findings.

4. **User Action Framework (UAF):** usability findings are structured according to the seven stages of actions, from Norman (2013), describing whether a finding relates to the *planning*, *translation*, *physical actions*, *outcome* and *assessment* categories. Such categories are cyclical, after the *assessment* category comes the *planning category* again. The UAF allows practitioners to compare similarity of findings based on a comparison of categories [Yusop *et al.* 2017; Vilbergsdottir *et al.* 2014; Hartson and Pyla 2012]. Its goal is to guide the interaction design.

The literature also presents the Classification of Usability Problems (CUP), the Root Cause Defect Analysis (RCA), the Orthogonal Defect Classification (ODC) and the Usability-Error Ontology (UEO). The CUP, the ODC and the RCA aim provide a feedback for developers in order to help them to correct such findings. Such classifications are domain-free and provide information as the trigger (or root cause) of a usability finding. On the other hand, the UEO is the unique domain specific classification among these classifications. The UEO is an ontology focused on health systems, created after a survey with professionals of the field [Elkin *et al.* 2013]. The UEO may be appropriate to enhance the process of matching usability finding descriptions, but we can speculate it only for the domain of health systems. Although such a variety of classifications exists in the literature, the UAF stands among the few classifications that are domain-free and was approached in relevant recent case studies [Vilbergsdottir *et al.* 2014; Yusop 2017].

Finally, Petrie and Buykx (2010) and Petrie and Power (2012) adopted the following criteria for matching usability finding descriptions:

- **Relaxed matching criteria:** usability findings are considered similar if they refer to the same finding, or to the same design element, independent of the level of abstraction. If the same underlying finding is described, two usability findings are considered as similar.
- **Strict matching criteria:** usability findings are considered similar only if they refer to the same finding, to the same element of design, and the description is at the same level of abstraction.

The *strict* and *relaxed* criteria are similar to the matching process *similar change* [Hornbæk and Frøkjær 2008]. However, analyzing data with *strict* and *relaxed* criteria highlights two distinct levels of similarity, instead of only one as in the *similar change*.

The next section presents the UAF-HE classification framework.

3. The *User Action Framework for the Heuristic Evaluation Domain* (UAF-HE)

The *User Action Framework for the Heuristic Evaluation Domain* (UAF-HE) aims to classify usability findings from different approaches of HE considering characteristics from the open world (e.g. online communities). We created the UAF-HE classification as an extension for the UAF framework to help methods of matching usability findings in the literature about HE.

We adopted the UAF as basis for our classification because it is domain-free, and it is the classification with most recent and relevant case studies [Vilbergsdottir *et al.* 2014]. In addition to the UAF, we created a classification based on the Typical Usability Defects (TUD), from the ISO/IEC 25066 (2016), and on the standard usability heuristics of Nielsen (1994).

We used Protégé to generate an illustrative graph of the relations among the three main classes of our classification, as shown at Figure 1. The code of the first version of UAF-HE ontology is available online[a].

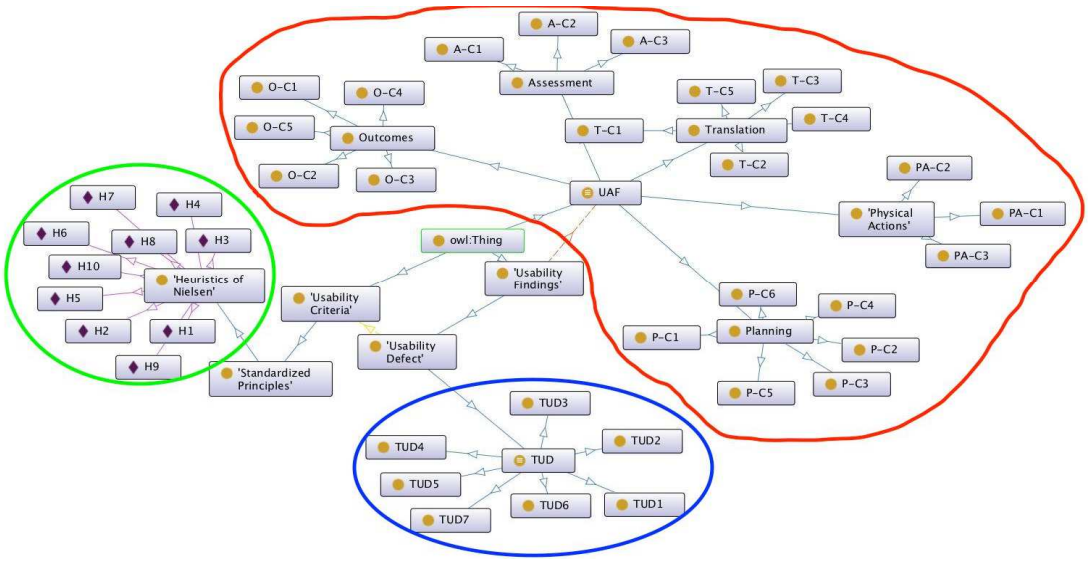


Figure 1 The UAF-HE distribution of main classes

The UAF-HE has 35 classes and 10 instances. The three classes *UAF*, *TUD* (Typical Usability Defect) and *Usability Criteria* are the conceptual backbone of our classification. The Table 1 shows the labels adopted for the UAF-HE, with each respective meaning. Such labels are also presented in the rest of this paper to facilitate the visualization of the results and analyses.

[a] UAF-HE ontology (OWL):
https://drive.google.com/file/d/1XvTCBTCs2NxkUUbXzTF_yrpTv2jQeHSo/view?usp=sharing

Table 1 UAF-HE labels and its respective meanings

UAF: User Action Framework
Planning Contents
P-C1: *User model and high-level understanding of system.*
P-C2: *Goal decomposition.*
P-C3: *Task/step structuring and sequencing, workflow.*
P-C4: *User work context, environment.*
P-C5: *User knowledge of system state, modalities, and especially active modes.*
P-C6: *Supporting learning at the planning level through use and exploration.*
Translation Contents
T-C1: *Existence of a cognitive affordance to show how to do something.*
T-C2: *Presentation (of a cognitive affordance).*
T-C3: *Content, meaning (of a cognitive affordance).*
T-C4: *Task structure, interaction control, preferences and efficiency.*
T-C5: *Support of user learning about what actions to make on which UI objects and how through regular and exploratory use.*
Physical Action Contents
PA-C1: *Existence of necessary physical affordances in user interface.*
PA-C2: *Sensing UI objects for and during manipulation.*
PA-C3: *Manipulating UI objects, making physical actions.*
Outcome Contents
O-C1: *Existence of needed functionality or feature (functional affordance).*
O-C2: *Existence of needed or unwanted automation.*
O-C3: *Computational error.*
O-C4: *Results unexpected.*
O-C5: *Quality of functionality.*
Assessment Contents
A-C1: *Existence of feedback or indication of state or mode.*
A-C2: *Presentation (of feedback).*
A-C3: *Content, meaning (of feedback).*
TUD: Typical Usability Defects
TUD1: *Additional unnecessary steps not required as part of completing a task.*
TUD2: *Misleading information.*
TUD3: *Insufficient and/or poor information on the user interface.*

TUD4: *Unexpected system responses.*
TUD5: *Limitations in navigation.*
TUD6: *Inefficient use error recovery mechanisms.*
TUD7: *Physical characteristics of the user interface that are not suitable for the physical characteristics of the user.*

Heuristics of Nielsen
H1: *Visibility of system status.*
H2: *Match between system and the real world.*
H3: *User control and freedom.*
H4: *Consistency and standards.*
H5: *Error prevention.*
H6: *Recognition rather than recall.*
H7: *Flexibility and efficiency of use.*
H8: *Aesthetic and minimalist design.*
H9: *Help users recognize, diagnose, and recover from errors.*
H10: *Help and documentation.*

The *UAF* class is the union of five disjoint sub-classes, as indicated by red circles at Figure 1, each one to represent the first level categories of the UAF interaction cycle: *Planning*, *Translation*, *Physical Actions*, *Outcomes* and *Assessment*. Each of these five sub-classes has its respective disjoint sub-classes (22 in total) to represent its respective contents, according to the standard UAF framework [Hartson and Pyle 2012, p. 677-685].

The *Usability Findings* class is the super-class of *Usability Defect* class, which is the super-class of *TUD* according to the terminology presented by the ISO/IEC 25066 (2016). The *TUD* class is the union of seven disjoint sub-classes, each one representing a TUD, as indicated by the blue circle at Figure 1.

Finally, the *Usability Criteria* class is the super-class of *Standardized Principles*, according to the ISO/IEC 25066 (2016). Therefore, we included a sub-class of *Standardized Principles*, named *Heuristics of Nielsen*, to model the 10 usability heuristics of Nielsen, each one represented by an instance at the model (see the green circle at Figure 1). The heuristics of Nielsen are not disjoint, and one usability finding may be classified among more than one heuristic [Nielsen 1994]. To enforce a unique classification and, therefore, enhance the classificatory power of the UAF-HE, we suggested the fits first strategy. The fits first strategy is based on the fact that the heuristics of Nielsen were sorted by their explanatory power (the probability of explaining different usability findings). Therefore, the fits first strategy implies that a

practitioner must classify a usability finding with the first possible heuristic among the ten instances, respecting its order.

The next section presents the methods and material of this research.

4. Methods and Material

This section describes the methods adopted for our research. We aimed to answer the following question:

> ***Research Question**: How to classify usability findings to enhance the matching process for validation of new heuristics?*

The goal of this research was to create a usability finding classification to enhance the matching process for future researches about HE. In addition, we elaborated three hypotheses (H0, H1 and H2), and each respective test is also described at this test (see Section 4.3).

Our methods aimed to demonstrate the feasibility of the UAF-HE. In this direction, we conducted a case study classifying 173 usability findings from the online project of Firefox Focus at GitHub. The remaining of this section describes the Firefox Focus browser, the protocol of our case study and the procedures of data analysis.

4.1. The Firefox Focus Browser

The Firefox Focus[b] is a private mobile browser created by Mozilla, launched in the market of in 2017. Firefox Focus browser is available for both Android and iOS platforms. One of the main features of Firefox Focus is to block hidden trackers, as analytic and content trackers. In addition, Firefox Focus also has an access button to erase recent browsing history. Figure 2 shows a screenshot of the settings interface of Firefox Focus.

[b] Firefox Focus website: https://www.mozilla.org/en-US/firefox/mobile/

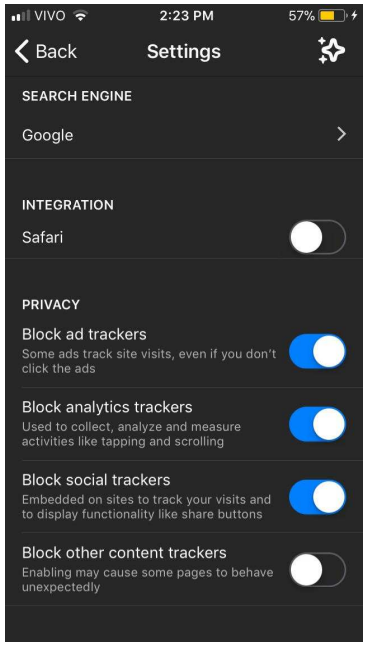


Figure 2 Settings screen of Firefox Focus mobile browser for iOS

We choose to adopt the Focus browser in this study because of the availability of project issues (which includes usability issues) in online datasets. The Firefox Focus development team used GitHub to discuss the Android[c] and the iOS[d] project issues online. Their team created a "UX" label to refer to project issues classified by them as related to user experience. Although their team referred to UX issues, the issues referred to usability aspects of the interface. We reinforce that it was out of the scope of this study to discuss the differences and similarities between user experience and usability.

Firefox Focus also represents a popular interest, because it is part of the Firefox family, one of the most popular web browsers. The popularity of Firefox brings comments from all over the world to the Firefox Focus project. In this context, contributors' descriptions of usability findings are dynamic and cannot be known a priori, a characteristic of the open world [Bendale and Boult 2015]. For this reason, conducting a case study with Firefox Focus was adequate to test the employment of our classification.

The following section describes the protocol we adopted during the case study.

4.2. Case Study

To evaluate the UAF-HE classification, we employed it in a set of usability issues retrieved from the Firefox Focus' projects on GitHub. We considered both the Android and the iOS projects. We only collected issues' information; individual information

[c] https://github.com/mozilla-mobile/focus-android/issues

[d] https://github.com/mozilla-mobile/focus-ios

about the contributors, which posted the respective issues in the platform, was not collected.

We conducted the analysis of issues between November 6th, 2017, and November 10th, 2017. Because the Firefox Focus project did not receive only usability issues, we filtered the project's issues by the label "UX". The UX label was the unique label created by the Focus' development team to indicate potential usability related issues. In sequence, we filtered the issues by those already closed ("closed" tag). This second filter was necessary to ensure that we would analyze issues that were potentially solvable, because they were already closed. In consequence, we obtained 173 potential usability issues, 126 from the Android project and 47 from the iOS project.

To classify each of the usability issues, we prepared an online form so that the authors could access collaboratively and classify the issues. The classification protocol for each issue was as follows:

STEP 1: To inform the specific project (Android or iOS).

STEP 2: To inform the issue's id as presented in the GitHub project.

STEP 3: To read and understand the description of the issue.

STEP 4: To classify the issue among the 22 UAF classifications.

STEP 5: To classify the issue among the seven (7) TUD classifications.

STEP 6: To classify the issue in the first possible heuristic among the ten (10) heuristics of Nielsen. Notice that the heuristics of Nielsen are not disjoint [Nielsen 1994], and that we did not classified the issues with all possible heuristics. We chose to classify by the first possible heuristic because the heuristics of Nielsen are ordered by its coverage (probability of covering a usability finding).

After the classification protocol, we saved the form's responses in a spreadsheet for data analysis. The next session explains the procedures of data analysis.

4.3. Data Analysis

This section explains the procedures of data analysis conducted among the results of this study. The primary goal of our analysis was to test each hypothesis of this study.

The hypothesis H0 (Null) was "*Employing the UAF-HE cannot enhance the matching process in comparison to employing only the UAF"*. We tested this hypothesis by testing the other two hypotheses. Accepting the hypothesis H1 or H2 rejects the H0.

The hypothesis H1 was "*Adding the TUD classification to the UAF can enhance the matching process in comparison to employing only the UAF"*. To test this hypothesis, we compared the number of issue sets formed after employing the UAF classification (n_{set1}) against the number of issue sets formed after the UAF and the TUD classifications together (n_{set2}). Therefore, to accept the hypothesis H1 it is necessary to have an increase in the number of issue sets formed after with the UAF and the TUD classifications together in comparison to the number of sets formed after the UAF

classification alone. In other words, to accept the hypothesis H1 the following formula must be verified:

$$n_{set2} > n_{set1} \quad (1)$$

Finally, the hypothesis H2 was "*Employing the UAF-HE can enhance the matching process in comparison to employing only the UAF*". To test this hypothesis, we compared the number of issue sets formed after employing the UAF and the TUD classifications (n_{set2}) against the number of issue sets formed after the UAF, the TUD and the Heuristic classifications together (n_{set3}). Therefore, to accept the hypothesis H1 it is necessary to have an increase in the number of issue sets formed after with the UAF and the TUD classifications together in comparison to the number of sets formed after the UAF classification alone. In other words, to accept the hypothesis H2 the following formula must be verified:

$$n_{set3} > n_{set2} \quad (2)$$

To test this hypothesis, we compared the size and number of sets of issues formed after employing both the UAF and the TUD on the issues against the respective values after employing the classification with the UAF, the TUD and Nielsen's heuristics together.

The following section presents the results of our study, its analysis and discussions.

5. Results and Discussion

This section presents the results of the case study conducted. The goal of this study was to create a usability finding classification to support comparisons (during the matching process) of HE approaches by increasing the UAF potential to indicate dissimilarities among usability finding descriptions. In this direction, we elaborated three hypothesis, which were tested and the reslts presented at this section.

We collected 173 project issues from Firefox Focus project on GitHub. 127 of these issues were retrieved from the Android project at Firefox Focus directory in GitHub; while 46 issues were from the iOS project. Such a difference may be due to similar tasks between the projects, some changes requested in the Android project may have been also applied to the iOS project. Figure 3 shows an initial analysis of the 173 issues according to the heuristics of Nielsen, employing the fits first strategy. In this context, only one (1) of the issues could not be referred by a heuristic. The title of such issue is "*[meta][ux] iterate on the onboarding experience*"; such description was too broad to be classified among the heuristics and also among the TUD variables.

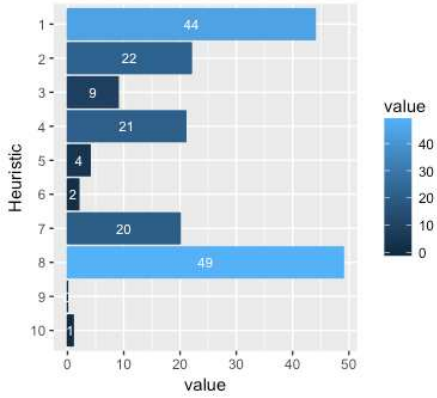
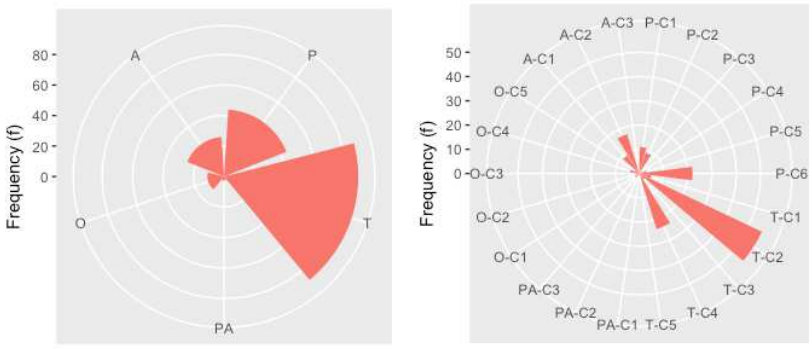


Figure 3 Issues classified by each of the Nielsen's heuristics, employing the fits first strategy

Figure 3 indicated a colored map according to the value of issues fitted in each heuristic. Although we have employed the fits first strategy, which could impplicate in a preferrence for the initial heuristics, the heuristic 8 "*Aesthetic and minimalist design*" had the highest value of issues. In addition, we can divide the 10 heuristics in three groups (according to the color scale). Such groups may indicate that the explanatory power of each of Nielsen's heuristics is different in the domain of usable privacy for mobile browsers. Future studies can explore in deeper this new hypothesis. The three groups of heuristics are as follows:

- Group 1: *Visibility of system status* (#1) and *Aesthetic and minimalist design* (#8). These were the most referred heuristics in our study.
- Group 2: *Match between system and the real world* (#2), *Consistency and standards* (#4) and *Flexibility and efficiency of use* (#7). This group had less coverage than group 1.
- Group 3: *User control and freedom* (#3), *Error prevention* (#5), *Recognition rather than recall* (#6), *Help users recognize, diagnose, and recover from errors* (#9) and *Help and documentation* (#10). This group had less coverage than group 1 and 2.

Figure 4 plots the frequency (f) of issues classified with the first (Figure 4a) and second (Figure 4b) level of the standard UAF according to the interaction cycle. Considering Figure 4a, the *Planning (P)* category classifies 44 issues, the *Translation (T)* category classifies 88 issues, the *Physical Actions (PA)* category classifies three (3) issues, the *Outcomes (O)* category classifies eleven (11) issues and the *Assessment (A)* category classifies 26 issues. We verified a significant dependency ($X^2 = 132.83$; $df = 4$; $p\text{-value} < 0.05$) among the frequency of issues by the categories. Considering our sample of filtered issues, this fact represents a predominance of translation issues among usability findings in the context of Firefox Focus. This may also indicate a predominance of translation issues among usability findings in the context of private browsers. Comparing Figure 4a with Figure 4b, one can see that the main issues classified in the planning category were also classified in the P-C6 category; and the main issues classified in the translation category were also classified in the T-C2 content.



(a) **(b)**

Figure 4 Frequency (f) of issues classified with the standard UAF by each stage of the interaction cycle

The Table 2 shows the distribution of frequency (f) of issues classified by each of the content levels, the second level of the standard UAF (*mean* ≈ *7.86; median = 2.5; sd* ≈ *12.88*). In total, such classification generated 17 sets (n_{set1} = *17*). We highlight the diversity of the issues classified, because the value of *sd* was higher then both the mean and the median. For this reason, we speculate that P-C6, T-C2 and T-C4 may be predominant types of usability issues that Firefox Focus' project will face in the future and, maybe, they are predominant types of usable privacy issues among private browsers.

Table 2 Frequency (f) of issues classified with the content levels of the UAF.

Content	*f*
P-C1	11
P-C2	9
P-C3	0
P-C4	0
P-C5	3
P-C6	22
T-C1	5
T-C2	56
T-C3	3
T-C4	24
T-C5	0
PA-C1	1
PA-C2	0
PA-C3	2
O-C1	2
O-C2	1
O-C3	2
O-C4	4
O-C5	2
A-C1	9
A-C2	17
A-C3	0

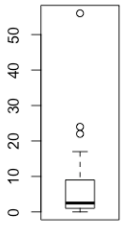


Figure 5 Boxplot analysis of the frequency of issues in each set of n_{set2}

The folowing sections were structured according to the test of each hypothesis. It presents the results from the tests and discussions about it. Because the test of hypothesis H0 is dependent on the test of the other two hypothesis, we present it as the last one.

5.1. Test of the Hypothesis H1

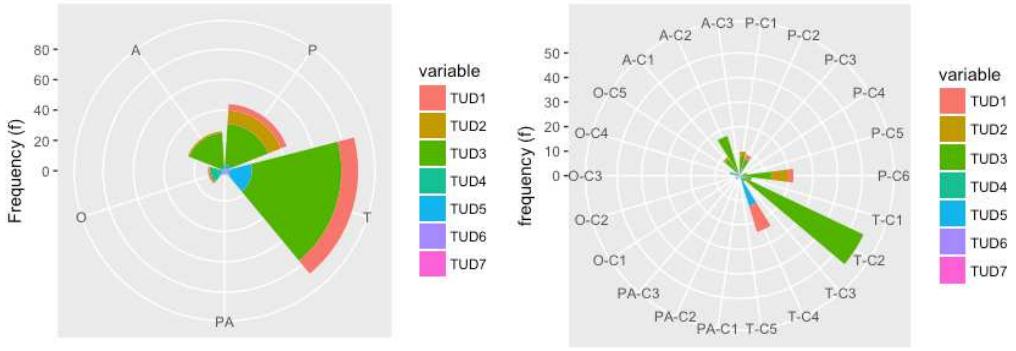
The hypothesis H1 was "*Adding the TUD classification to the UAF can enhance the matching process in comparison to employing only the UAF"*. To test this hypothesis, we compared the number of issue sets formed after employing the UAF classification (n_{set1}) against the number of issue sets formed after the UAF and the TUD classifications together (n_{set2}). The value of n_{set1} was 17; this section shows the calculus of n_{set2}.

The Table 3 shows the frequency of issues in n_{set2} according to each of the first level categories of the UAF (each category is represented by a row in the table) and its respective classification according to TUD (each one represented by a column in the table). As shown at Table 3, the category T (translation) had the highest frequencies among rows; the TUD3 had the highest frequencies among the columns; and the intersection of both T and TUD3 had the highest value of the Table 3. In summary, this classification formed 37 ($n_{set2} = 37$); an increase of 20 sets in comparison to n_{set1}.

Table 3. Number of issues classified by the first level categories of the standard UAF and the TUD classification.

Categories	TUD1	TUD2	TUD3	TUD4	TUD5	TUD6	TUD7
P	4	9	26	0	3	1	1
T	11	0	59	0	14	3	1
PA	0	0	0	0	1	0	2
O	1	0	1	4	3	2	0
A	0	1	23	2	0	0	0

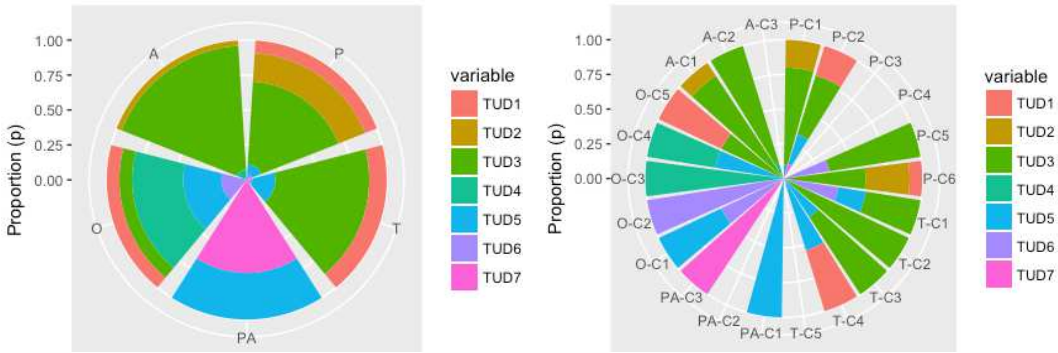
As an alternative to visualize the n_{set2} is shown at Figure 6, which can also be compared to Figure 4. Figure 6a shows the frequency (f) of issues resulted from employing the TUD classification in addition to the first level of the standard UAF classification, and the Figure 6a shows the frequency (f) of issues resulted from employing the TUD classification in addition to the second level of the standard UAF classification. The same analysis discussed for the Table 3 can be applied to discuss the results shown at Figure 4.



(a) **(b)**

Figure 6 Frequency (f) of issues resulted from employing the TUD classification in addition to the first level categories of the standard UAF classification

Similarly, Figure 7a shows the proportion (p) of issues resulted from employing the TUD classification in addition to the first level of the standard UAF classification, and Figure 7b shows the proportion (p) of issues resulted from employing the TUD classification in addition to the second level of the standard UAF classification. Such proportional analyses are important to normalize the values of issues among each classification. These figures made clear the impact of employing each TUD variable among the first level categories of the standard UAF classification. As shown at Figure 7a, the highest impact the TUD classification was located at the *Outcome (O)* category. Figure 7b indicated that the last four (4) TUD variables were more common between the contents T-C4 and O-C4; while the first three (3) variables of TUD were more common between the contents O-C5 and T-C3. These findings raised the following questions: (i) "*Are the TUD variables organized according to categories of the interaction cycle?*" and (ii) "*What is the cost/benefit of employing each of the TUD variables for each of the categories and contents of the standard UAF?*".



(a) **(b)**

Figure 7 Proportion (p) of issues resulted from employing the TUD classification in addition to the first level categories of the standard UAF classification

Finally, based on the analysis of this section, we accepted the hypothesis H1 because adding the classifications based on TUD to the standard UAF increased its potential to indicate dissimilarity among usability findings ($n_{set2} > n_{set1}$). The next section presents the test of the hypothesis H2.

5.2. Test of the Hypothesis H2

The hypothesis H2 was "*Employing the UAF-HE can enhance the matching process in comparison to employing only the UAF*". To test this hypothesis, we compared the number of issue sets formed after employing the UAF and the TUD classifications (n_{set2}) against the number of issue sets formed after the UAF, the TUD and the Heuristic classifications together (n_{set3}). The value of n_{set2} was 17; this section shows the calculus of n_{set3}.

Figure 8 shows the sets of issues formed after the classification with both the UAF and TUD classifications (n_{set2}), indicating each of the 37 sets. The vertical axis (y) represents the 22 variables (P-C1 to A-C3) of the UAF, while the horizontal axis (x) represents the seven (7) variables of TUD. The issue set 14 (T-C2, TUD3), points out as the largest set, as showed at Figure 6b.

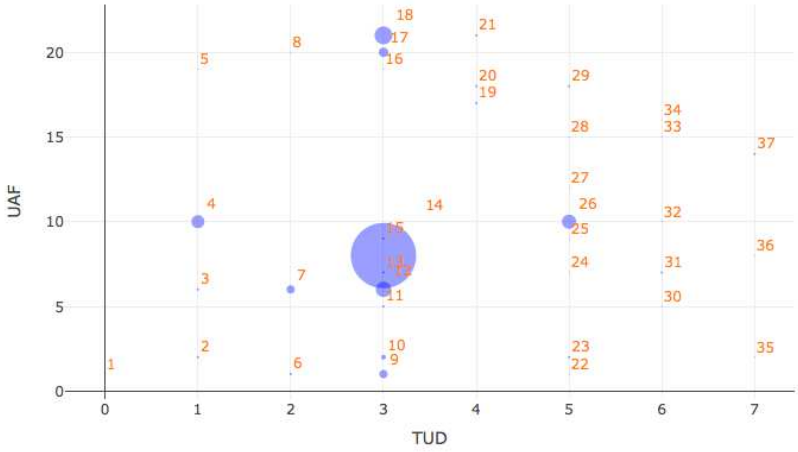


Figure 8 Sets of issues formed after both UAF and TUD classifications

Similarly, Figure 9 shows the sets of issues formed after the UAF, the TUD and the Heuristic classifications (n_{set3}). In total, Figure 9 shows 65 sets of issues ($n_{set3} = 65$), each of these sets was indicated at Figure 9. The z-axis represents the 22 variables (P-C1 to A-C3) of the UAF, the x-axis represents the seven (7) variables of TUD and the y-axis represents the ten heuristics of Nielsen. The largest set among the sets shown at Figure 9 is set 59 (T-C2, TUD3, Heuristic 8) with 32 issues. As shown at Figure 10, most of the sets showed at Figure 9 have less than five (5) issues classified, and only set 59 has more than 15 issues classified. Because set 59 (see Figure 9) had excessive issues in comparison to the other sets shown at Figure 9, we analyzed its issues to suggest additional classifications that could enhance the degree of dissimilarities among such issues. In this regard, we found that indicating the animation, color pallets and

iconography may help to better discriminate such issues and reduce the high number of issues in set 59.

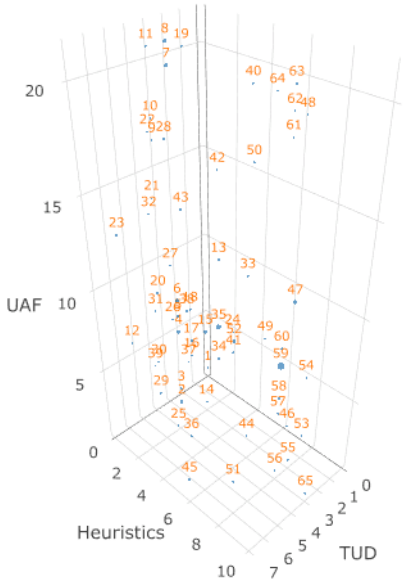


Figure 9 Groups of issues formed after the UAF, the TUD and the Heuristic classifications

The analyses of this section showed that the employment of the heuristics of Nielsen as an additional classification to both UAF and TUD classification could increase the capacity of indicating dissimilarities among usability findings ($n_{set3} > n_{set2}$). For this reason, we accepted the hypothesis H2.

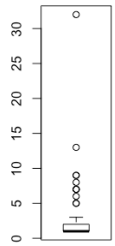


Figure 10 Boxplot analysis of the frequency of issues in each set of n_{set3}

The next section presents the test of the hypothesis H0.

5.1. Test of the Hypothesis H0 (Null)

The hypothesis H0 (Null) was "*Employing the UAF-HE cannot enhance the matching process in comparison to employing only the UAF*". We tested this hypothesis by testing hypothesis H1 and H2. Because we accepted both the hypothesis H1 and H2, we therefore rejected the hypothesis H0.

6. Implications for Design

Based on the observations of this study, we found that the heuristics of Nielsen can be grouped among three (3) separated groups according to its coverage on Firefox Focus' usability issues. The three groups are as follows:

- Group 1: *Visibility of system status* (#1) and *Aesthetic and minimalist design* (#8). These were the most referred heuristics in our study.
- Group 2: *Match between system and the real world* (#2), *Consistency and standards* (#4) and *Flexibility and efficiency of use* (#7). This group had less coverage than group 1.
- Group 3: *User control and freedom* (#3), *Error prevention* (#5), *Recognition rather than recall* (#6), *Help users recognize, diagnose, and recover from errors* (#9) and *Help and documentation* (#10). This group had less coverage than group 1 and 2.

We suggest to practitioners that, during HE on private browsers, they should employ the heuristics of Nielsen on such a group sequence, because the original order of the heuristics was defined according to their coverage on usability issues from other domains but privacy.

7. Conclusions

This study aimed to create a usability finding classification to enhance the matching process for researches about HE. Although different classifications were proposed in the literature, there is no widely adopted classification to describe usability findings [Hornbæk 2010; Yusop *et al.* 2017]. In this regard, the following research question guided our work:

> ***Research Question****: How to classify usability findings to enhance the matching process for validation of new heuristics?*

To reach our goal and answer this question, we created the UAF-HE, an extension for the standard UAF. Therefore, we elaborated three (3) hypotheses. We tested these hypotheses based on a case study classifying 173 usability findings from an online project of Firefox Focus. Only one of the 173 usability findings could not be classified among the extensions provided by the UAF-HE, but it was classified with the standard UAF.

The most part of usability findings classified in our case study was related to the translation contents of the UAF (T-C2), to the "*Insufficient and/or poor information on the user interface*" type of usability defect (TUD3) and to the heuristic "*Aesthetic and minimalist design*" (H8). These findings indicated that these classes of usability findings have been the main challenge for the design of Firefox Focus. Future studies can explore such fact with other private browsers.

The data analysis of our study supported the acceptation of H1 ("*Adding the TUD classification to the UAF can enhance the matching process in comparison to employing only the UAF*") and H2 ("*Employing the UAF-HE can enhance the matching process in comparison to employing only the UAF*"), and the rejection of H0

("*Employing the UAF-HE cannot enhance the matching process in comparison to employing only the UAF*"). In addition, we discussed potential new extensions to the UAF-HE by indicating the animation aspect, the color pallet and the iconography characteristics. Such indications may help to better discriminate the differences among usability findings classified among T-C2, TUD3 and H8.

In conclusion, the UAF-HE can enhance the process of matching usability finding descriptions for the domain of HE and, potentially, fill out the gap in the literature. The UEO [Elkin *et al.* 2013] may also be appropriate to enhance the process of matching usability finding descriptions, during comparisons among HE variations, but we can speculate it only for the domain of health systems. Because none of the other popular classifications (the CUP, the RCA, the ODC and the UEO) was focused on the HE domain, we suggest the UAF-HE as the appropriate classification to support the matching process in future comparisons of HE proposals [Vilbergsdottir *et al.* 2014; Yusop 2017; Elkin *et al.* 2013].

The main limitation of our study regards to the knowledge of the researchers that employed the UAF-HE to classify the usability findings. It is possible that different researchers could employ different classifications with the UAF-HE for the same usability findings. Therefore, we suggest some practices to enhance the internal validity of future studies employing the UAF-HE. We suggest that at least two different researchers are needed to employ the UAF-HE with the same usability findings; while a third researcher is needed in cases of disagreements between the first two to achieve a majority for the classification. In addition, because we focused on the HE domain, we suggest that the same researchers must employ the UAF-HE for both the controlled HE and the HE under test. Finally, the external validity of the UAF-HE is initially supported by its heritage from the UAF and its compendium of relevant case studies [Vilbergsdottir *et al.* 2014]. Nevertheless, future studies can validate the UAF-HE with different datasets.

7. Acknowledgements

This study was supported by the grants 2015/24525-0, 2017/15239-0 and 2018/26038-8, São Paulo Research Foundation (FAPESP).

This study was financed in part by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil* (CAPES) - Finance Code 001.

This study was supported by The Brazilian National Council for Scientific and Technological Development (CNPq - MCTIC).

References

Araujo, R. M. (2017). Information Systems and the Open World Challenges. In: Boscarioli, C.; Araujo, R. M.; Maciel, R. S. P.[Eds.]. . *GranDSI - BR – Grand Research Challenges in Information Systems in Brazil 2016 - 2026*. Special Committee on Inform ation Systems (CE - SI). Brazilian Computer Society (SBC). p. 42–51.

Bendale, A. and Boult, T. (2015). Towards open world recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Cranor, L. F. and Buchler, N. (nov 2014). Better Together: Usability and Security Go Hand in Hand. *IEEE Security Privacy*, v. 12, n. 6, p. 89–93.

De, L. A. and Zezschwitz, E. Von (2016). Usable privacy and security. *it - Information Technology*, v. 58, n. 5, p. 215–216.

Elkin, P. L., Beuscart-Zephir, M.-C., Pelayo, S., Patel, V. and Nøhr, C. (2013). The usability-error ontology. *Beuscart-ZéphirM. JaspersM. KuziemskyC. NøhrC. AartsJ.(Eds.), Context sensitive health informatics: Human and sociotechnical approaches*, p. 91–96.

Garfinkel, S. and Lipford, H. R. (2014). *Usable Security: History, Themes, and Challenges*. Morgan \& Claypool Publishers. v. 5

Hartson, H. R., Andre, T. S. and Williges, R. C. (2001). Criteria for evaluating usability evaluation methods. *International journal of human-computer interaction*, v. 13, n. 4, p. 373–410.

Hartson, R. and Pyla, P. S. (2012). *The UX Book: Process and guidelines for ensuring a quality user experience*. Elsevier.

Hermawati, S. and Lawson, G. (2016). Establishing usability heuristics for heuristics evaluation in a specific domain: Is there a consensus? *Applied Ergonomics*, v. 56, p. 34–51.

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*, v. 33, n. 2, p. 144–162.

Hornbæk, K. (2010). Dogmas in the assessment of usability evaluation methods. *Behaviour & Information Technology*, v. 29, n. 1, p. 97–111.

Hornbæk, K. and Frøkjær, E. (dec 2008). Comparison of techniques for matching of usability problem descriptions. *Interacting with Computers*, v. 20, n. 6, p. 505–514.

ISO: International Organization for Standardization (2016). ISO/IEC 25066:2016(en), Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Common Industry Format (CIF) for Usability — Evaluation Report. . https://www.iso.org/obp/ui/#iso:std:iso-iec:25066:ed-1:v1:en, [accessed on Oct 6].

Kawakani, C. T., Barbon, S., Miani, R. S., Cukier, M. and Zarpelão, B. B. (12 mar 2017). Discovering Attackers Past Behavior to Generate Online Hyper-Alerts. *iSys - Revista Brasileira de Sistemas de Informação*, v. 10, n. 1, p. 122–147.

Lavery, D., Cockton, G. and Atkinson, M. P. (1997). Comparison of evaluation methods using structured usability problem reports. *Behaviour & Information Technology*, v. 16, n. 4–5, p. 246–266.

Lewis, J. R. (2014). Usability: Lessons Learned … and Yet to Be Learned. *International Journal of Human–Computer Interaction*, v. 30, n. 9, p. 663–684.

Nielsen, J. (1994). Enhancing the Explanatory Power of Usability Heuristics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. , CHI '94. ACM. http://doi.acm.org/10.1145/191666.191729.

Norman, D. (2013). *The design of everyday things: Revised and expanded edition*. Basic Books (AZ).

Preece, J., Sharp, H. and Rogers, Y. (2015). *Interaction design: beyond human-computer interaction*. 4. ed. John Wiley & Sons.

Still, J. D. (apr 2016). Cybersecurity Needs You! *interactions*, v. 23, n. 3, p. 54–58.

Vilbergsdottir, S. G., Hvannberg, E. T. and Law, E. L.-C. (2014). Assessing the reliability, validity and acceptance of a classification scheme of usability problems (CUP). *Journal of Systems and Software*, v. 87, p. 18–37.

Yusop, N. S. M., Grundy, J. and Vasa, R. (sep 2017). Reporting Usability Defects: A Systematic Literature Review. *IEEE Transactions on Software Engineering*, v. 43, n. 9, p. 848–867.