UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Adaptations to the Heuristic Evaluation (HE) method for novice evaluators

André de Lima Salgado

Dissertação de Mestrado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP SÃO CARLOS

André de Lima Salgado

Adaptations to the Heuristic Evaluation (HE) method for novice evaluators

Master dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, in partial fulfillment of the requirements for the degree of the Master Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Profa. Dra. Renata Pontin de Mattos Fortes

USP – São Carlos
August 2017

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

de Lima Salgado, André
d164a Adaptations to the Heuristic Evaluation (HE) method for novice evaluators / André de Lima Salgado; orientadora Renata Pontin de Mattos Fortes. -- São Carlos, 2017.
100 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2017.

1. Usability. 2. Heuristic Evaluation. 3. Novice Evaluators. 4. Evaluator-Effect. 5. Expertise-Effect. I. Pontin de Mattos Fortes, Renata , orient. II. Título.

André de Lima Salgado

Adaptações ao método de Avaliação Heurística (AH) para avaliadores novatos

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientadora: Profa. Dra. Renata Pontin de Mattos Fortes

USP – São Carlos
Agosto de 2017

For all

ACKNOWLEDGEMENTS

I would like to thank God, my family, my fiancée and everyone who helped me along this Master. In addition, I thank all professors, colleagues and my supervisor Renata P. M. Fortes for their kindly help, support and collaboration.

This study was supported by the grant 2015/09493-5, São Paulo Research Foundation (FAPESP).

“The combination of good observational skills and good design principles is a powerful tool, one that everyone can use, even people who are not professional designers. Why? Because we are all designers in the sense that all of us deliberately design our lives... ”

Norman (2013, p. xi-xii)

RESUMO

SALGADO, A. L. **Adaptações ao método de Avaliação Heurística (AH) para avaliadores novatos**. 2017. 100 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2017.

A Avaliação Heurística (AH) é um método popular de inspeção de usabilidade. Entretanto, seus resultados são dependentes da experiência dos avaliadores. Este estudo explorou e descreveu a diferença na qualidade de resultados (relatórios) de AH colaborativa conduzida por grupos de avaliadores de composição distinta, considerando diferentes quantidades de avaliadores experientes em cada grupo. Vinte e sete (27) avaliadores contribuíram voluntariamente com este estudo, nove (9) experientes e 18 novatos. Assim, foram organizados sete (7) grupos de AH, de acordo com quatro (4) níveis diferentes do fator "*presença de avaliador experiente*", variando de nenhum experiente até três (3) avaliadores experientes no mesmo grupo. Cada grupo de avaliadores concordou em entregar seus relatórios de AH para este estudo. A partir de tais relatórios, foi conduzida uma análise comparativa baseada em métodos específicos da área, e também baseado em uma análise de agrupamento com base em medidas de similaridade. Como resultado, descreveu-se as medidas F (*F-measure*) referentes ao relatório de cada grupo respeitando critérios estritos e relaxados de comparação. Além disto, foram descritos os dendrogramas resultados das análises de agrupamento. Os resultados mostraram que a qualidade de relatórios de AH colaborativas conduzidas por avaliadores experientes e novatos juntos pode ser mais similar à qualidade de relatórios de AH tradicional conduzida por múltiplos avaliadores experientes (*Grupo Benchmark*) do que à qualidade de relatórios de AH colaborativa conduzida por grupos formados apenas por avaliadores novatos (*Grupo Baseline*). Finalmente, discutiu-se resultados adicionais e implicações para pesquisas futuras na área.

Palavras-chave: Usabilidade, Avaliação Heurística, Avaliadores Novatos, Efeito-Avaliador, Efeito-Experiência.

ABSTRACT

SALGADO, A. L. **Adaptations to the Heuristic Evaluation (HE) method for novice evaluators**. 2017. 100 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2017.

Heuristic Evaluation (HE) is a popular method of usability inspection. However, its outcomes are dependent on the expertise of evaluators. This study explored and described the difference in quality of outcomes (reports) of a collaborative HE conducted by evaluator groups of distinct composition, regarding different numbers of expert evaluators in each group. Twenty-seven (27) evaluators voluntarily contributed with this study, nine (9) expert and 18 novice evaluators. Thus, I organized seven (7) HE groups according to four (4) different levels of the factor "*presence of an expert*", which ranged from no expert up to three (3) experts in the same group. Each group agreed to provide their reports for this study. Thereafter, I conducted a comparative analysis on the reports based on standard methods of the field and on a cluster analysis of similarities. I described the *F-measure* for each group report according to a relaxed and a strict criteria. Also, I described the dendrograms formed from the cluster analysis and the respective similarities indicated by each cluster. The results showed that the quality of reports from collaborative HE conducted by experts and novices together can be more similar to the quality of reports from a traditional HE with multiple expert inspectors (*Benchmark Group*) then to the quality of reports from a collaborative HE conducted by a group composed only by novice evaluators (*Baseline Group*). Finally, I discuss additional findings and implications for future studies in the field.

Keywords: Usability, Heuristic Evaluation, Novice Evaluators, Evaluator-Effect, Expertise-Effect.

LIST OF FIGURES

LIST OF TABLES

CONTENTS

CHAPTER

1

INTRODUCTION

1.1 Preamble

This dissertation aims to be an original study, with an exploratory goal and its procedures are quasi-experimental. Moreover, its results are based on the coherence theory of justification (WAZLAWICK, 2014; SHADISH; COOK; CAMPBELL, 2002).

This study is placed in the intersection of the following research areas: *Web and Interactive Multimedia Systems* and *Human-Computer Interaction*. The main approaches, methods and techniques used here come from Human-Computer Interaction and were applied in the context of developing usable Web and Interactive Multimedia Systems.

The following section presents an introduction to this study.

1.2 Heuristic Evaluation and Novice Evaluators

Usability is an important aspect in software design. It relates to quality and ergonomics of software (ISO/TR 9241-100, 2010; ISO/IEC 25066, 2016; FERNANDEZ; INSFRAN; ABRAHÃO, 2011). Adoption of usable systems may reduce monetary loses and increase productivity, achievement of goals and profits (BARUA; MANI; MUKHERJEE, 2012; NIELSEN, 2012). To develop usable software, Usability Evaluation Methods (UEM) play an essential role (HORNBÆK, 2010).

The ISO/IEC 25066 (2016) shows that UEMs can be categorized as *user-based evaluation* or as *inspection-based evaluation*. *User-based* methods require the participation of potential users. On the other hand, *inspection-based* methods require the participation of inspectors (e.g. usability professionals), also called evaluators. Designers can choose among different UEMs according to their needs. In initial stages of the development, *inspection-based* methods can be more appropriated if the current interface lacks interactivity. When interfaces still lack

interactivity, users can face difficulties to use it, while evaluators can still evaluate it because of their expertise in the field. On the other hand, when the interface has sufficient interactivity, *user-based* methods can be highly valuable providing evidences on usability issues that users will really care about (PETRIE; POWER, 2012; PREECE; SHARP; ROGERS, 2015).

This study addresses one of the most popular usability *inspection-based* methods, the Heuristic Evaluation (HE) (GEORGSSON; WEIR; STAGGERS, 2014; MARTINS *et al.*, 2014; JOHANNESSEN; HORNBÆK, 2014; FØLSTAD; LAW; HORNBÆK, 2012). On early studies about HE, Nielsen (1992) shows that HE has its best outcomes from the participation of evaluators with great expertise (called specialists). However, for distinct reasons (e.g. costs), counting on evaluators with low expertise (novices) is common among practitioners (FERNANDEZ; INSFRAN; ABRAHÃO, 2011; RENZI *et al.*, 2015; HUANG, 2012; BRUUN; STAGE, 2014; BRUUN; STAGE, 2015). Similarly to other inspection methods, the HE is dependent on evaluators' knowledge and opinion (COCKTON; LAVERY; WOOLRYCH, 2009; COCKTON; WOOLRYCH, 2001). Thus, individual differences, including different expertise, can influence on outcomes from an inspection method as HE. Effects caused on HE outcomes due to evaluators' individual differences is called evaluator-effect, while effects caused on HE outcomes due to evaluators different expertise is called expertise-effect (BRAJNIK; YESILADA; HARPER, 2011; HERTZUM; JACOBSEN, 2001). Therefore, it is crucial to develop adaptations for the HE method that make the method less biased by evaluator or expertise effect.

MacFarlane and Pasiali (2005), MacFarlane, Sim and Horton (2005), Salian, Sim and Read (2013), Salian and Sim (2014) and Read (2015) investigated adaptations of HE for children evaluators. According to them, children present specific characteristics (proper of their age) that are important to consider during HEs. These authors proposed simplifications of heuristics description, as well as adaptations to traditional severity rating scale in order to enhance HE conducted by children. However, their contributions are focused on children, and the extent which it is applicable to other groups of novice evaluators still needs investigation.

Likewise, Wodike, Sim and Horton (2014) explored adaptations of HE for teenager evaluators. The authors explored the outcomes of a HE conducted by novice teenagers tutored by one teenager with previous training. According to the authors, the performance of the teenagers during the cases studied was not satisfactory. The authors argued that ways that make evaluation more playful and friendly to teenagers were still necessary. Studies that investigate the applicability of such adaptations to other samples of novice evaluators are still needed.

Besides, Buykx (2009) and Petrie and Buykx (2010) proposed a collaborative alternative for HE, called Collaborative Heuristic Evaluation (CHE). The primary goal of developing the CHE was not to adapt HE for novice evaluators. However, during discussion about their results, the authors suggested that future studies could investigate the applicability of CHE as a training for novice evaluators when conducting the method in collaboration with expert evaluators. In this case, novices could possibly learn from the expert as CHE occurs. This study presented great

insights on how to adapt HE in order to reduce its bias caused by the evaluator and expertise effects. However, further studies still need to validate such a possibility.

Therefore, the following research problem remains: *Due to the importance of HE, and the presence of evaluator and expertise effects, it is still necessary studies that explore adaptations for HE regarding the broad profile of novice evaluators*.

1.3 Objective

Based on the related studies, this study aimed to determine if an CHE conducted by expert and novice evaluators together (*Mixed Group*) has qualified outcomes in comparison to standard HEs. To understand such goal, I adopted the following terms regarding compositions of evaluator groups in HEs: *Baseline Group*, *Mixed Group* and *Benchmark Group*. A *Baseline Group* is composed only by novice evaluators; a *Mixed Group* is composed by novice evaluators and, at least, one expert evaluator; and a *Benchmark Group* is composed only by expert evaluators. For this reason, the objective of this study can be understood as to answer the following *research question*:

> *"Can a CHE performed by a Mixed Group result in outcomes whose quality can be considered more similar to the quality of outcomes from a traditional HE with multiple expert inspectors (Benchmark Group) than to the quality of outcomes from a CHE conducted only by novice inspectors (Baseline Group)?"*

The following section presents the hypothesis developed to answer this question.

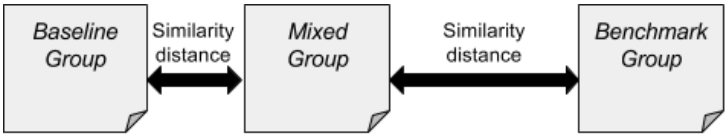
1.3.1 Hypothesis

To answer such research question, I elaborated three hypotheses. The first hypothesis, the **Hypothesis H0 (null)**, assumed that the quality of the outcomes from a *Mixed Group* was more similar to the quality of the outcomes of a *Baseline Group* (composed only by novice evaluators) than to the quality of outcomes from a *Benchmark Group* (composed only by expert evaluators), as represented in Figure 1. I considered this hypothesis as null because its contributions to the literature would be less significant than the acceptance of the following hypothesis.

The second hypothesis, the **Hypothesis H1**, assumed that the quality of the outcomes from a *Mixed Group* was equally similar to the quality of the outcomes of a *Baseline Group* and to the quality of outcomes from a *Benchmark Group*, as represented in Figure 2.

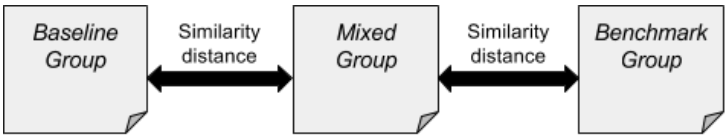
Finally, the third hypothesis, the **Hypothesis H2**, assumed that the quality of the outcomes from a *Mixed Group* was more similar to the quality of the outcomes of a *Benchmark*

Figure 1 – Representing the Hypothesis H0 (null).



Source: Elaborated by the author.

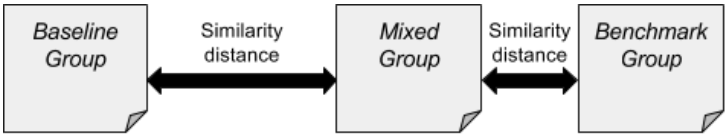
Figure 2 – Representing the Hypothesis H1.



Source: Elaborated by the author.

Group than to the quality of outcomes from a *Baseline Group*, as represented in Figure 3.

Figure 3 – Representing the Hypothesis H2.



Source: Elaborated by the author.

This study tested these hypothesis regarding multiple analysis, showed at Chapter 5. The next section presents additional as appended hypotheses emerged during this study.

Additional Objectives

In direction to the main objective, I included additional objectives as appended hypotheses emerged. Such objectives were achieved in collaboration with colleagues and are published in the literature or in process of publication. The additional objectives (AOs) were as follows:

AO1: To identify difficulties faced by novice inspectors when performing HEs (SALGADO; FORTES, 2016).

AO2: To structure alternative adaptations of HE for novice inspectors (SALGADO; FORTES, 2016; SALGADO *et al.*, 2016a).

AO3: To verify the need of Brazilian enterprises to count on HE appropriated for novice inspectors (SALGADO *et al.*, 2016b).

AO4: To verify the importance of HE for new contexts and audiences (SALGADO; RODRIGUES; FORTES, 2016).

AO5: To review the importance of usability evaluation methods, focusing on the organization of HE for novice practitioners, for Rich Internet Application (FORTES; ANTONELLI; SALGADO, 2016a; FORTES; ANTONELLI; SALGADO, 2016b).

AO6: To validate new heuristics for the hot topic *elderly interaction with mobile technologies* (SALGADO *et al.*, 2017b).

AO7: To discuss the applicability of HE for evaluation of usability for people with wide range of characteristics in games (FORTES *et al.*, 2017).

AO8: To report the application of HE in a case study of User Centered Design (UCD) (SALGADO *et al.*, 2017a).

The following section discuss the value of this research.

1.4 Value of the Research

This research may help, especially, organizations that cannot count on multiple expert evaluators to perform HEs. It may also help novice evaluators themselves by providing means for them to perform their work with better quality. As a consequence, it may help communities that depend on such organizations and professionals.

Nonetheless, this study has the potential to help the popularization of usability practices among organizations that cannot count on multiple expert evaluators and, for this reason, did not apply proper methods for the development of usable technologies.

1.5 Organization of this Dissertation and Final Remarks

The remaining of this text is structured as follows:

- Chapter 2 presents a literature review on User Centered Design (UCD) process and traditional UEMs.
- Chapter 3 presents a review on HE for novice evaluators.
- Chapter 4 presents a methodological review on assessment of distinct UEMs.
- Chapter 5 presents the methods and materials adopted in this research.
- Chapter 6 presents evaluation of the results from the conducted experiments.

- Chapter 7 presents discussions about results obtained from the experiments.
- Chapter 8 presents conclusions and suggestions of future works in the field of this dissertation.

This chapter presented an introduction for this Master's dissertation. Therefore, it showed a brief of related works, motivation for the present study, objectives of this dissertation and the structure of this text. The following chapter presents a review on traditional UEMs from the HCI field.

CHAPTER

2

LITERATURE REVIEW

2.1 Introduction

This chapter presents a review on Usability definition, the User Centered Design (UCD) process and traditional Usability Evaluation Methods (UEM). Finally, this chapter describes the following UEMs: *Testing with Users*, *Cognitive Walkthrough*, *Guidelines Review*, *Heuristic Evaluation* and *Collaborative Heuristic Evaluation*.

2.2 Usability

Usability was initially understood as synonym of easy to use, an aspect that could improve programmers' productivity (DEMERS, 1981). Later, Nielsen (2012) defined usability based on five (5) quality components, as follows:

Learnability: "*How easy is it for users to accomplish basic tasks the first time they encounter the design?*"

Efficiency: "*Once users have learned the design, how quickly can they perform tasks?*"

Memorability: "*When users return to the design after a period of not using it, how easily can they reestablish proficiency?*"

Errors: "*How many errors do users make, how severe are these errors, and how easily can they recover from the errors?*"

Satisfaction: "*How pleasant is it to use the design?*"

In sequence, it became recognized as one of the important factors that can impact the quality and ergonomics of a software (ISO/IEC 25066, 2016; ISO 9241-210, 2010). Thus,

different series of ISO also defined usability. The ISO/IEC 25066 (2016) defined usability as a subset of quality in use in accordance with ISO 9241-210 (2010), as follows:

> "*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 TR 25066 (2016) described the following terms that compose its usability definition:

User: "*person who interacts with a system, product or service*".

Goal: "*intended outcome*".

Effectiveness: "*accuracy and completeness with which users achieve specified goals*".

Efficiency: "*resources expended in relation to the accuracy and completeness with which users achieve goals*".

Satisfaction: "*freedom from discomfort, and positive attitudes towards the use of the product*".

Context of use: "*users, tasks, equipment (hardware, software and materials), and the physical and social environments in which a product is used*".

Task: "*activities required to achieve a goal*".

Other usability definitions were proposed in the literature. However, I adopted the definition showed by ISO/IEC 25066 (2016) for this study because of its international reputation and because it was relatively recent when this text was written.

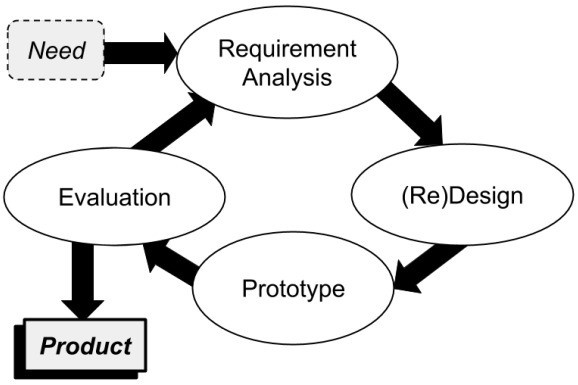
2.3 User Centered Design

The ISO 9241-210 (2010) shows User-Centered Design (UCD) as a particular comprehension of Human-Centered Design (HCD). The term "*User-Centered Design*" means that user is a central source of information for all UCD stages. According to the standard, UCD refers to design of interfaces for stakeholders that can be considered as users (see the definition for *user* in the previous section of this chapter). UCD is basis for developing usable interfaces (PREECE; SHARP; ROGERS, 2015; NORMAN, 2013; KATZEFF *et al.*, 2012).

UCD process enhances the development of usable interfaces because it proposes the consideration of human factors/ergonomics and usability techniques through its cycles (ISO 9241-210, 2010). Since the study of Gould and Lewis (1985), UCD is based on the following principles (PREECE; SHARP; ROGERS, 2015; KATZEFF *et al.*, 2012):

- *Early focus on users and Tasks*: understanding who the users are, their characteristics (e.g. their cognitive and behavior) and the nature of their work (e.g. how it is accomplished by tasks).
- *Empirical Measurement*: examining real users using the product (or a prototype), observing and analyzing their performance and reactions.
- *Iterative Design*: the UCD must be a cycle. When problems are discovered from the evaluation activities, they may be corrected in a sequent cycle (see Figure 4).

Figure 4 – The UCD cycling process, adapted from the interaction design model showed by *usability.gov* portal.



Source: Adapted from the U.S. Department of Human Health & Human Services (*usability.gov* portal) at: <https://www.usability.gov/what-and-why/user-centered-design.html>.

Figure 4 shows the UCD cycle. This figure was retrieved and adapted from *usability.gov* portal. It shows the common stages of a UCD process: *Requirement Analysis*, *(Re)Design*, *Prototype* and *Evaluation*. Each stage of the UCD cycle receives outcomes from the previous stage, and provides outcomes for the following one.

The UCD process starts by establishing new requirements from users' needs (see the balloon "*Need*" at Figure 4). These requirements can be generated by previous evaluation stage (previous cycle), or by initial requirement analysis using other tools (e.g.: questionnaires). Thereafter, the *(Re)Design* stage begins. At such a stage, the practitioners can design different alternatives that might prosper considering the requirements specified. The sequent stage is the *Prototype* stage. At the *Prototype* stage, practitioners can prototype each alternative design proposed. Finally, the UCD process has a stage for evaluating the current version of the prototype (or the interface itself). The goal of the *Evaluation* stage is to identify limitations among prototypes evaluated. Commonly, outcomes from the *Evaluation* stage are related to usability

characteristics of the product, and are originated after the conduction of a Usability Evaluation Method (UEM) (PREECE; SHARP; ROGERS, 2015). Popular UEMs are presented in the next section of this chapter.

2.4 Usability Evaluation Methods

The process of usability evaluation should not occur only once during the development process. In contrast, it should be periodically done over UCD cycles. Different cycles of a UCD process may require distinct alternatives of evaluation. As users are a central source of information in UCD process, their participation during evaluation stages is valuable for the design. However, some cycles cannot embrace the participation of potential users (e.g.: because the prototype does not have sufficient interactivity to be used by such users) (DIX *et al.*, 2003; HORNBÆK; STAGE, 2006; PREECE; SHARP; ROGERS, 2015; FERNANDEZ; INSFRAN; ABRAHÃO, 2011).

Due to the wide variety of UEMs proposed in the literature, categorizing groups of UEMs is an important task that can help practitioners to understand which method fits better for each situation. In this context, the ISO/IEC 25066 (2016) classifies UEMs between *user-based evaluation* and *inspection-based evaluation*. According to the ISO/IEC 25066 (2016), *user-based evaluations* involve a representative sample of end users to perform pre-defined tasks using the interface. Nevertheless, the representativeness of such user samples, especially the minimum number of users needed to perform such evaluations, is subject of discussion in the literature (BORSCI *et al.*, 2013). The ISO/IEC 25066 (2016) defines *user-based evaluations* as:

> "*evaluation that involves representative users performing tasks with the system to enable identification of usability problems and/or measurements of efficiency, effectiveness, user satisfaction or other user experiences*"

Most of terms used in the definition of *user-based evaluations* are part of the definition of usability, showed previously. The other terms are presented by the ISO/IEC 25066 (2016) as follows:

Usability problem: "*situation during use resulting in poor effectiveness, efficiency or satisfaction*".

User Experience (UX): "*a person's perceptions and responses that result from the use and/or anticipated use of a product, system or service*".

The term user experience is still discussed in the literature, as highlighted by Bevan *et al.* (2016). However, I presented such definition in order to be in accordance to the definition of *user-based evaluations* as presented by the ISO/IEC 25066 (2016).

2.4.1 User-based Evaluations

Preece, Sharp and Rogers (2015) show that testing usability of an interface with users can involve different methods, such as interviews, questionnaires and observation. Among such methods, observing users can reveal usability problems that have a more severe impact on users (PETRIE; POWER, 2012). Observations can be performed in laboratory or following ethnographic procedures (KRUMM, 2016; PREECE; SHARP; ROGERS, 2015). Regarding laboratory tests, using the Think Aloud protocol, eye-tracking and video recording users' interaction can enhance the quality of outcomes (ERICSSON; SIMON, 1980; DIX *et al.*, 2003; PREECE; SHARP; ROGERS, 2015).

The Think-Aloud protocol aims to help users to verbalize their interaction with an interface, providing insights of usability problems. This protocol requires the presence of a test moderator, who is responsible for helping and motivating users to verbalize thoughts related to their interaction during test sessions. As users may feel uncomfortable to reveal their thoughts, they must be informed about the test purpose, their roles and rights during the testing sessions, including their right of leaving test sessions whenever they want. In addition, practitioners must ask for users consent about participating in the tests (ERICSSON; SIMON, 1980; PREECE; SHARP; ROGERS, 2015).

Video recording tests allow practitioners to post analyze usability issues in deeper, analyzing users' behavior carefully. Additionally, eye-tracking technologies allow practitioners to verify where users were looking at during each stage of the interaction (PREECE; SHARP; ROGERS, 2015). Examples of video recording and eye-tracker tools are:

- **Video-recording**
  - *Morae* - a popular user testing video recording tool. Morae is a commercial product (payed solution) and belongs to TechSmith Corporation[1].
  - *UserZoom* Remote Usability Testing - a software solution for remote usability test on Websites. This software is also a commercial product (payed solution) and belongs to UserZoom[2].
- **Eye-tracking**
  - *Eye Tribe* Tracker - the Eye Tribe Tracker is a series of eye-tracker equipments developed by The EyeTribe enterprise[3].
  - *Tobii* - the Tobii eye-tracker series is developed by Tobii AB enterprise. This group of eye-trackers has traditional computer based eye-trackers and also a smart glasses

[1] Retrieved June 13th, 2016, from Morae Website at: <www.techsmith.com/morae.html>
[2] Retrieved June 13th, 2016, from UserZoom Website at: <www.userzoom.co.uk/software/remote-usability-testing/#content-read=true>
[3] Retrieved June 13th, 2016, from The EyeTribe Website at: <theeyetribe.com/>

solution, enabling practitioners to eye-track multiple interfaces through an environment[4].

The following section presents methods of *inspection-based evaluation*.

2.4.2 Inspection-based Evaluations

Inspection-based evaluations are based on the judgment of evaluators (typically usability specialists) respecting determined criteria (ISO/IEC 25066, 2016). Popular inspection criteria are guidelines, standards, principles, good practices (ISO/IEC 25066, 2016). The ISO/IEC 25066 (2016) presents a list of main examples of usability inspection methods, including *Cognitive Walkthrough*, *Guidelines Review* and *Heuristic Evaluation*. The following sections define each one of such inspection methods.

Cognitive Walkthrough

Usability walkthrough methods have a particular characteristic: it can involve usability specialists, end users or other professionals. In this context, a walkthrough method involves the evaluator, or a group of them, playing the roles of users interacting with a specific interface in order to identify usability problems related to achieving the goal of pre-defined tasks (ISO/IEC 25066, 2016).

Cognitive Walkthrough (CW) is a popular method among usability walkthrough methods. Lewis *et al.* (1990) proposed the first version of CW. Thereafter, Polson *et al.* (1992) and other authors proposed sequent review for the CW method (MAHATODY; SAGAR; KOLSKI, 2010). A CW has two successive phases: preparation and evaluation.

The first phase of a CW is called preparation phase. At this phase, practitioners must choose the tasks that will be base for the CW. In sequence, they must divide (describing) each chosen task into specific actions that compose the respective task. Thus, evaluators will follow such actions to proceed the CW. The preparation phase is also the phase for organizers to explain for evaluators about user profile and context of use (PREECE; SHARP; ROGERS, 2015; MAHATODY; SAGAR; KOLSKI, 2010).

The second phase of a CW is the evaluation phase. In such period, an evaluator must play the role of a user conducting the defined set of pre-defined actions, as determined since the preparation phase (JADHAV; BHUTKAR; MEHTA, 2013). Evaluator(s) must conduct each prescript action answering four questions that are related to users' behavior, as follows (MAHATODY; SAGAR; KOLSKI, 2010):

Question 1 - "*Will the user try to achieve the right effect?*" – This question refers to what users may be thinking when the action begins.

[4] Retrieved June 13th, 2016, from Tobii AB Website at: <www.tobii.com/group/>

Question 2 - "*Will the user notice that the correct action is available?*" – This question refers to whether users would be able to locale the command.

Question 3 - "*Will the user associate the correct action with the effect that user is trying to achieve?*" – This question refers to whether users would be able identify the specific command.

Question 4 - "*If the correct action is performed, will the user see that progress is being made toward solution of the task?*" – This question refers to users ability to understand the possible given feedback.

The CW questions are broad enough to be applied among different interfaces. The appropriate employment of such questions is dependent on the evaluators knowledge. Also, the CW may be conducted by evaluators of distinct characteristics. However, the CW evaluation has a limited approach since it evaluates only the ease of learning, that is a part of usability (WHARTON *et al.*, 1994). According to Mahatody, Sagar and Kolski (2010), many variants of CW method were proposed in the literature. However, it was out of the scope of the present chapter to describe all of them. Instead, we focused on describing the traditional and widely applied CW method.

Guidelines Review

In Guidelines Review, evaluators inspect the interface following a set of specific usability criteria, called guidelines. The Web Content Accessibility Guidelines (WCAG 2.0)[5] is a popular guidelines set that can be used for evaluation of usability for users with the widest range of characteristics and capabilities. The difference between usability and accessibility is still a theme for investigation in the area, and such a discussion is out of the scope of this study. However, some definitions show that usability and accessibility may share common aspects, as content related to a widest range of characteristics and capabilities of users (ISO/IEC 25066, 2016).

WCAG 2.0 is a document of many guidelines organized among four principles: *Perceivable*, *Operable*, *Understandable* and *Robust*. The *Perceivable* principle cautions that any information and user interface component must be implemented in order to be perceived by users in some way. The *Operable* principle cautions that any feature of the Web system must be available through keyboard access. The *Understandable* principle refers to making the Web content readable and understandable by users. Finally, the *Robust* principle refers to Web content being interpretable by user agents, including Assistive Technology. For each of these principles, the WCAG 2.0 shows a subset of guidelines considered as essential for implement the respective principle in the Web. Each WCAG 2.0 guideline has its respective success criteria with levels

[5] Retrieved from Web Content Accessibility Guidelines (WCAG) 2.0 Website at: <www.w3.org/TR/WCAG20/#guidelines>

ranging among A, AA, and AAA (further information about WCAG 2.0 success criteria levels can be found at the WCAG portal[6]).

Reviewing large sets of guidelines may be time consuming for human evaluators. For this reason, the literature presents Automatized Tools for Guidelines Checking that implement most of the work of a human evaluator during a guideline review method. In this context, the **Web Accessibility Checker (achecker)**[7] and the **TAW tool**[8] are examples of automatized tools for checking part of WCAG 2.0 guidelines.

The WCAG 2.0 is a popular set of guidelines. However, other usability guidelines exist, as page design, typography, charts, diagrams, graphics and icons (WATZMAN; RE, 2009).

Heuristic Evaluation

Similarly to Guidelines Review, Heuristic Evaluation (HE) involves evaluators inspecting an interface based on a set of usability principles, called heuristics. In contrast to other guidelines, that usually represent specific usability issues, heuristics are general (high level) usability rules (NIELSEN, 1995; PREECE; SHARP; ROGERS, 2015). Usability heuristics are usually generated from a large set of usability issues. As an example, Nielsen (1994) enhanced the explanatory power of his traditional usability heuristics after a factor analysis of 249 usability problems.

HE is one of the most popular *inspection-based evaluations* (FØLSTAD; LAW; HORNBÆK, 2012; MARTINS *et al.*, 2014; PETRIE; POWER, 2012; PAZ; POW-SANG, 2016). Nielsen and Molich (1990) proposed the HE method for the first time, with an initial set of nine (9) usability heuristics based on the professional experience of the authors regarding usability issues. This method consists on multiple evaluators independently inspecting an interface in order to discover disagreements between its design and a set of usability heuristics (NIELSEN; MOLICH, 1990; NIELSEN, J., 1994; PAZ; POW-SANG, 2016; NIELSEN, 1992).

After a sequence of studies on usability heuristics, Nielsen and his colleagues composed a set of usability heuristics that remains popular in the field (NIELSEN; MOLICH, 1990; NIELSEN, 1994; NIELSEN, J., 1994). Since these studies, Nielsen's 10 heuristics became as a standard for conducting HEs. However, NIELSEN, J. (1994) reinforced that a HE does not require the use of his 10 heuristics as a standard and that a more appropriate set of heuristics can be used instead, according to recommendations from an expert. A complete list of Nielsen's 10 heuristics can be found at Nielsen Norman Group portal[9]. The following list presents the title for each of Nielsen's 10 traditional heuristics:

[6] WCAG 2.0 success criteria Webpage: <www.w3.org/TR/UNDERSTANDING-WCAG20/conformance.html#uc-levels-head>
[7] <achecker.ca/checker/index.php>
[8] <www.tawdis.net/ingles.html?lang=en>
[9] <https://www.nngroup.com/articles/ten-usability-heuristics/>

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.
10. Help and documentation.

In order to produce its best outcomes, HE should be conducted by group of evaluators (NIELSEN, J., 1994). In this context, Nielsen (1992) shows that a HE should be conducted by at least three (3) evaluators. However, the ideal number of evaluators in a HE is still under wide debate in the literature (BORSCI *et al.*, 2013).

Preece, Sharp and Rogers (2015) explain the HE method dividing it among three sections: *briefing session*, *evaluation period* and *debriefing session*. During the *briefing session*, the evaluators receive all the guiding information about how the HE should be conducted. In the *evaluation period*, the evaluators actually conduct the usability evaluation working independently through the interface looking for possible violations of any of the heuristics considered. Finally, at the *debriefing session*, the evaluators discuss their findings with each other and prepare a final list of usability problems with suggestions of possible improvements for the interface. In addition, at this session, evaluators are required to rate severity for each usability problem reported. NIELSEN, J. (1994) suggests the following severity scale:

0 - Not a usability problem: it is not a usability problem at all.

1 - Cosmetic problem: it is only a cosmetic problem. Its correction may be made only if extra time is available in the project timeline.

2 - Minor problem: the correction of this kind of problem may receive low priority.

3 - Major problem: the correction of this kind of problem may receive a high priority.

4 - Usability catastrophe: these must be the first problems to be corrected. They must be corrected before the product be released.

Figure 5 – Representing the difference between HE and CHE at the evaluation period.

(a) Standard Heuristic Evaluation (HE). (b) Collaborative Heuristic Evaluation (CHE).

Source: Elaborated by the author.

Despite being an affordable method, the quality of HE outcomes is dependent on evaluators' skills. In consequence, evaluators with low expertise in usability area (also called novices) can report outcomes with low quality (NIELSEN, 1992). On the other hand, counting on evaluators with great expertise in usability field still represents an elevated cost for some organizations (COCKTON; LAVERY; WOOLRYCH, 2009; LOWRY; ROBERTS; ROMANO, 2013; BORYS; LASKOWSKI, 2014; JOHANNESSEN; HORNBÆK, 2014; SCHELLER; KÜHN, 2015; ALJOHANI; BLUSTEIN, 2015; PAZ; PAZ; POW-SANG, 2015). In consequence, some organizations rarely count on more multiple expert evaluators and, instead, they frequently count on novice evaluators (SLAVKOVIC; CROSS, 1999; KOUTSABASIS *et al.*, 2007; BUYKX, 2009; BRUUN; STAGE, 2014; BRUUN; STAGE, 2015). Such organizations still need a qualified solution to supply the lack of expert participation in their HEs. The effect caused by the participation of novice evaluators in HEs is called the expertise-effect, and is part of the evaluator-effect (HERTZUM; JACOBSEN, 2001; BRAJNIK; YESILADA; HARPER, 2011).

Collaborative Heuristic Evaluation

The Collaborative Heuristic Evaluation (CHE) was proposed by Buykx (2009) and Petrie and Buykx (2010). The CHE method instructs that evaluators could perform the *evaluation period* of a HE collaboratively, and not individually as in the traditional HE. In the CHE case, scribes help evaluators to report their findings, which is similar to the role of observers as described by Nielsen (1995). Such difference is illustrated in Figure 5, representing evaluators performing a traditional HE (Figure 5a) and a CHE (Figure 5b). At Figure 5, I represented evaluators (Evl.), the Web interface, list of problems, discussion or debriefing process (Disc.), the CHE scribe (Scr.) who is responsible for taking notes and preparing CHE report, and the report. As showed in the figures, one can see that evaluators could inspect the same sequence of screens during the CHE.

Studies about CHE are still a few in the literature. This lack of studies limits the understanding about pros and cons of a CHE. For this reason, it is still difficult to describe in deep the benefits of the collaborative evaluation process. On the other hand, this method opens a venue for evaluators to learn with each other as the evaluation period goes on. In summary, the CHE is a potential method and studies about its contributions are valuable to the literature.

The next chapter, Chapter 3, describes studies related to the effect of novice participation on HEs and CHEs.

2.5 Final Remarks

In this chapter I presented a literature review of the main concepts involved in this study. For this reason, I referred to definitions of *usability*, described concepts of *User Centered Design* and distinct *Usability Evaluation Methods* that can occur during such design process. In addition, I presented some of the most popular methods of usability evaluation: *Testing with Users*, *Cognitive Walkthrough*, *Guidelines Review* and *Heuristic Evaluation*. Next chapter presents studies related to evaluator and expertise effect in the context of HEs.

CHAPTER

3

HEURISTIC EVALUATION AND NOVICE EVALUATORS

3.1 Introduction

This chapter presents content about the effect of novice evaluator participation in HE sessions. Moreover, I show results and discussions from studies about how to classify usability evaluators and distinct studies about adapting HE for novice evaluators.

3.2 Classifying Experts and Novices in HEs

A level classification for usability evaluator expertise is still rare in the field. Nielsen (2002) suggests that three characteristics are common among expert usability evaluators:

(i) Knowledge on interaction theory and on UEMs, especially regarding *test with users*.

(ii) "*High brain power*".

(iii) "*10 years' experience running user tests and other usability activities, such as field studies*".

Although Nielsen (2002) suggests characteristics of usability experts, he did not provide a classification of novice evaluators. It is intuitive to understand that novice evaluators would be those professionals that do not have such characteristics. However, how far from such characteristics are novice evaluators? Such a question remains to be answered. In this context, MacDonald and Atwood (2013) argue that understanding the skills needed to be an expert evaluator in usability area remains as a gap in the literature.

Indeed, a wide accepted classification for usability evaluator expertise levels is still required. Nevertheless, the study of Botella, Alarcon and Peñalver (2014) helps to fill this gap. Botella, Alarcon and Peñalver (2014) proposed a classification of five (5) levels of usability expertise, as follows:

Novice: "*person without a university degree but with at least one training course on HCI and few hours of practice in usability evaluation.*"

Beginner: "*professionals without university degree but with several training courses in HCI or with less than 2,500 hours of practice in usability evaluation.*"

Intermediate: "*professionals with a bachelor's degree or lees than 5,000 hours of practice in usability evaluation.*"

Senior: "*professionals with master's degree or less than 7,500 hours of practice in usability evaluation.*"

Expert: "*professionals with master's degree, and optionally a doctorate, and with more than 10,000 hours of professional practice deliberate (10 years) in the field of usability evaluation.*"

The classification levels showed by Botella, Alarcon and Peñalver (2014) are in accordance with Nielsen (2002) in different aspects. Both Nielsen (2002) and Botella, Alarcon and Peñalver (2014) show time of experience with UEMs as a requirement to differentiate levels of expertise among usability evaluators, and indicate that 10 years would be an ideal time for a usability professional to become an expert. However, only Botella, Alarcon and Peñalver (2014) indicate that expert evaluators should hold at least a master's degree. Moreover, Nielsen (2002) and Botella, Alarcon and Peñalver (2014) consider knowledge on HCI theories as another aspect that differentiate evaluator expertise levels. On the other hand, only Nielsen (2002) presents "*brain power*" as a characteristic of expert evaluators.

Although Botella, Alarcon and Peñalver (2014) present contributions to the field about aspects that could be considered in order to understand the expertise of an usability evaluator (or inspector), further studies are still needed in order to validate their proposal and investigate the impact of each expertise of evaluator on HE outcomes.

For the purposes of this study, we considered as novices those professionals that had, at least, an introductory course on Human-Computer Interaction, could understand the terms that compose usability and the heuristics of NIELSEN, J. (1994), and were conducting a HE for the first time after a training session. I adopted such criteria in order to identify professionals that have the minimum requirements to perform a HE as usability professionals. On the other hand, I considered as expert those evaluators with at least four (4) years of experience in usability area, one publication authored in a vehicle of HCI area, previous usability evaluations and were engaged on researches related to usability. I adopted this criteria based on the Brazilian scenario

and on common characteristics, the most experienced evaluators that agreed to participate as voluntary in this study.

3.3 Adapting Heuristic Evaluation for Novice Evaluators

Nielsen (1992) shows that evaluators of different expertise can find different usability problems as outcomes of a HE. In addition, he showed that evaluators with higher expertise in usability area are capable of finding more usability problems than evaluators with less expertise. Slavkovic and Cross (1999) are among the first to argue that the literature should investigate variations for the HE in order to support novice evaluators. In such study, the authors analyzed data from HEs conducted by 43 novices. According to them, novice evaluators tend to focus on specific parts of a complex interface when evaluating it, instead of focusing on the entire application. Slavkovic and Cross (1999) also show that novice evaluators may ignore entire aspects of the interface during HEs.

After a literature review, I identified some studies that adapted HE for some profile of novice evaluator. Among such studies, the major part studied HE for children evaluators. MacFarlane and Pasiali (2005) argue that, although children cannot be considered usability experts, they have specific characteristics that are not present in adults and are important for usability evaluation on interfaces designed for children.

In the context of HE for children evaluators, MacFarlane and Pasiali (2005) proposed simplifying Nielsen's heuristic descriptions, removing jargon and simplifying language, in order to adapt HE for children. During their experiment, almost all children showed great interest in participating in a HE. Nevertheless, the authors argue that further rephrasing of heuristics are still needed. Similarly to MacFarlane and Pasiali (2005), Salian, Sim and Read (2013) observed 14 children conducting HE of a mobile game with the heuristics of Korhonen and Koivisto (2006). According to the results of Salian, Sim and Read (2013), children have difficulties to understand severity ratings, linking heuristics with a problem found and identifying similar problems as the same. In a sequent study, Salian and Sim (2014) propose an adaptation of HE for older children as evaluators. The adaptations proposed by the authors were simplifying heuristic descriptions and introducing a new severity scale (the "*Bad Scale*") based on three different smile figures, representing levels of satisfaction: *Bad*, *Very Bad* and *Awful*. In their study, children were able to find usability problems, but they could not identify the validity of simplified heuristics and neither of the "*Bad Scale*".

Besides the group of studies on HE for children evaluators, the study of Wodike, Sim and Horton (2014) explored HEs adapted for teenager evaluators. Wodike, Sim and Horton (2014) proposed the presence of a facilitator, with greater knowledge in HE, guiding a group of teenager evaluators during a HE on a game interface. In their study, evaluators could play the game during 15 minutes after each 30 minutes of evaluation. The results from MacFarlane

and Pasiali (2005) were not satisfactory in order to verify the adaptations proposed as effective. According to MacFarlane and Pasiali (2005), facilitators had difficulties to inform their peers about the HE process, and most part of evaluators appeared to prefer discussing their results in the game instead of discussing HE findings. Nevertheless, the authors argue that heuristic description, severity scale and reporting forms should be more suitable for teenagers.

Despite the contributions from studies about children and teenager evaluators in HEs, the extent of its results to adult novice evaluators still need to be explored. For this reason, the author of this study and colleagues conducted two exploratory studies on how to adapt HE for adult novice evaluators. In a first study, Salgado *et al.* (2016a) explored tactics that expert evaluators apply during HEs. We collected 38 tactics after a survey with four (4) experienced evaluators. These tactics were divided by each of Nielsen's heuristics, composing a roadmap for novice evaluators during their first HEs. We understand that novice evaluators could apply these tactics during their first HEs, until they get enough expertise to conduct HE by themselves. In a sequent study, the author of this dissertation and his supervisor conducted a survey with 13 professors from the usability field and 15 novice evaluators in order to explore novice difficulties on distinguishing Nielsen's heuristic descriptions (SALGADO; FORTES, 2016). Our results showed that a novice evaluator may have more difficulties in distinguishing between heuristic 3 and 7. For this reason, we proposed adaptations for the description of such heuristics (SALGADO; FORTES, 2016, p. 395), as follows:

> **Heuristic 3 - Control to undo and redo actions:** "*Users often choose system functions by mistake - e.g. after actions of trial and error - and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.*"
>
> **Heuristic 7 - Accelerators, shortcuts and efficiency of use:** "*Accelerators (e.g. shortcuts) - unseen by the novice user - may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.*"

Finally, another proposal with potential as HE adaptation for novice evaluators is the Collaborative Heuristic Evaluation (CHE) (BUYKX, 2009; PETRIE; BUYKX, 2010). The primary goal of Buykx (2009) and Petrie and Buykx (2010) was not to adapt HE for novice evaluators. However, Buykx (2009) suggested that CHE had a potential to be used as training for novice evaluators through the composition of group mixed by expert and novice evaluators. Furthermore, the following section presents a literature review on studies about CHE and novice evaluators.

3.3.1 Collaborative Heuristic Evaluation and Novice Evaluators

The interaction among evaluators during the *evaluation period* of a CHE suggest the employment of such method as a training for novice evaluators when they are part of a mixed group (interacting with expert evaluators during the *evaluation period*). A few works explored the participation of novice evaluators in CHE. Huang (2012) compared the differences of performing CHE remotely (called Remote CHE - rCHE) or not. The authors also compared the performance of novice and expert evaluators on discovering usability problem with rCHE, based on outcomes from test with eight (8) users. Such evaluations were conducted on four (4) websites. Huang (2012) compared the performance between novice and expert groups based on a relaxed (same underlying problem) and on a strict criteria (problems that referred to the same interface element, at the same level of abstraction). Considering a relaxed matching criteria, novice group covered 12.% of usability problems identified with user-based evaluations, while expert group covered 21.6%. Under a strict matching criteria, the novice group covered 10.1% of usability problems found during user-based evaluations, while expert group covered 19.9%. Also, according to Huang (2012), novices were not always precise in describing problems.

Othman *et al.* (2014) investigated the difference between the performance of a group of novice evaluators and a group of expert evaluators conducting a CHE of 3 mobile applications. In contrast to Huang (2012), Othman *et al.* (2014) did not compare the outcomes from CHEs with outcomes from test with users. Othman *et al.* (2014) organized five (5) groups of evaluators, but their composition (in terms of evaluator expertise) was not clear. In the case study of Othman *et al.* (2014), novice evaluators covered between 30% to 35% of the usability problems reported by expert evaluators. Although Huang (2012) and Othman *et al.* (2014) investigated aspects of the participation of novice evaluators in CHE, exploring the performance of mixed groups (groups composed by both expert and novice evaluators) remains as a gap in the field.

3.4 Final Remarks

This chapter reviewed the topic of HE adapted for novice evaluators. Such review shows that researches about this topic are still a few, and further studies are needed in order to better adapt HE for novice evaluators (especially adult evaluators). In this direction, validating such adaptations is a requirement to move forward in the field. For this reason, the next chapter reviews methods for validating new UEMs.

CHAPTER

4

METHODS FOR COMPARISON OF USABILITY EVALUATION METHODS

4.1 Introduction

This chapter presents a review on methodology referred in the literature about assessment of UEMs. Consequently, it shows procedures for comparing outcomes from distinct UEMs as reviewed by Hartson, Andre and Williges (2001). Such methods still present limitations, as summarized by Hornbæk (2010). Moreover, Fernandez, Abrahão and Insfran (2012) argued that UEMs should not be compared only in number, but in quality as well. Nevertheless, the methods indicated by Hartson, Andre and Williges (2001) provide frameworks for evidence that are still adopted in the field (FERNANDEZ; ABRAHÃO; INSFRAN, 2012).

4.2 Comparing UEMs

Hartson, Andre and Williges (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 problems*" with the set of usability problems found by the UEM being assessed. Hartson, Andre and Williges (2001) show that such comparison can be performed through distinct means, as:

Comparison against a standard list of usability problems. This method adopts the premise that a list of all usability problems of an interface exists and is known. Thus, the outcomes from the UEM assessed can be compared to such a list. Hartson, Andre and Williges (2001) show that traditional user-based evaluations conducted in laboratory are usually accepted as a gold standard.

Determining the realness of usability problems by expert review and judgment. For this method, usability experts review the list of usability problems originated from the assessed UEM in order to judge the realness of each problem.

Determining the realness of usability problems by end-users review and judgment. For this approach, a sample of end-users review and judge the realness of each usability problem reported from a UEM.

According to Hartson, Andre and Williges (2001), the literature commonly compares different UEMs based on the measures: *Reliability*, *Thoroughness*, and *Validity*. *Reliability* is the consistency of the UEM outcomes, independent of evaluator or expertise effect. *Thoroughness* is how close the outcomes of a UEM is to a standard set of usability problems. Finally, *Validity* is how correct are the outcomes of a UEM, also evaluating its realness.

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

***Hits*:** usability problems reported by the assessed UEM that exist in the standard set of usability problems.

***Misses*:** usability problems not reported by the UEM that exist in the standard set of usability problems.

***False alarms*:** usability problems reported by the UEM that do not exist in the standard set of usability problems.

Considering these metrics, the formulas for *Thoroughness* (HARTSON; ANDRE; WILLIGES, 2001, p. 390), *Validity* (HARTSON; ANDRE; WILLIGES, 2001, p. 392) and *Effectiveness* (HARTSON; ANDRE; WILLIGES, 2001, p. 394) are as follows:

$$Thoroughness = \frac{hits}{standard} \tag{4.1}$$

$$Validity = \frac{hits}{hits + false\ alarms} \tag{4.2}$$

$$Effectiveness = Thoroughness \times Validity \tag{4.3}$$

Hartson, Andre and Williges (2001, p.394) argue that *Validity* and *Thoroughness* have a preference over other measures. Thus, they described a weighed combination of *Validity* and

Thoroughness, called *F-measure*, adapting it from Manning, Schütze *et al.* (1999). The formula for *F-measure* is based on an α value, as follows:

$$F-measure = \frac{1}{\alpha(\frac{1}{Validity}) + (1-\alpha)(\frac{1}{Thoroughness})} \tag{4.4}$$

Considering Equation 4.4, for an α equals to 0.5, both *Validity* and *Thoroughness* receive the same weight. Thus, Equation 4.4 could be described as the following:

$$F-measure = \frac{2 \times Validity \times Thoroughness}{Validity + Thoroughness} \tag{4.5}$$

This Equation 4.5 was developed by (HARTSON; ANDRE; WILLIGES, 2001, p. 394).

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

4.2.1 Matching Usability Problems

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

1. **Similar changes:** problems 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, each usability finding must be based on respective usability problems, which could be considered as similar.

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

4. **User Action Framework (ANDRE *et al.*, 2001):** usability problems are structured according to the seven stages of actions, from Norman (2013), describing whether a problem relates to the *planning*, *translation*, *physical actions*, *outcome and system functionality*, or *assessment* stages. Also, problems can be described as independent from the interaction cycle. Such problem description allows practitioners to compare similarity of problems based on a comparison of categories.

Furthermore, Buykx (2009), Petrie and Buykx (2010), Babajo and Petrie (2012), Huang (2012) and Petrie and Power (2012) adopt similar matching criteria for assessment involving CHE outcomes. According to these authors, similarity can be analyzed through a relaxed or a strict criteria, as follows:

Relaxed matching criteria: usability problems are considered similar if they refer to the same problem, or to the same design element, independent of the level of abstraction. If the same underlying problem is described, two usability problems are considered as similar.

Strict matching criteria: usability problems are considered similar only if they refer to the same problem, 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*, as referred by 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*. Due to the affinity of our researches, I adopted the relaxed and strict criteria showed by Buykx (2009) and Petrie and Buykx (2010) in the present study.

4.3 Final Remarks

This chapter presented a review of methods for assessment of UEMs. Among such methods, I presented popular methods commonly adopted to compare UEMs, along with its respective measures and metrics showed in Hartson, Andre and Williges (2001). Additionally, I presented processes adopted in the literature for matching usability problems. Therefore, the next chapter is based on this review and presents the methods of this study.

CHAPTER

5

METHODS AND MATERIAL

5.1 Introduction

This study aimed to determine if CHE conducted by expert and novice evaluators has qualified outcomes in comparison to standard HEs. In this direction, this chapter presents methods and material adopted in this study. Such methods are based in a literature review, as showed in the previous chapter (see Chapter 4). These methods are adopted in order to evaluate a new adaptation of the HE method. Therefore, this study aimed to answer the following *research question*:

> *"Can a CHE performed by a Mixed Group result in outcomes whose quality can be considered more similar to the quality of outcomes from a traditional HE with multiple expert inspectors (Benchmark Group) than to the quality of outcomes from a CHE conducted only by novice inspectors (Baseline Group)?"*

In this context, the following hypotheses are presented:

Hypothesis H0 (null): the quality of the outcomes from a *Mixed Group* was more similar to the quality of the outcomes of a *Baseline Group* than to the quality of outcomes from a *Benchmark Group*.

Hypothesis H1: the quality of the outcomes from a *Mixed Group* was equally similar to the quality of the outcomes of a *Baseline Group* and to the quality of outcomes from a *Benchmark Group*.

Hypothesis H2: that the quality of the outcomes from a *Mixed Group* was more similar to the quality of the outcomes of a *Benchmark Group* than to the quality of outcomes from a

Baseline Group.

To answer the *research question* and evaluate the *hypotheses*, evaluators were invited to voluntarily contribute with this study. Their participation was voluntarily, anonymous and restricted to providing CHE written reports expressing their public opinion (GUERRIERO, 2016) about a product (the website evaluated). Additionally, the reports provided by evaluator groups had to present only aggregated information, with no possibility of identifying an individual evaluator.

Therefore, I organized the study design based on the contributions from voluntary evaluators. To control differences among groups, I planned repetitions for each kind of group composition (Baseline, Mixed and Benchmark). The evaluations (reports) provided by the participants were analyzed in order to identify the usability problems pointed out in them. Thereafter, the group reports were analyzed as set of usability problems, and I conducted set operations to calculate metrics and measures planned for data analysis. In summary, I structured a quasi-experimental design as shown in the following section.

5.2 Quasi-Experimental Design

This study has a quasi-experimental design because it lacks random assignment. The assignment of groups of evaluators was based on their self-selection as experts or not, after a careful justification given to the author of this study and his supervisor (SHADISH; COOK; CAMPBELL, 2002, p. 13-14).

To answer the *research question*, I prepared a factorial design (see Table 1). Such a design is based on the reports provided by the groups. This factorial design has four (4) distinct levels of the factor: *presence of an expert in a CHE group*. At each level, I indicate the respective kind of group as expected in the hypothesis. Thus, *level 0* indicates the presence of no expert, *level 1* indicates the presence of only one (1) expert, *level 2* indicates the presence of two (2) experts, and *level 3* indicates the presence of three (3) experts. A *Level 3* is the maximum level considered in the factorial design, because of the difficulty of finding experts that could contribute with this study and, also, join the other evaluators to compose groups and perform the evaluations. For this reason, I assume the premise that such *level 3* produces a standard list of usability problems. Also, such group was asked to perform a standard HE due to our *research question*.

Table 1 – Factorial Design of this study.

	Level 0	**Level 1**	**Level 2**	**Level 3**
Factor	*Baseline*	*Mixed*	*Mixed*	*Benchmark*

Source: Elaborated by the author.

5.3 Website Evaluated

The website evaluated in this study was the Brazilian Ministry of Health web portal, named "*Portal Saúde*". Due to the difficulty of finding evaluators that could voluntarily contribute with this study, and their limitation of time, it was not possible to evaluate more websites. I understand that this fact can limit the results. However, such a choice was made after an indication of an specialist, arguing that this website had sufficient number of usability issues, which would be important for comparisons and data analysis. Also, I chose this website because of its importance for Brazilian society. Therefore, I froze a copy of the entire portal (May, 2016) at a local server so that all evaluations could occur with the same version of the website. Figure 6 presents a screenshot of "*Portal Saúde*".

Figure 6 – Screenshot of "*Portal Saúde*", the website evaluated.

Source: Available on the Internet in May, 2016.

5.4 Participants

A total of 27 evaluators consented to voluntarily contribute with this study. Among them, nine (9) were usability experts and 18 were novices. For the purposes of this study, we considered as novices those professionals that had, at least, an introductory course on Human-Computer Interaction, could understand the terms that compose usability and the heuristics of NIELSEN, J. (1994), and were conducting a HE for the first time after a training session. I adopted such criteria in order to identify professionals that have the minimum requirements to perform a HE as usability professionals. On the other hand, I considered as expert those evaluators with at least

four (4) years of experience in usability area, one publication authored in a vehicle of HCI area, previous usability evaluations and were engaged on researches related to usability. I adopted this criteria based on the Brazilian scenario and on common characteristics, the most experienced evaluators that consented to participate as voluntary in this study.

The 27 evaluators were randomly assigned into seven (7) CHE groups, from G1 to G7, as shown in Table 2. The period of evaluation ranged from May, 2016, to October, 2016. This period was necessary in order to count with the participation of all 27 evaluators. Also, I asked all evaluators to limit their evaluation to one (1) hour. Following the definitions adopted in section 1.3, groups G1 and G2 are *Baseline Groups*, groups G3, G4, G5 and G6 are *Mixed Groups*, and group G7 is a *Benchmark Group*. In this study, such identifications (G1 up to G7) were used to refer to group opinion reports.

Table 2 – Factorial Design including groups of evaluators.

	Level 0	**Level 1**	**Level 2**	**Level 3**
Factor	G1 G2	G3 G4	G5 G6	G7

Source: Elaborated by the author.

Table 3 shows the composition of groups from the same level in the factorial design (columns) according to the participation of novices and experts. Thus, each group indicated in the first row of Table 3 has a composition as described in the respective cell bellow it (same column).

Table 3 – Group compositions.

	G1, G2	**G3, G4**	**G5, G6**	**G7**
Composition	4 novices and 0 experts	3 novices and 1 experts	2 novices and 2 experts	0 novices and 3 experts

Source: Elaborated by the author.

As I could count on the participation of 27 evaluators, the factorial design had only two (2) repetitions per level. Due to the difficulty of finding expert evaluators, I could assign only one group to level 3. Because these experts could not find an available time to get together and conduct a CHE, I asked them to perform a standard HE and send their individual reports. Therefore, I united the usability problems of such reports in order to have the report G7. There were no similar usability problems (considering both relaxed and strict criteria) among the reports provided. Therefore, the G7 is the union of individual reports from three (3) expert evaluators and was considered as the *Benchmark Group*.

Because the hypothesis of this study did not contemplate outcomes from test with users, it was not necessary to include test with users among the methods of this study. The literature usually adopts comparisons against test with users to identify which method (besides the test

with users) has the best outcomes rather than to identify similarities among them. In addition, such comparisons against results from test with users are usually considered the gold standard, but are not the unique standard in the literature. Standard HEs conducted by at least three (3) evaluators are commonly a standard for HEs (NIELSEN, 1992; PREECE; SHARP; ROGERS, 2015). For this reason, at this study, the G7 was considered as a benchmark method and results from other reports (G1 up to G6) were compered to it.

All groups received the same instructions on how to proceed for the evaluations. I describe such procedure in the following section.

5.5 Procedure for the HEs

All evaluations were conducted following the same protocol. The evaluators' evaluation based on the traditional usability heuristics of Nielsen, as rules for evaluation, considering the descriptions showed at Nielsen Norman Group website[1].

Before evaluations, I informed the evaluators about the user profile to be considered, as follows:

> "*For the purpose of this study, we will consider that users with and without disabilities, the elderly and foreign people (English and Spanish speakers) that aim to access and use important information about Brazilian health system.*"

Thereafter, I informed evaluators about particularities of the website evaluated and was responsible for indicating the end of the evaluation period. The time limit for each CHE was one hour.

I provided ten (10) website's tasks for the evaluations. Such an elevated number of tasks was necessary in order to gather enough data for comparisons, because I expected that novice evaluators would find fewer usability problems. Evaluators were free to stop the evaluation before the time limit. The ten (10) website's tasks were as follows:

1. To access information about registration of new users in the Brazil's publicly funded health care system (SUS).
2. To access information about SUS users' rights.
3. To watch an informative video about Brazilian Ministry of Health.
4. To access news about H1N1 flu.
5. To send a message to the Brazilian Ministry of Health through the contact form.

[1] <https://www.nngroup.com/articles/ten-usability-heuristics/>

6. To access information about health for people with disabilities.
7. To access information about elderly health.
8. To access information about how to fight dengue fever.
9. To access information about HIV prevention.
10. To free explore the website.

During evaluations, every group wrote a report with a list of usability problems, their descriptions and locale, and heuristic(s) affected. It was out of the scope of this study to compare severity ratings among groups, because different evaluators can rate the severity differently (SAURO, 2014, p. 24). However, I asked to the experts of groups G3, G4, G5 and G6 (*Mixed Groups*) to indicate and remove any false problem reported by the novices of their groups. Also, they indicated the local of each problem by the URL (Uniform Resource Locator) of its page. The URL was required in order to serve for the matching process during data analyses, described in the next section.

5.6 Data Analysis

The data analysis of this study was organized as follows:

First analysis: *Relaxed criteria analysis*.

Second analysis: *Strict criteria analysis*.

Third analysis: *Cluster Analysis*.

The first and second analysis are composed by two (2) methods: *Matching usability problems* and *Calculating metrics and measures*. Thus, I analyzed the data according to the *F-measure*, as showed in Chapter 4. These two first analysis are based on the premise that the report G7 can be considered as a benchmark (standard). However, it is difficult to affirm that groups G3, G4, G5 and G6 could report a *false alarm*, because they had at least one expert in their composition who reviewed the list of problems before providing the report. Thus, if we assume that experts (as G7) do not report *false alarms*, and considering that the experts invited are indeed experts, no *false alarm* could be found in reports from G3, G4, G5 and G6. To mitigate such limitations, I conducted a *Cluster Analysis*. The *Cluster Analysis* (third analysis) was conducted to compare similarity of reports without classifying *hits*, *misses* and *false alarms*. Instead, I based the *Cluster Analysis* on the discovery of problems from distinct categories of Web usability problems. The categories considered were showed by Petrie and Power (2012):

Physical Presentation, *Content*, *Information Architecture* and *Interactivity*. After, I conducted a *Cluster Analysis* based on the discovery of the most sever problems.

The *cluster analysis* was also planned because it allows to analyze multivariate data. In this context, different variables could not be controlled during the experiment of this study, e.g.:

- Level of influence among evaluators of the same group.
- Level of learning among novices.
- Heuristics used by each evaluator to find the same problem, because distinct heuristics can cover the same usability problem (NIELSEN, 1994).
- Different groups may conduct the CHE through different paths in the website.

The following sections describe each of the methods planned for data analysis.

5.6.1 Matching Usability Problems

For matching usability problems, I adopted the relaxed and strict criteria as showed by Buykx (2009) and Petrie and Power (2012) because they have been applied in similar studies about CHE (HUANG, 2012; BABAJO; PETRIE, 2012; PETRIE; POWER, 2012; PETRIE; BUYKX, 2010; BUYKX, 2009). The URL of usability problems was used in the strict criteria matching in order to verify whether two listed problems occurred at the same page. Two usability researchers conducted this stage independently. In sequence, they discussed the findings and provided a final list of matching.

5.6.2 Calculating Metrics and Measures

This method used the outcomes from *Matching usability problems* to analyze the metrics *hits*, *misses* and *false alarms* of each report (G1 to G7). From such metrics, I calculated the measures *Validity*, *Thoroughness* and *F-measure* of each report. Therefore, I conducted a linear regression of *F-measure*, regarding the distinct levels of the factor, to test if the *F-measure* increased along with the factor.

5.6.3 Cluster Analyses

The goal of a cluster analysis is to identify groups among observations (HAIR *et al.*, 2010; FERREIRA, 2008). According to Hair *et al.* (2010, p.415), cluster analysis is a multivariate method that cluster objects of an observation that are "*more similar to one another then they are to objects in other clusters*". Thus, such a method can indicate whether similarities among reports are in accordance to the hypothesis defined.

The research problem of this cluster analysis was to reveal relationships (similarity) among reports, considering the number of problems pointed out by each group of evaluators. Initially, I based the *cluster analysis* on the discovery of usability problems considering different types of usability problem, as suggested by Hornbæk (2010). For this reason, I adopted the categories of usability problems showed by Petrie and Power (2012): *Physical Presentation*, *Content*, *Information Architecture* and *Interactivity*. I adopted such categories were adopted because they were derived from usability evaluations in governmental websites, as this study.

To match usability problems with a respective category, my supervisor and I compared each usability problem from the reports against the examples of usability problem showed by Petrie and Power (2012, p. 2110-2111) for each category. To classify a problem in a specific category, we performed the following four (4) steps:

Step 1: to choose one usability problems from the reports (G1 up to G7) that was not yet compared against the usability problems showed by Petrie and Power (2012, p. 2110-2111).

Step 2: to compare the chosen problem against all the usability problems showed by Petrie and Power (2012, p. 2110-2111).

Step 3: to discuss and identify which of the usability problems showed by Petrie and Power (2012, p. 2110-2111) was the most similar to the chosen problem.

Step 4: to consider the chosen problem as belonging to the same category as the usability problem showed by Petrie and Power (2012, p. 2110-2111) that was identified as the most similar to the chosen problem.

Thereafter, I classified all usability problems listed in all reports as severe (corresponding to major and catastrophic problems (NIELSEN, J., 1994)) and non-severe problems. My supervisor reviewed such classification. In sequence, I performed a *cluster analysis* considering the similarity of reports considering the number of severe problems indicated in each group report. For such analysis, I considered the list of problems after the relaxed criteria analysis. Such analysis was needed in order to ensure that reports that indicated few, but broadly described problems had the same weight of reports that described many, but specifically described, problems.

To conduct an analysis that would be less sensitive to outlier values (FERREIRA, 2008, p. 400), I adopted an *Average Linkage* method for clustering, with the method Unweighted Pair Group Method with Arithmetic Mean (UPGMA).Also, I adopted the *euclidean distance* as similarity measure. The *euclidean distance* was appropriate to indicate the distance among numbers of problems discovered by each category in a two coordinates space (HAIR *et al.*, 2010, p.431). Finally, I analyzed whether the first four (4) clusters formed were in accordance to the four (4) levels of the factor, as organized in the factorial design (see Table 2).

5.7 Preferences for Evaluation of Hypothesis

At this section, I describe the preferences considered to evaluate the hypothesis of this study. Because this study had three analysis, I organized a sort of preferences to consider the results from each analysis for the evaluation of the hypothesis. Thus, because the *Cluster Analysis* were more appropriated to identify similarities among the reports, I gave it a higher preference than the *F-measure* during the evaluation of hypothesis. Among the different *Cluster Analysis*, the *Cluster Analysis* based on the discovery of more severe problems had also a preference over the others because it is more interesting to practitioners to identify usability problems that have higher severity. Also, all *Cluster Analysis* based on category of usability problems received the same preference in evaluations.

5.8 Final Remarks

In this chapter, I presented descriptions for methods and materials adopted in this study. For this reason, I reviewed the objectives of this study, described the experimental design, showed the website evaluated, described participants, described the procedures adopted for the CHE and for data analysis. Thus, the next chapter presents results of this study according to data analysis considered.

CHAPTER

6

RESULTS

6.1 Introduction

In this chapter, I show the results from the analysis conducted in this study with the CHE reports. The following section describes the usability problems indicated among the reports from the CHE conducted by each group. Moreover, I show the results according to each data analysis (relaxed and strict) as planned in section 5.6.

6.2 HEs Reports

This section describes the raw number of problems (which is referred as *n* at this study) indicated by each group report. As indicated in the previous chapter (Chapter 5), each group of inspectors provided a written CHE report, or a HE report (G7), after their evaluation. Differently from the observations of Huang (2012), the precision of problems described by novice inspectors was similar to the precision of problems described by expert inspectors.

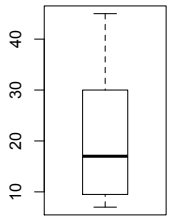
Table 4 shows the number of usability problems (*n*) listed in each report according to the factorial design. In addition, Figure 7 shows a box plot analysis for the distribution of number of problems listed among CHE reports. As shown at such figure, no outlier was detected (*Min.* = 7.00, *1st Qu.* = 9.50, *Md* = 17.00, *3rd Qu.* = 30.00, *Max.* = 45.00).

Table 4 – Number of usability problems listed in each report according to its respective level of the factorial design.

	Level 0	**Level 1**	**Level 2**	**Level 3**
Factor	G1: 7 G2: 8	G3: 45 G4: 17	G5: 11 G6: 22	G7: 38

Source: Elaborated by the author.

Figure 7 – Box plot for the number of problems listed among group reports. No outlier was detected.



Source: Elaborated by the author.

The following sections describe the three stages of data analysis (see section 5.6) conducted with the reports.

6.3 Relaxed Criteria Analysis

This section shows data analysis conducted with CHE and HE reports considering the relaxed criteria. For this reason, all subsections presented here follow such criteria. For these criteria, usability problems were considered similar if they referred to the same problem, or to the same design element, independent of the level of abstraction. If the same underlying problem is described, two usability problems were considered as similar (see subsection 5.6.1).

This analysis was conducted by the author of this study and, later, revised by his supervisor. At the beginning of this analysis, we noticed that some reports had fewer problems because groups did broad descriptions, while other groups reported more problems because they did narrow descriptions. Thus, to perform a fair analysis, we first compared all usability problems without distinguishing them among the groups. Consequently, narrow problems described in the same report could be considered as similar if they satisfied the relaxed criteria.

Table 5 shows the results from relaxed analysis. Thus, it shows a comparison between raw number of problems (n) listed against the number of distinct problems after relaxed criteria matching ($n2$). Moreover, Table 5 shows an *index of reduction (IR)*. Such index is described at Equation 6.1. The *IR* represents the percentage of n that was relaxed to $n2$. Thus, it allowed us to compare how narrow and how broad were the descriptions of each report. Thus, narrow descriptions described a specific usability problem in a specific element of the interface, while a broad description indicated a general usability problem that could be verified in different usability problem instances, in different levels of abstractions and among distinct elements of

the interface.

$$IR = 100 - 100 * n2/n \tag{6.1}$$

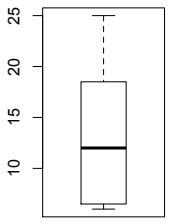
Table 5 – Distribution of raw numbers of usability problems (*n* (raw)), number of distinct usability problems after relaxed matching (*n2* (distinct)) and the *Index of Reduction (IR)* for each report.

Reports	***n***	***n2***	**IR (%)**
G1	7	6	≈ 14.29
G2	8	6	25
G3	45	21	≈ 53.33
G4	17	12	≈ 29.41
G5	11	7	≈ 36.36
G6	22	16	≈ 27.27
G7	38	25	≈ 34.21

Source: Elaborated by the author.

Comparing the distributions *n* and *n2*, an initial analysis showed that the standard deviation (s) *n2* ($\bar{x} \approx 13.29$, $s \approx 7.65$) is almost 50% lower than the standard deviation of set *n* ($\bar{x} \approx 21.14$, $s \approx 14.98$). This fact shows that values of *n2* are closer among each other than values from the distribution *n*. Similarly, Figure 8 shows a box plot analysis for values of *n2* among reports (*Min.* = 6.00, *1st Qu.* = 6.50, *Md* = 12.00, *3rd Qu.* = 18.50, *Max.* = 25.00). As shown in Figure 8, no outlier was detected among *n2* values. Also, one can see that the quartiles in Figure 8 are shorter than the quartiles showed previously at Figure 7.

Figure 8 – Box plot for the number of distinct problems per group reports after relaxed matching (*n2*). No outlier was detected (*Min.* = 6.00, *1st Qu.* = 6.50, *Md* = 12.00, *3rd Qu.* = 18.50, *Max.* = 25.00).



Source: Elaborated by the author.

Furthermore, a comparison between the distributions *n* and *IR* can indicate whether reports with higher number of usability problems also had the highest *IR* values. Such comparison can be showed through a correlation analysis. Thus, a Shapiro-Wilk test was first conducted and showed that the distributions *n* ($p - value \approx 0.2$) and *Index of Reduction* ($p - value \approx 0.75$)

might come from a normal distribution. Therefore, a Pearson Correlation test showed a strong positive correlation between values of n and the IR ($r \approx 0.79$) among reports. This indicates that, for the observations of this study, reports that had more problems (higher n) also had a higher IR, and reports that had fewer problems (lower n), also had a lower IR.

The analyses showed in this section were initial and complementary to the planned analyses. The goal of such analyses was to present an initial description of how relaxed criteria processed initial lists of usability problems showed at the reports. Thus, the values of *n2* distribution were used for the planned relaxed analysis. Thereafter, the following subsections show calculus of metrics (*Hits*, *Misses* and *False Alarms*) and measures (*Validity*, *Thoroughness* and *F-measure*) based on *n2* values (see Table 5) among reports from all groups.

6.3.1 Hits, Misses and False Alarms

As planned in section 5.6, this section presents results from relaxed analysis on metrics *hits*, *misses* and *false alarms* for each report. As planned, such calculus considered report G7 as a benchmark/standard (see section 4.2). Table 6 shows the number of *hits*, *misses* and *false alarms* accounted from each group report.

Table 6 – Number of *hits*, *misses* and *false alarms* by each group report.

Reports	Hits	Misses	False Alarms
G1	2	23	4
G2	3	22	3
G3	9	16	12
G4	5	20	7
G5	1	24	6
G6	6	19	10

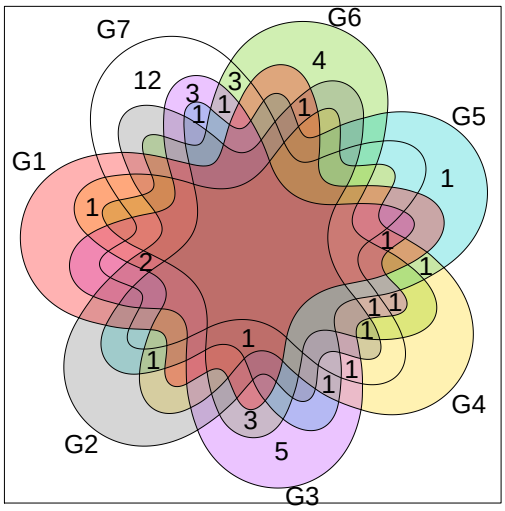
Source: Elaborated by the author.

To better describe similarities and differences among reports, I plotted Figure 9, Figure 10a, Figure 10b and Figure 10c. Figure 9 is a Venn diagram showing intersection numbers among usability problems listed in reports from all groups (zero (0) values are not shown for better presentation). The reason of Figure 9 is to illustrate two important facts: *number of problems uniquely reported in the benchmark report (G7)* and *number of problems reported by all reports*. As a result, it shows that report G7 listed 12 usability problems that were not listed in any other report, considering the relaxed analysis. Also, it shows that no usability problem was reported by all reports (represented by the empty spaces in the intersections among all reports).

Figure 10a, Figure 10b and Figure 10c show detailed Venn diagrams comparing reports from the same level in the factorial design with the benchmark report (G7).

10a shows the intersections among reports G1 and G2 (from level 0 of the factorial design) and report G7. In this case, G1 and G2 had two (2) *hits* in common. Also, G2 had a

Figure 9 – Venn diagram for intersections among all reports considering usability problems listed.



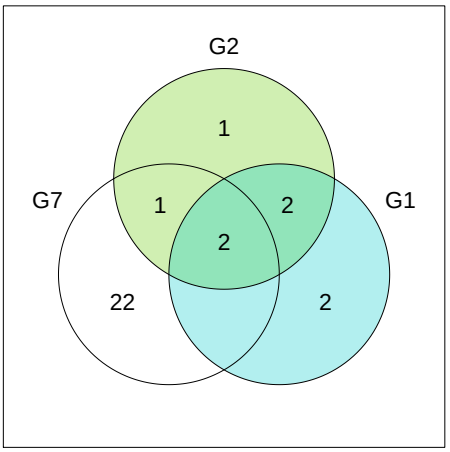
Source: Elaborated by the author.

unique *hit*, while G1 had no one unique *hit*. The union of reports G1 and G2 could only hit three (3) out of the 22 usability problems listed in report G7 (only 12% of the problems listed in G7). This coverage of novice reports in expert reports is lower than the coverage described by Othman *et al.* (2014), when novice inspectors found from 30% to 35% of the problems found by experts after a CHE. Also, uniting reports G1 and G2 results in five (5) *false alarms*. Such *false alarms* were reviewed by the author of this study and his supervisor. Thereafter, we discussed and concluded that all of these five (5) usability problems indeed exist in the website evaluated. This apparent contradiction, novice inspectors reporting existent usability problems that experts did not report, may have occurred for different reasons. As an example, such novices may have gone through a different path of the website where other problems existed. Also, other expert groups might have found such problems, but the sample of this study was limited.

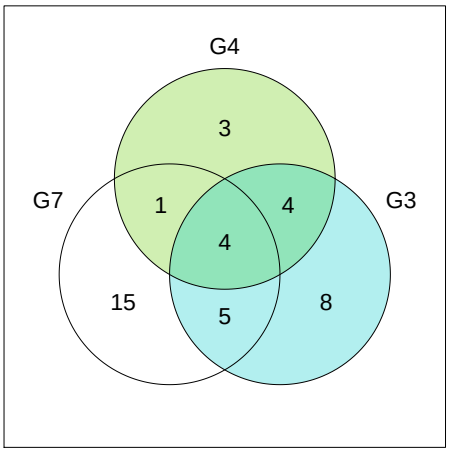
In sequence, Figure 10b shows the intersections among reports G3 and G4 (from level 1 of the factorial design), and report G7. Reports G3 and G4 had four (4) *hits* in common. Between G3 and G4, G3 had more unique *hits* (five (5)) than G4 (one (1)). The union of reports G3 and G4 has ten (10) *hits* out of the 25 usability problems listed in G7 (40% of the usability problems listed in report G7). This coverage of novice reports in expert reports is higher than the coverage described by Othman *et al.* (2014), when novice reports listed from 30% to 35% of the problems listed in expert reports. Also, the union of reports G3 and G4 had 15 *false alarms*.

Finally, Figure 10c represents the intersections among reports G5 and G6 (from level 2 of the factorial design), and report G7. In this comparison, G5 and G6 had no *hits* in common. Also, G6 reported six (6) unique *hits*, while G5 reported only one (1). The union of reports G5 and G6 had seven (7) *hits* (28% of the usability problems listed in report G7). This coverage is similar to the coverage reported by Othman *et al.* (2014), when novice inspectors found from 30% to

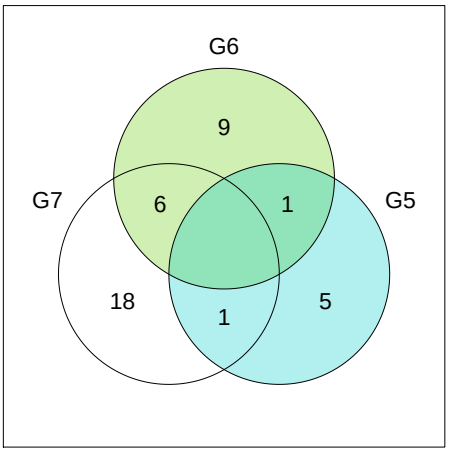
Figure 10 – Venn diagrams representing intersections among reports of the same level from the factorial design with the benchmark report (G7).



(a) Intersections among G7, G1 and G2



(b) Intersections among G7, G3 and G4



(c) Intersections among G7, G5 and G6

Source: Elaborated by the author.

35% of problems listed by experts. However, four (4) expert and four (4) novice inspectors wrote reports G5 and G6. Also, such finding is contradictory because such union leads to more experts (4) than the number of experts that wrote report G7 (3). However, it may represent the *evaluator-effect* among such reports, evidencing individual differences among experts. This fact is also interesting because the union of G3 and G4 (involving only two (2) experts) had 12% more *hits* than the union of G5 and G6, this may also be due to the *evaluator-effect*.

Based on the metrics calculated at this section, the following section presents calculus for the measures *Validity*, *Thoroughness* and *F-measure*.

6.3.2 Validity, Thoroughness and F-measure

This section presents calculus of *Validity*, *Thoroughness* and *F-measure* based on the metrics of *hits*, *misses* and *false alarms* showed in the previous section. As mentioned before, these metrics were calculated after a relaxed matching procedure. The formulas for such measures were presented at section 4.2.

Table 7 presents the measures of *Validity*, *Thoroughness* and *F-measure* per group report. As the *F-measure* describes a weighted combination of *Validity* and *Thoroughness* ($\alpha = 0.5$, see section 4.2), I plotted Figure 11 as a mean of illustrating the differences on *F-measure* according to each level of the factorial design (see section 5.2). As shown in Figure 11, among the observations of this study, the distance from the mean *F-measure* increased as the level of the factor increased. This may have occurred because of the limited sample of reports.

Table 7 – Measures of *Validity*, *Thoroughness* and *F-measure* by each group report.

Reports	Validity	Thoroughness	F-measure
G1	0.3333333	0.08	0.1290323
G2	0.5	0.12	0.1935484
G3	0.4285714	0.36	0.3913043
G4	0.4166667	0.2	0.2702703
G5	0.1428571	0.04	0.0625
G6	0.375	0.24	0.2926829

Source: Elaborated by the author.

The following section presents strict analysis of data. As planned, it shows metrics and measures based on such matching criteria.

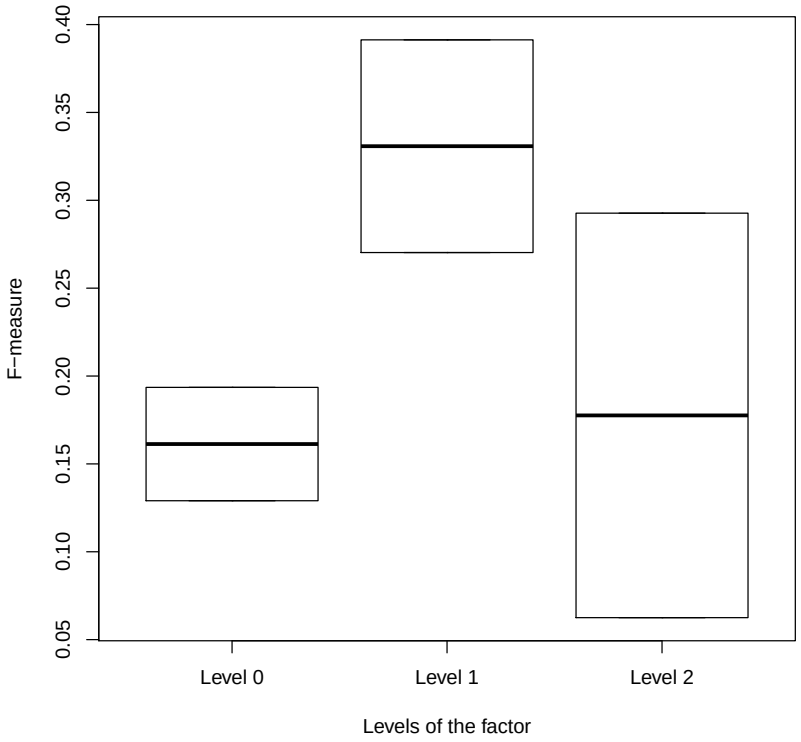
6.4 Strict Criteria Analysis

This section shows results from the strict data analysis planned. All metrics and measures presented here were based on the strict matching procedure. The strict matching criteria is a more rigorous criteria compared to the relaxed one. By the strict matching criteria, usability problems are considered similar only if they refer to the same problem, to the same element of design, and the description is at the same level of abstraction (see subsection 5.6.1). For this reason, in contrast to the relaxed analysis, no relaxation (dispensation) is allowed.

The strict analysis was performed by the author of this study and revised by his supervisor. Table 8 shows the number of usability problems listed by each report provided by the groups. The values in column *n* represent the raw number of usability problems reported, this is the same distribution as showed in Table 5. As showed in Figure 7, no outlier exists among values of *n*.

The following subsections show calculus of metrics (*Hits*, *Misses* and *False Alarms*) and measures (*Validity*, *Thoroughness* and *F-measure*) based on the strict matching process.

Figure 11 – F-measure of reports grouped by each level of factorial design.



Source: Elaborated by the author.

6.4.1 Hits, Misses and False Alarms

This section shows values of *Hits*, *Misses* and *False Alarms* based on the strict matching procedure. Table 9 shows the values of each metric according to its respective group report. Such table shows an interesting fact: CHE reports done only by novice inspectors had more *hits* than two of the *Mixed Group* reports (G4 and G5). This fact is interesting because G4 and G5 together had three (3) expert inspectors. Considering the observations of this study, it is not possible to

Table 8 – Number of usability problems (*n*) listed by each report as provided by the groups.

Reports	***n***
G1	7
G2	8
G3	45
G4	17
G5	11
G6	22
G7	38

Source: Elaborated by the author.

understand the cause of such difference. This may be due to the small sample of reports and websites considered. Also, the mentioned difference is of only two (2) usability problems (one from each G1 and G2). This may be explained by the path among the website pages that G1 and G2 chose to conduct the inspection at the website.

Table 9 – Number of *hits*, *misses* and *false alarms* by each group report.

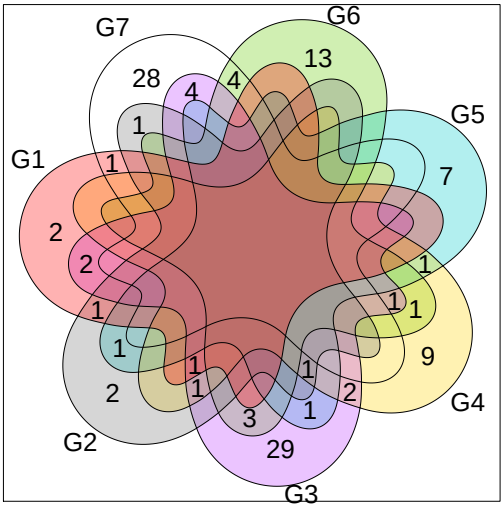
Reports	Hits	Misses	False Alarms
G1	1	37	6
G2	1	37	7
G3	4	34	41
G4	0	38	17
G5	0	38	11
G6	4	34	18

Source: Elaborated by the author.

As a complement for Table 9, I plotted Figure 12, Figure 13a, Figure 13b and Figure 13c to provide detailed information about intersections and differences among reports.

Figure 12 illustrates the number of usability problems that, considering the strict criteria, were uniquely listed by the benchmark report (G7). At Figure 12, one can see that report G7 listed 28 usability problems that were not listed in any other report. In other words, more than 70% of the problems listed in G7 were not listed in any other report. Figure 12 also indicates the number of usability problems listed uniquely by each one of the other reports.

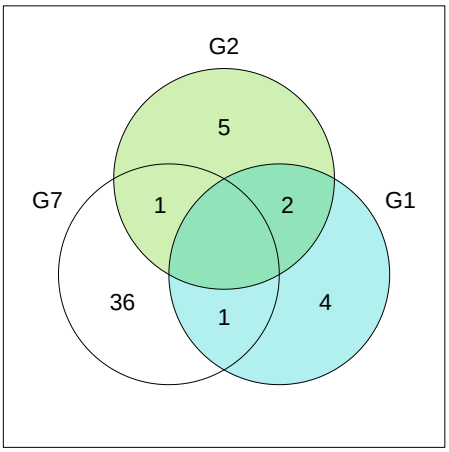
Figure 12 – Venn diagram for intersections among all reports considering usability problems listed.



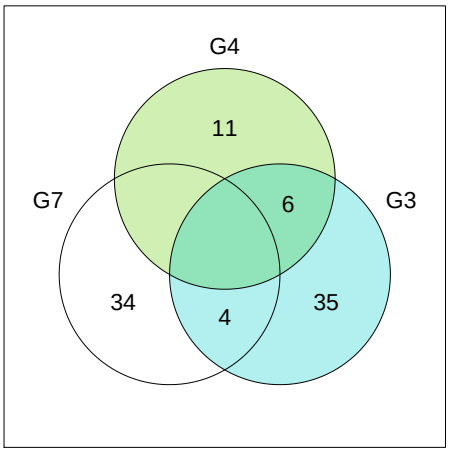
Source: Elaborated by the author.

Similarly, Figure 13a, Figure 13b and Figure 13c provide a detailed view of intersections among reports from the same level of the factorial design against the benchmark report (G7).

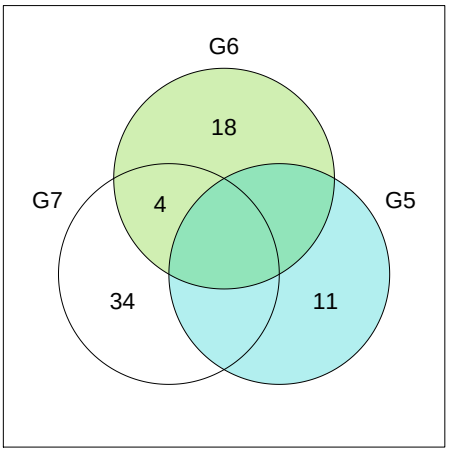
Figure 13 – Venn diagrams representing intersections among reports of the same level from the factorial design with the benchmark report (G7).



(a) Intersections among G7, G1 and G2



(b) Intersections among G7, G3 and G4



(c) Intersections among G7, G5 and G6

Source: Elaborated by the author.

Figure 13a shows the intersections among reports G1 and G2 (from the level 0 of the factorial design), and G7. In such comparison, G1 and G2 had no *hits* in common. On the other hand, G1 and G2 had one (1) distinct *hit* each. The union of both reports (G1 and G2) resulted in two (2) *hits* ($\approx$ 5% of the problems listed in G7), which is much less than the coverage showed by Othman *et al.* (2014). In addition, such union had 11 *false alarms* ($\approx$ 85% of problems listed in both G1 and G2).

In sequence, Figure 13b shows the intersections among reports G3 and G4 (from level 1 of the factorial design), and G7. As shown in the figure, G3 had four (4) *hits* while G4 had no one. In this case, the union of reports from level 1 cover less than 11% of problems listed in report G7, which is about one/third of the coverage showed by Othman *et al.* (2014) for groups composed only by novice evaluators. On the other side, the matching criteria adopted by Othman

et al. (2014) is not clear enough.

Considering Figure 13b, we analyzed the description of the *hits* from report G3. Thus, we found that two out of these four *hits* were usability problems for users with disabilities. Thus, among these observations, knowledge about accessibility played an important role in the performance of groups with only one expert. Figure 13b also shows that the union of reports G3 and G4 had 52 *false alarms* ($\approx$ 93% of problems listed in both G3 and G4).

Figure 13c represents the intersections among G5 and G6 (from level 2 of the factorial design), and report G7. In this comparison, G6 had four (4) *hits* while G5 had no one. An interesting fact is that G5 and G6 had no intersection. The union of G5 and G6 produced the same number of *hits* that the union of G3 and G4 produced. This fact may indicate that the difference in quality of reports from CHE conducted by groups with only one expert are similar to the quality of reports produced by CHE groups with two (2) experts. The percentage of *false alarms* produced by reports G5 and G6 ($\approx$ 88%) was also similar to the percentage of *false alarms* produced in reports G3 and G4 ($\approx$ 93%). Such facts raise an additional question: *Are the reports produced by CHE groups composed by only one expert significantly similar to CHE reports produced by groups composed by two (2) experts?* The results of this study indicate that such reports may be similar, depending on confirmatory analysis of the similarity among inspectors' expertise. Future studies are needed in order to understand the probability of such occurrence.

The following section basis on the metrics showed here to calculate measures of *Validity*, *Thoroughness* and *F-measure*.

6.4.2 Validity, Thoroughness and F-measure

At this section, I present the values of *Validity*, *Thoroughness* and *F-measure* calculated for each report in comparison to the benchmark (G7). The section 4.2 presented the formulas for each of these measures. As a result, Table 7 indicates each of the measures and its respective report. Because reports G4 and G6 had no *hits* under the strict criteria, such reports did not have the F-measure calculated.

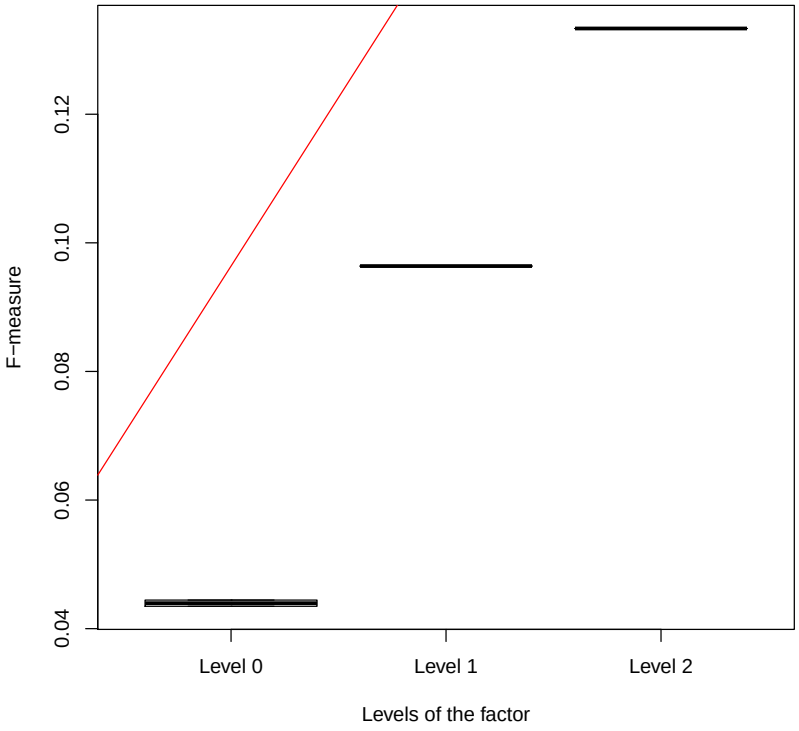
Table 10 – Measures of *Validity*, *Thoroughness* and *F-measure* by each group report.

Reports	Validity	Thoroughness	F-measure
G1	0.1428571	0.02631579	0.04444444
G2	0.125	0.02631579	0.04347826
G3	0.08888889	0.10526316	0.09638554
G4	0	0	–
G5	0	0	–
G6	0.1818182	0.1052632	0.1333333

Source: Elaborated by the author.

Figure 14 illustrates the F-measure per level of the factor, reports G4 and G5 could not be considered. Analyzing such figure, one can see that the F-measure increases as the level of the factor increases. To verify such correlation, I conducted a Linear Regression (indicated at Figure 14 by the red crossing line). For our observations, there was a significant correlation between presence of an expert and the level of F-measure ($p - value \approx 0.009$). However, future studies are needed to explore the probability of such model.

Figure 14 – F-measure of reports grouped by each level of factorial design. Linear Model for F-measure versus levels of the factor represented by the red crossing line ($p - value \approx 0.009$).



Source: Elaborated by the author.

The next section presents results from data analysis using *cluster analyses* of reports.

6.5 Cluster Analysis

At this section, I present the results of *cluster analyses* conducted. The goal of this *cluster analysis* was to verify whether the similarities among CHE reports according to different aspects of quality (category of problems listed and its severity). Also, the results from the *cluster analysis* were compared against similarities expected: *reports of the same level of the factorial design should list similar usability problems* (see Table 2).

The first set of *cluster analyses* based on categories of problems reported. Thus, the similarity considered referred to the number of problems, from each category, listed in each report. The categories of problems adopted were: *Physical Presentation*, *Content*, *Information Architecture* and *Interactivity*. Such categories were retrieved from the study of Petrie and Power (2012), which also referred to the evaluation of governmental websites.

In this context, Table 11 shows the number of problems listed in each report divided by each category. This *cluster analysis* required the use of problem lists without relaxation (as in the strict analysis), because a relaxed matching could join problems from distinct categories if they referred to the same design element, or the same underlying problem. For this reason, this analysis adopt the *n* distribution of raw problems, as in the strict criteria analysis at section 6.4.

Table 11 – Number of usability problems discovered by category (PETRIE; POWER, 2012) and the respective group report.

Reports	Physical Presentation	Content	Information Architecture	Interactivity	Total (n)
G1	1	1	3	2	7
G2	2	2	1	3	8
G3	4	15	3	23	45
G4	1	4	6	6	17
G5	0	4	2	5	11
G6	9	5	3	5	22
G7	14	9	5	10	38

Source: Elaborated by the author.

Thereafter, I conducted a *cluster analysis* comparing similarity based on the discovery of severe problems. At this time, I needed to consider the set of problems after the relaxed matching. This was needed because similar problems, or the same underlying problem, could not be rated with different severity. Also, reports that listed high number of narrow described problems would not be in advantage compared to reports that listed low number of broad described problems. Thus, the distribution *n2* was considered for this analysis. Table 12 shows the number of severe problems listed in each report.

The following sections present the dendrograms indicating the clusters formed after each *cluster analysis*. As indicated at subsection 5.6.3, the UPGMA method was adopted for such analysis.

6.5.1 Discovery of Physical Presentation Problems

Figure 15 shows the number of *Physical Presentation* problems listed in reports, according to each level of the factorial design. Analysing such figure, one can see that a linear correlation may occur between number of *Physical Presentation* problems listed and the level of the factorial design. However, I carried out a linear regression for such data and no significance

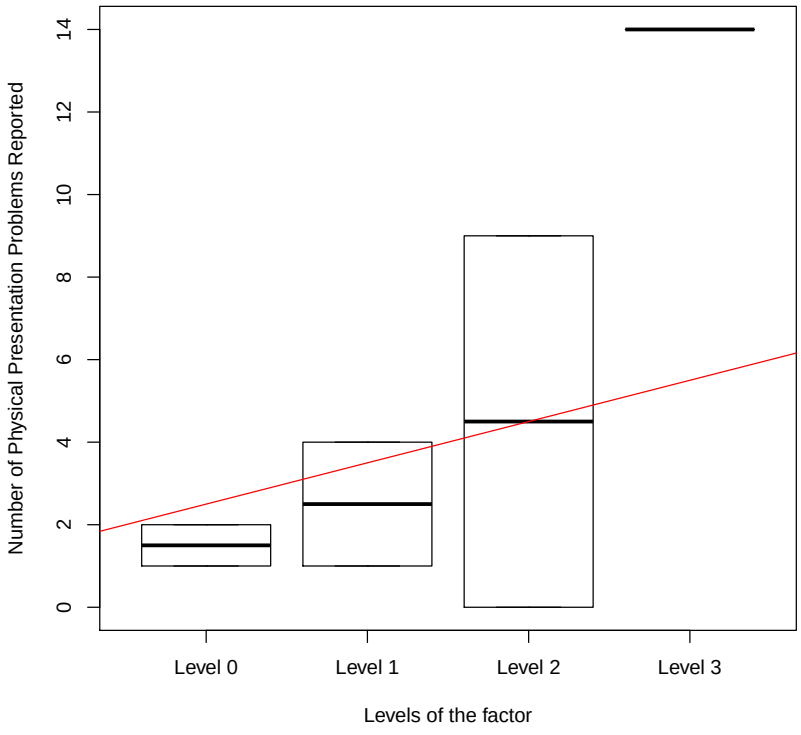
Table 12 – Number of severe problems listed by each report according to the relaxed criteria (*n2*).

Report	Severe	n2
G1	4	6
G2	4	6
G3	14	21
G4	9	12
G5	7	7
G6	13	16
G7	17	25

Source: Elaborated by the author.

for such linear model was found ($p - value = 0.2308$). Further studies are needed in order to understand into more depth the probability of such linear model.

Figure 15 – Distribution of the number of *Physical Presentation* problems reported by each level of the factorial design. Linear regression indicated by the red crossing line ($p - value = 0.2308$).



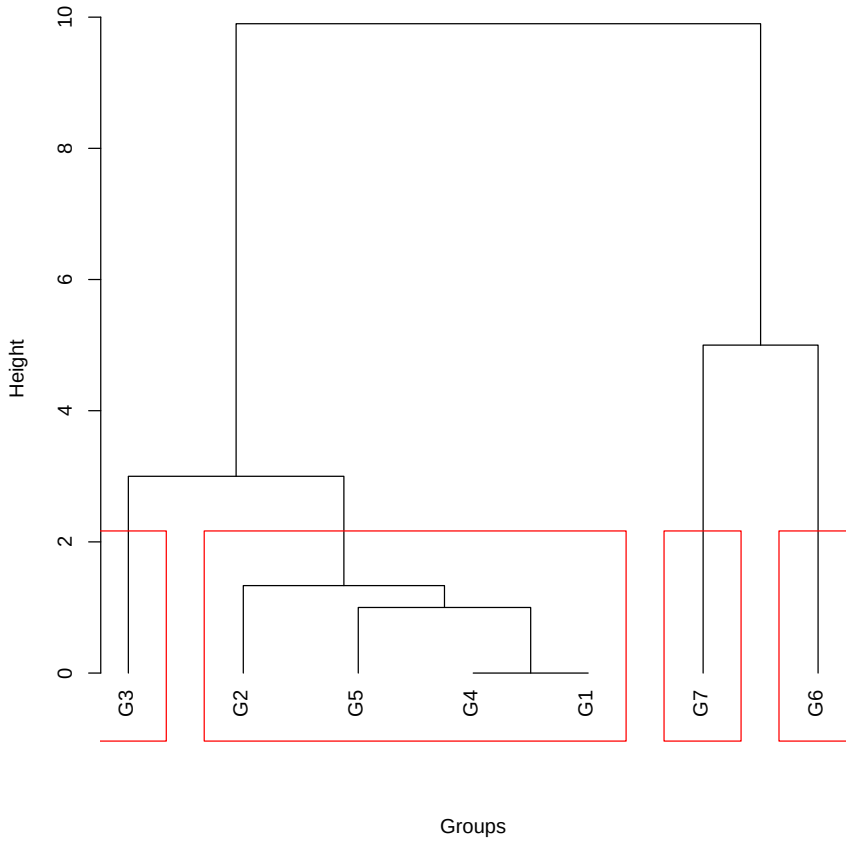
Source: Elaborated by the author.

Another interesting fact from Figure 15 is that the distance from the mean discovery of *Physical Presentation* problems increased as the level of the factor increased. Although this study approached a limited sample of reports and websites, this fact may show evidence that the *evaluator-effect* plays higher influence in reports from groups composed with more experts.

Future studies can investigate such evidence into more depth.

The values represented in Figure 15 served as basis for the *cluster analysis*. Figure 16 shows the results from the UPGMA cluster analysis considering such values. The clusters were indicated with the red square surrounding reports.

Figure 16 – UPGMA dendrogram considering similarity of reports in discovery of usability problems that belong to the *Physical Presentation* category.



Source: Elaborated by the author.

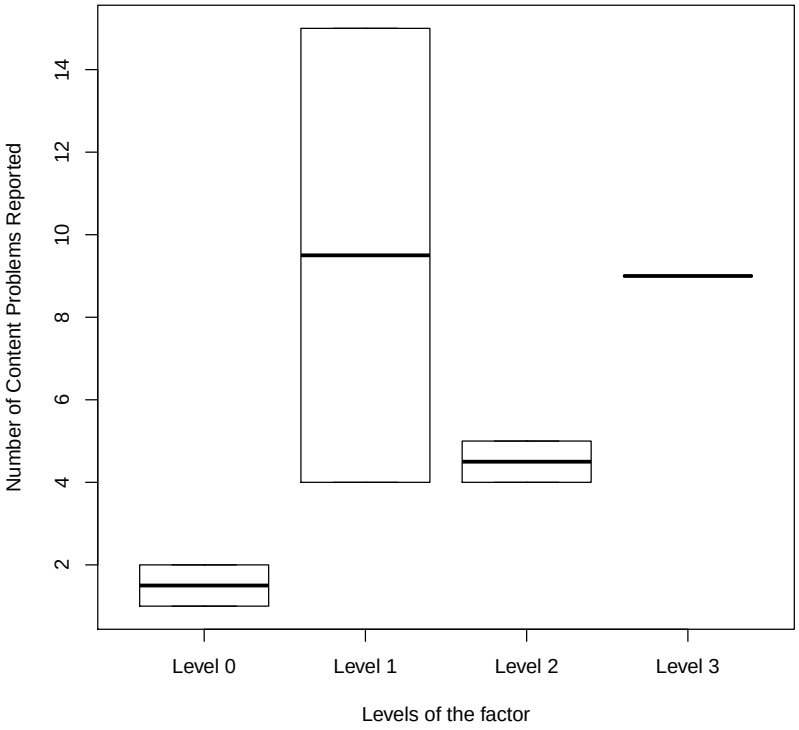
The dendrogram of Figure 16 shows that G6 is the most similar to G7 in the discovery of *Physical Presentation* problems. In sequence, G3 is the most similar to G6 and G7. All other reports were grouped in the same cluster. Regarding the factorial design organized, only G1 and G2 (level 0 of the factorial design) and G7 (level 3 of the factorial design) were clustered in accordance to the factorial design.

Finally, Figure 16 shows that a CHE *Mixed Group* composed by one (1) or two (2) experts can produce reports of quality compared to reports produced only by novice inspectors (see the cluster G2, G5, G4 and G1). Future studies can explore these findings into more depth and understand the probability of such occurrences.

6.5.2 Discovery of Content Problems

Figure 17 shows the number of *Content* problems listed in reports of each level of the factorial design. As showed in the figure, it was observed a higher variance between values from level 1 (G4 and G3 in comparison to values of other levels. For this reason, I conducted a box plot analysis to verify the existence of an outlier (see Figure 18). As a result, the box plot showed that the number of *Content* problems listed in report G3 may be an outlier among such distribution. Nevertheless, as explained in subsection 5.6.3, the UPGMA cluster analysis is among the cluster analysis methods that are less sensitive to the outlier values.

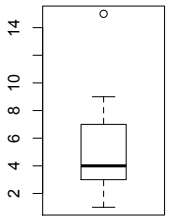
Figure 17 – Distribution of number of *Content* problems reported by each level of the factorial design.



Source: Elaborated by the author.

Thereafter, Figure 19 shows the dendrogram resulted form the UPGMA cluster analysis considering the discovery of *Content* problems. Although the UPGMA method is very insensitive to outlier values (FERREIRA, 2008, p. 400), it seems that in our analysis such outlier was determinant for cluster compositions. Despite the possible influence of the outlier, the other reports seems to be clustered in accordance to the factorial design. Additionally, the similarity observed between G4 and G5 may be an evidence that the quality of outcomes between a CHE *Mixed Group* composed by one (1) or two (2) experts can be highly similar. Future studies can explore the probability of such findings regarding larger samples of reports and websites.

Figure 18 – Box plot for the number of *Content* problems listed in each group report. One outlier detected, the G3 report (*Min.* = 1.00, *1st Qu.* = 3.00, *Md* = 4.00, *3rd Qu.* = 7.00, *Max.* = 15.00).



Source: Elaborated by the author.

6.5.3 Discovery of Information Architecture Problems

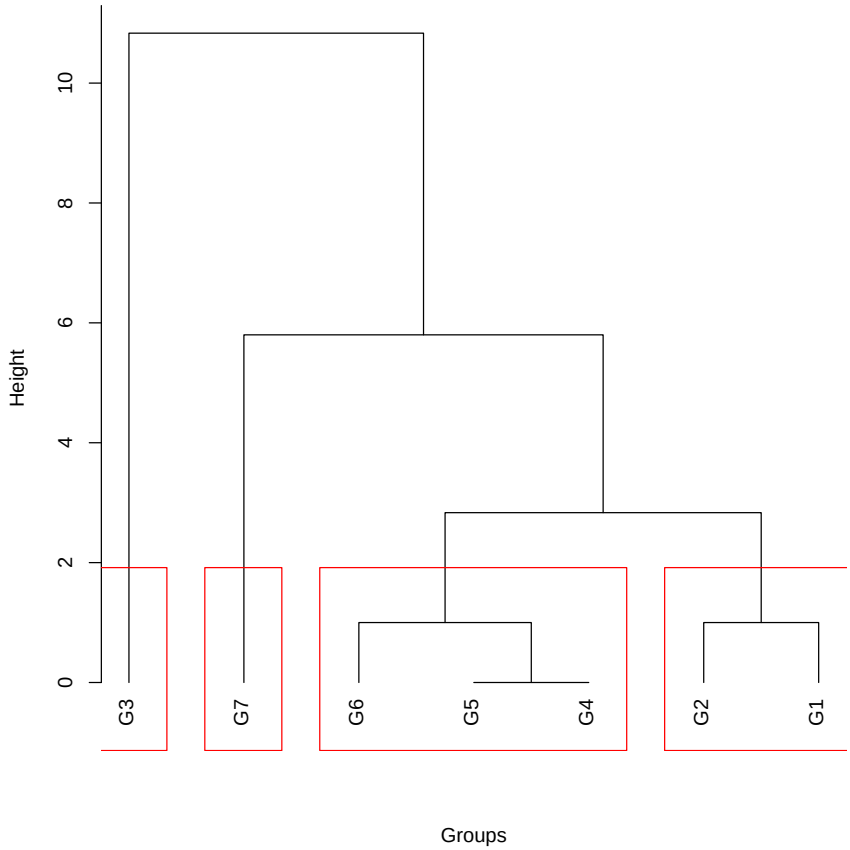
The number of *Information Architecture* problems listed by each report, according to the factorial design, is presented at Figure 20. At Figure 20, one can see that one of the reports from level 1 found a higher number of *Information Architecture* problems than report G7 (*Benchmark*). This fact may be due to the small sample of reports and websites considered in this study. In addition, as showed in the figure, one report from each of the first three level listed the same number of *Information Architecture* problems. This may show evidence that the discovery of *Information Architecture* problems is less dependent on the *expertise-effect* than the discovery of other categories of usability problems. Also, this fact may be due to the method (CHE), future studies can investigate whether HE methods are appropriate to find *Information Architecture* problems. Nevertheless, such hypothesis still needs exploration in future studies.

Figure 21 is the dendrogram resulted from the UPGMA cluster analysis regarding discovery of *Information Architecture* problems. This dendrogram is particularly interesting because none of the clusters formed were in accordance to the factorial design. This finding shows that CHE *Baseline* groups can produce reports of similar quality in terms of discovery of *Information Architecture* problems. As in Figure 20, this fact rises an additional question: *is the discovery of Information Architecture problems independent from the presence of an expert in a CHE group?* Future studies can explore this question.

6.5.4 Discovery of Interactivity Problems

Figure 22 shows the distribution of number of *Interactivity* problems listed by each report, regarding its respective level of factorial design. Analyzing this figure, one can see that the difference between values of level 1 is much higher than the difference of values from other levels. This may indicate the presence of an outlier. For this reason, I conducted a box plot analysis (see Figure 23), which indicated that report (G3) may be an outlier.

Figure 19 – UPGMA dendrogram considering similarity of reports in discovery of usability problems that belong to the *Content* category.



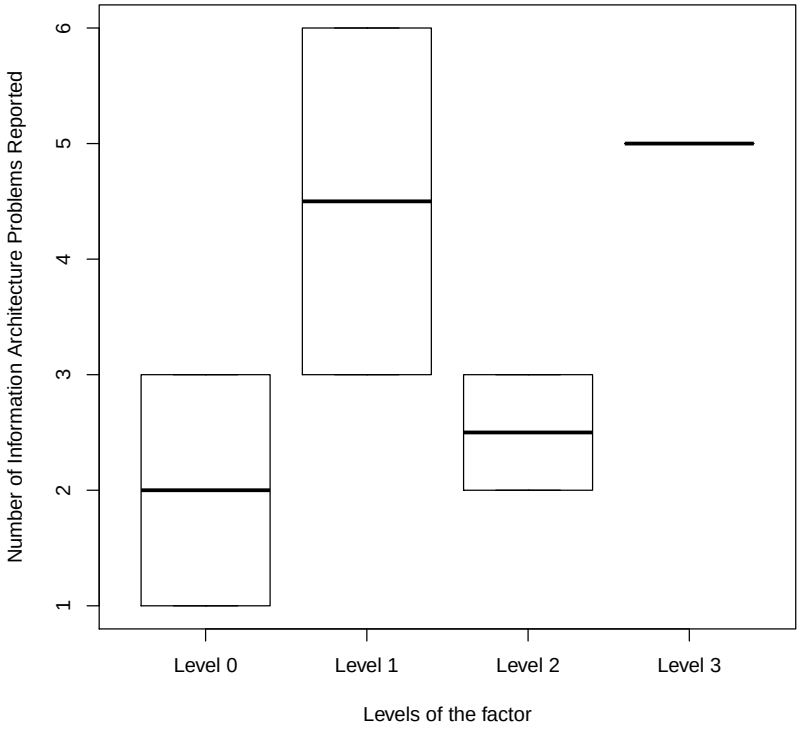
Source: Elaborated by the author.

Thereafter, Figure 24 shows the dendrogram resulted from the cluster analysis. Such dendrogram also seems to be impacted by the presence of the outlier (G3). For this reason, the position of G3 among the clusters formed may be biased. Despite this fact, the formed clusters with other reports were close to the factorial design. Nevertheless, G4 was closer to G7 than reports from level 2 (G5 and G6). This fact may be an evidence that the similarity on discovery of *Interactivity* problems by CHE *Mixed Groups* with one (1) or two (2) experts might not compensate the cost of having two (2) experts instead of only one (1) in a *Mixed Group*. Future studies can explore this topic.

6.5.5 Discovery of Severe Problems

The number of severe problems listed in each report was shown at Table 12. I plotted such distribution at Figure 25, regarding the respective levels of the factor. Figure 25 indicates that the values of reports from level 1 and 2 are close to each other. Also, it shows that the difference between values of level 1 and the difference between values of level 2 are similar. Therefore, I tested a linear model between the number of discovery of severe problems and the level of the factor. However, it resulted in a non significant linear model ($p - value = 0.1386$).

Figure 20 – Distribution of the number of *Information Architecture* problems reported by each level of the factorial design.

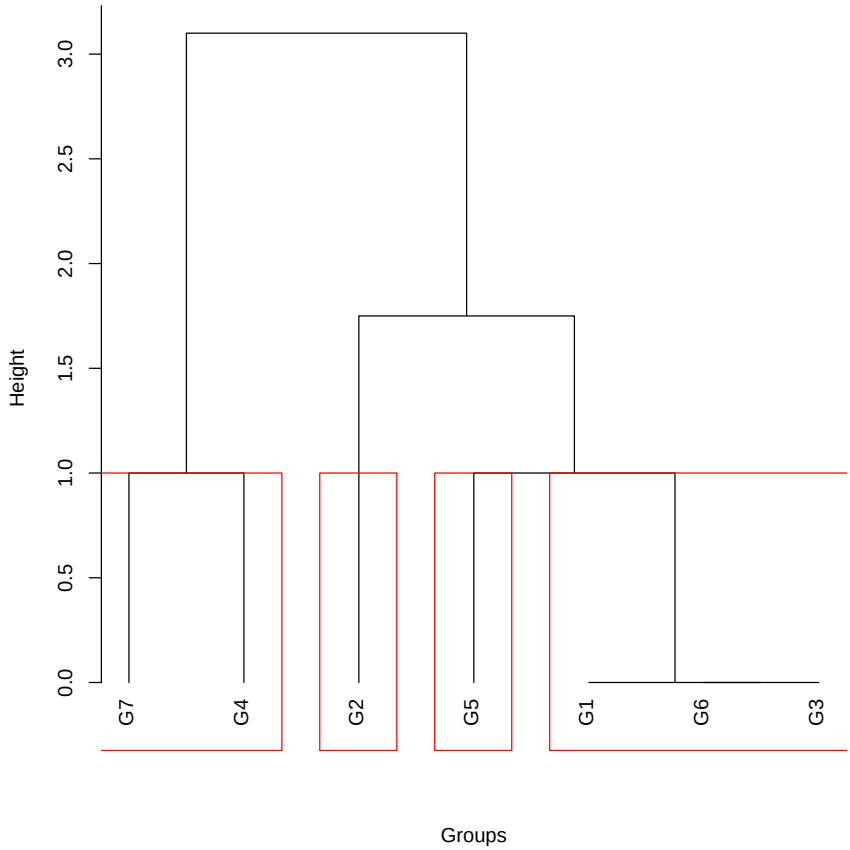


Source: Elaborated by the author.

On the other hand, this linear model is more significant than the linear model showed at Figure 15 (regarding the discovery of *Physical Presentation* problems). This fact raises a new question: *is the amount of severe problems discovered influenced by the number of experts in a Mixed Group?* Future studies can also explore this question.

Figure 26 presents the resulted UPGMA clusters from this analysis. Report G7 was not clustered with any other report, as explained by the factorial design. Additionally, the clusters showed that levels 1 and 2 had one report each among the most similar to G7; reports G6 and G3 were equally similar to G7. This finding may indicate that, despite the *evaluator-effect* among experts, the *expertise-effect* can play an important role in CHE groups of novice inspectors. In other words, the presence of only one expert among novices can enhance the quality of CHE outcomes, in terms of discovery of severe problems, and make it more similar to outcomes from a *benchmark* group than to outcomes from a *baseline* group. Future studies are required in order to understand the probability of such finding to occur, considering larger samples of reports and websites.

Figure 21 – UPGMA dendrogram considering similarity of reports in discovery of usability problems that belong to the *Information Architecture* category.

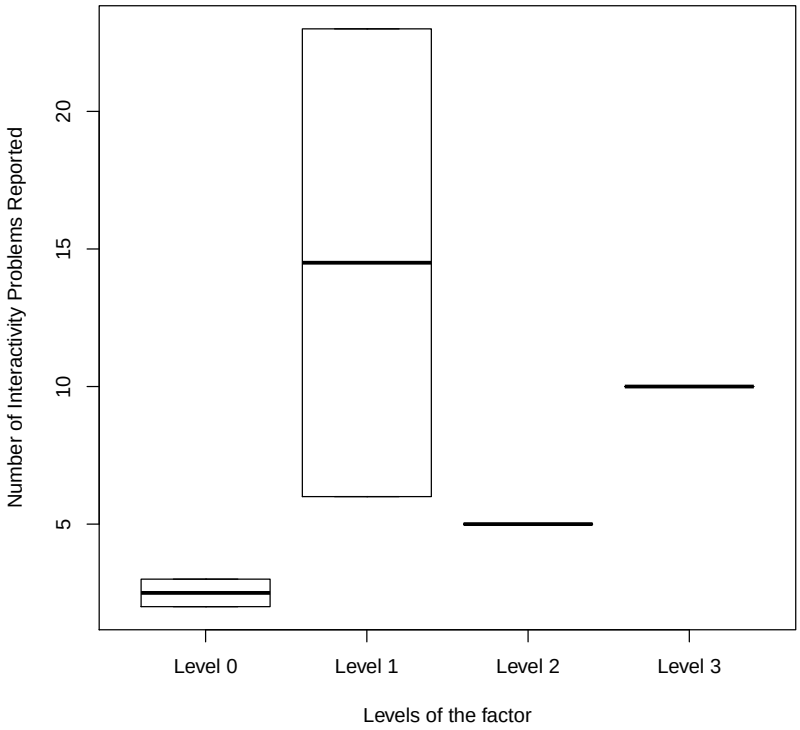


Source: Elaborated by the author.

6.6 Final Remarks

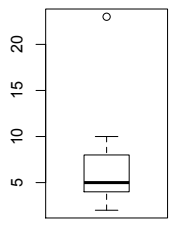
This chapter presented the results of this study. Traditional analysis of *hits*, *misses* and *false alarms* were conducted to compare group reports. For this reason, I presented values of *Validity*, *Thoroughness* and *F-measure* for each report approached. The results were limited by the small sample of reports and websites considered. However, they showed great insights to the literature. Such findings, insights and its impact in field are discussed in the following chapter. Also, the next chapter presents an evaluation of the hypothesis from this study.

Figure 22 – Distribution of number of *Interactivity* problems reported by each level of the factorial design.



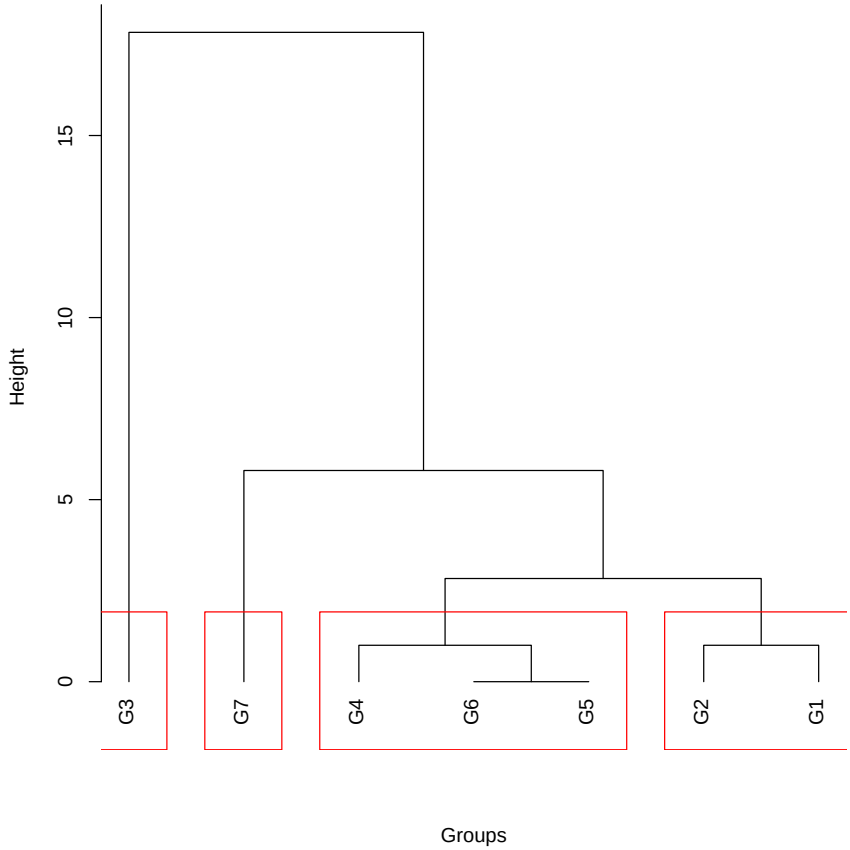
Source: Elaborated by the author.

Figure 23 – Box plot for the number of *Interactivity* problems listed in each group report. One outlier detected, the G3 report (*Min.* = 2.00, *1st Qu.* = 4.00, *Md* = 5.00, *3rd Qu.* = 8.00, *Max.* = 23.00).



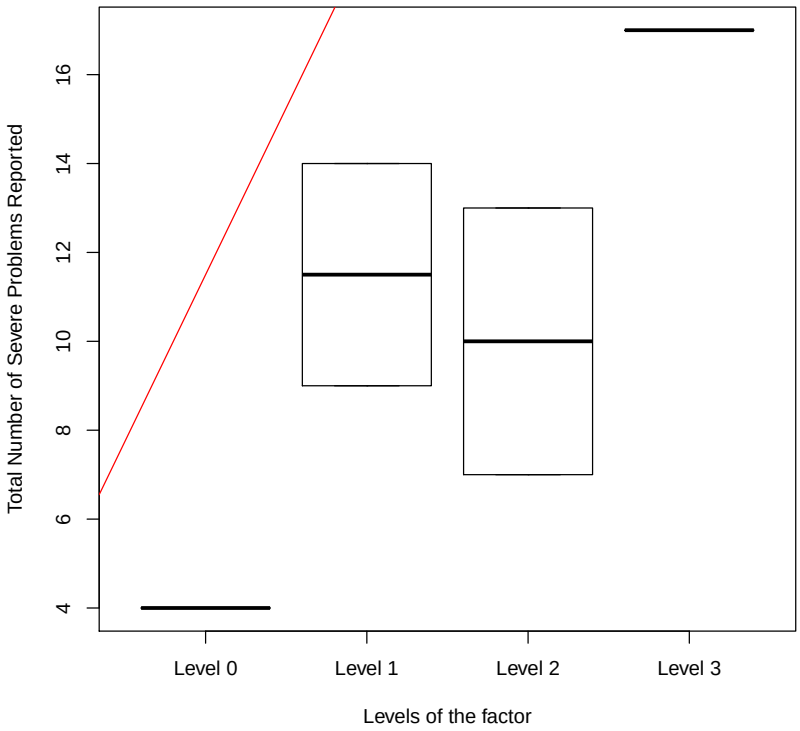
Source: Elaborated by the author.

Figure 24 – UPGMA dendrogram considering similarity of reports in discovery of usability problems that belong to the *Interactivity* category.



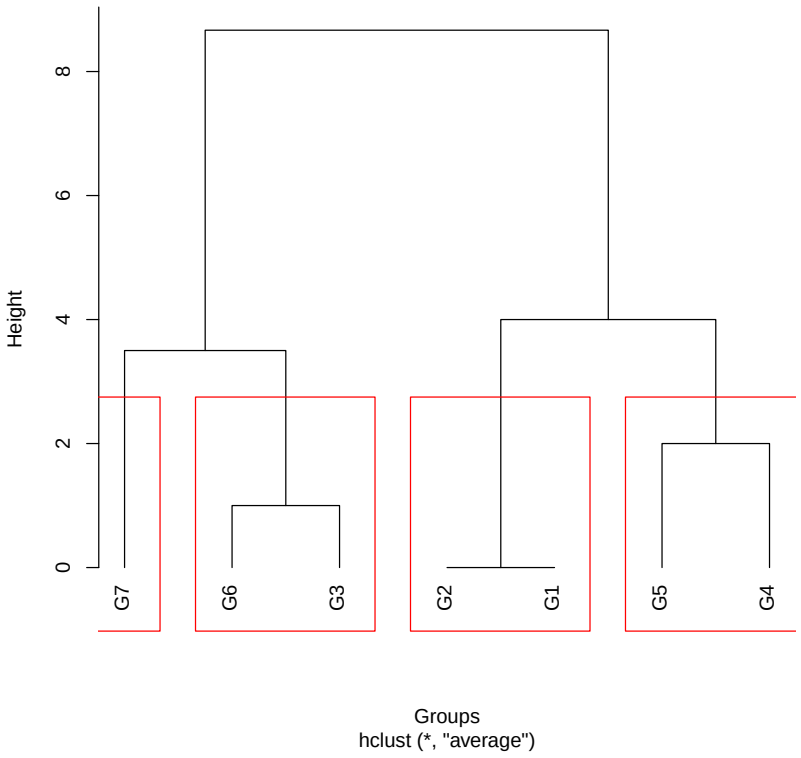
Source: Elaborated by the author.

Figure 25 – Distribution of number of severe problems reported by each level of the factorial design. Linear regression indicated by the red crossing line ($p - value = 0.1386$).



Source: Elaborated by the author.

Figure 26 – UPGMA dendrogram considering similarity of reports in discovery of severe usability problems.



Source: Elaborated by the author.

CHAPTER

7

DISCUSSIONS

7.1 Introduction

The goal of this chapter is to discuss results of this study. For this reason, I conduct an evaluation of its *Hypothesis* and *Research Question* (see section 1.3). In addition, I suggest topics for future studies based on the evidences from this study.

7.2 Evaluation of the Hypothesis

This section presents evaluation for each of the Hypothesis of this study (as presented in subsection 1.3.1). Such evaluations are based on the results showed at the previous chapter. The following subsections are organized by each Hypothesis.

7.2.1 Hypothesis H0

The **Hypothesis H0 (null)** assumed that the quality of the outcomes from a *Mixed Group* was more similar to the quality of the outcomes of a *Baseline Group* than to the quality of outcomes from a *Benchmark Group*, as represented in Figure 1.

The results showed that the F-measure, from both criteria, suggested the acceptance of this hypothesis. All F-measure calculated were lower than 0.5, which may indicate that they are less similar to the *Benchmark* than to the *Baseline*.

Regarding the *cluster analysis* by category of usability problems, all category analysis showed, at least, one case that could accept the Hypothesis H0 (null). Considering the discovery of *Physical Presentation* usability problems, the major part of *Mixed Group* reports were more similar to *Baseline* reports than to *Benchmark* report (see Figure 16). Considering the discovery of *Content* usability problems, the major part of *Mixed Group* reports were also more similar to *Baseline* reports than to *Benchmark* report (G7), as showed in Figure 19. Considering the

discovery of *Information Architecture* usability problems, reports from *Mixed Groups* (despite G4) were more similar to *Baseline* reports than to the *Benchmark* report (G7), as indicated in Figure 21. Considering the discovery of *Interactivity* usability problems, *Mixed Group* reports (despite G3) were more similar to *Baseline* reports than to the *Benchmark* report (G7), as shown in Figure 24.

Nevertheless, the results from *cluster analysis* considering severe problems showed that reports G3 and G6 (from *Mixed Group*) were more similar to the *Benchmark* report (G7) than to other reports (see Figure 26). In addition, this similarity was the strongest similarity indicated in the dendrogram.

Therefore, assuming that the discovery of severe problems is more valuable for practitioners, I conclude that the **Hypothesis H0 - null** can be reject by the analysis of this study.

7.2.2 Hypothesis H1

The **Hypothesis H1** assumed that the quality of the outcomes from a *Mixed Group* was equally similar to the quality of the outcomes of a *Baseline Group* and to the quality of outcomes from a *Benchmark Group*, as represented in Figure 2.

The results showed that the F-measure, from both criteria, suggested the acceptance of this hypothesis. All F-measure calculated were lower than 0.5, which may indicate that the similarity between a *Mixed Group* and a *Baseline Group* is higher than the similarity between a *Mixed Group* and a *Benchmark Group*. This suggest the rejection of Hypothesis H1.

Regarding the *cluster analysis* (see section 6.5), none of the dendrograms indicated that the "similarity between *Mixed Group* reports and *Baseline* reports" and the "similarity between *Mixed Group* reports and the *Benchmark* report" may have the same height. Thus, I conclude that, for the observations of this study, the **Hypothesis H1** can be rejected.

7.2.3 Hypothesis H2

This section discuss the evaluation of **Hypothesis H2**, which assumed that the quality of the outcomes from a *Mixed Group* was more similar to the quality of the outcomes of a *Benchmark Group* than to the quality of outcomes from a *Baseline Group*, as represented in Figure 3.

The results showed that the F-measure, from both criteria, suggested the rejection of this hypothesis. All F-measure calculated were lower than 0.5, which may indicate that they are less similar to the *Benchmark* than to the *Baseline*.

All the *cluster analysis* of this study (both analysis by category and by severity) showed, at least, one case that suggested the acceptance of this hypothesis. Figure 16 showed that G6, one of the *Mixed Group* reports from level 2, was more similar to the discovery of *Physical*

Presentation problems of the *Benchmark Group* report than to the discovery of such category of problems observed among *Baseline Group* reports. In sequence, Figure 19 showed that a *Mixed Group* report from level 1 (G3) was more similar to the discovery of *Content* usability problems of a *Benchmark Group* than to the discovery of *Content* usability problems of a *Baseline Group* report. In this case, G3 was detected as a possible outlier (see Figure 18). Figure 21 showed that report G4, a *Mixed Group* report from level 1, was more similar to the *Benchmark Group* report than to any *Baseline Group* report considering the discovery of *Information Architecture* usability problems. Moreover, Figure 24 showed that report G3, a *Mixed Group* report from level 1, was more similar to the *Benchmark Group* report than to any *Baseline Group* report regarding the discovery of *Interactivity* usability problems. Nevertheless, Figure 23 showed that G3 report was a possible outlier in the discovery of *Interactivity* usability problems. Finally,the results from *cluster analysis* considering severe problems showed that reports G3 and G6 (from *Mixed Group*) were more similar to the *Benchmark* report (G7) than to other reports (see Figure 26). In addition, this similarity was the strongest similarity indicated in the dendrogram.

Therefore, for the observations of this study, I conclude that **Hypothesis H2** can be accepted.

7.3 Evaluation of the Research Question

The **Research Question** of this study was: *"Can a CHE performed by a Mixed Group result in outcomes whose quality can be considered more similar to the quality of outcomes from a traditional HE with multiple expert inspectors (Benchmark Group) than to the quality of outcomes from a CHE conducted only by novice inspectors (Baseline Group)?*" This section presents discussions on the evaluation of such question.

Based on the observations of this study and the discussions of subsection 7.2.3, the **Research Question** can be answered by the acceptation of Hypothesis H2. For this reason, I assume that the quality of the outcomes from a *Mixed Group* was more similar to the quality of the outcomes of a *Benchmark Group* than to the quality of outcomes from a *Baseline Group*.

7.4 Implications for Design

The results of this project showed that a CHE performed by a *Mixed Group* (an expert and multiple evaluators) can be more similar to traditional HE performed only by experts than CHE performed only by novices. Thus, I suggest to practitioners to adopt the structure of *Mixed Group* of evaluators performing CHE when their budget is not sufficient to count on multiple experts performing traditional HE method. Nevertheless, in such cases, practitioners need to ensure that expert evaluators are indeed expert, which is out of the scope of this study.

7.5 Directions for Future Studies

The results from this study raised additional research questions. First, because this study had a limited sample size for the experiment, the first direction for future studies is to replicate our design of study with larger, or complementary, samples of evaluation reports and websites. Thereafter, the results of this study suggest the following topics for future researches on HE and novice evaluators:

Additional *Research Questions*:

1. **Motivation:** the findings of this study suggested that the *Index of Reduction (IR)* (see Table 5) may be correlated to the number of usability problems discovered. In other words, for the observations of this study, groups that reported more problems, also reported more narrow described problems (Pearson Correlation Test, $r \approx 0.79$).
   a) ***Research Question:*** *do HEs that report more usability problems also report more narrow usability problems?*
2. **Motivation:** as suggested in the results (see Figure 10), the union of reports from the level 1 (G3 and G4) had more *hits* than reports from the level 2 (G5 and G6) united. Despite the possibility of this fact be due to the method of assessment of UEMs, this fact may be due to the small sample size of reports and websites. On the other hand, this may indicate that other expertise may have impacted inn these results, which would classify some experts as double experts(NIELSEN, 1992).
   a) ***Research Question:*** *What is the best cost/benefit regarding the number of experts in a Mixed Group in order to influence the development of qualified CHE reports?*
   b) ***Research Question:*** *How to classify usability inspectors in the Brazilian context?*
   c) ***Research Question:*** *Is the classification showed by Botella, Alarcon and Peñalver (2014) appropriated for the Brazilian context?*
   d) ***Research Question:*** *What is the influence of double experts in Mixed Group of CHEs?*
3. ***Motivation:*** as shown at the relaxed criteria (Figure 11), among the observations of this study, the distance from the mean *F-measure* increased as the level of the factor increased.
   a) ***Research Question:*** *Do CHE groups with more experts spend more time discussing details of the inspection during inspection period while groups with less experts focus more on finding and reporting problems?*

4. **Motivation:** the strict analysis showed a significant positive correlation between presence of an expert and the value of *F-measure* (Pearson Correlation Test, $p-value \approx 0.009$).

   a) ***Research Question:*** *Is the value of F-measure positively correlated with the level of the presence of experts in a CHE?*

5. **Motivation:** a linear model may be applicable between the number of *Physical Presentation* problems listed and the level of the factorial design. However, I carried out a linear regression for such data and no significance for such linear model was found ($p-value = 0.2308$).

   a) ***Research Question:*** *Is the discovery of Physical Presentation problems positively correlated with the number of experts in a group of CHE?*

6. **Motivation:** according to the results shown in Figure 20, the reports from level 1 listed more *Information Architecture* problems than report G7 (*Benchmark*). In addition, as showed in the figure, one report from each of the first three level listed the same number of *Information Architecture* problems. This may evidence that the discovery of *Information Architecture* problems is less dependent on the *expertise-effect* than the discovery of other categories of usability problems.

   a) ***Research Question:*** *Is the discovery of Information Architecture problems independent of the expertise effect?*

7. **Motivation:** a linear regression comparing the discovery of *severe* problems and the number of experts in a CHE/HE group showed a non significant linear model ($p-value = 0.1386$).

   a) ***Research Question:*** *Is the amount of severe problems discovered influenced by the number of experts in a Mixed Group?*

As mentioned, these directions are based on the scope of the evidences from this study that could not be tested with this experimental design. Nevertheless, I believe that they are important evidences for the field. Thus, I presented them as suggestions for future studies in the field.

7.6 Final Remarks

This chapter presented evaluation for the hypothesis and research question of this study. Also, it presented a discussion for each evaluation and a discussion based on additional findings of this study. Thus, I gave suggestions of topics to be investigated and explored by future studies. Thereafter, the following chapter presents the conclusions of this study.

CHAPTER

8

CONCLUSIONS

This study aimed to determine whether a CHE conducted by expert and novice inspectors together result in qualified outcomes in comparison to standard HEs. Also, the goal of this study can be understood as answering the following *research question*:

> *"Can a CHE performed by a Mixed Group result in outcomes whose quality can be considered more similar to the quality of outcomes from a traditional HE with multiple expert inspectors (Benchmark Group) than to the quality of outcomes from a CHE conducted only by novice inspectors (Baseline Group)?"*

Based on the observations of this study, on the preferences for evaluation of hypothesis (see section 5.7) and the discussions of subsection 7.2.3, the *research question* can be answered by the acceptation of Hypothesis H2. Therefore, I conclude that the quality of the outcomes from a *Mixed Group* was more similar to the quality of the outcomes of a *Benchmark Group* than to the quality of outcomes from a *Baseline Group*.

The findings of this study showed new insights for the literature about HE for novice evaluators. Also, it is in accordance with the suggestion of Buykx (2009) of employing CHE as a training for novice evaluators. It also reinforce the suggestion of Wodike, Sim and Horton (2014) about having a facilitator among novice inspectors in order to enhance the quality of CHE reports. I suggest to practitioners that, in situations when the organization cannot count on multiple experts for a HE, they should consider the adoption of a *Mixed Group* of evaluators performing a CHE.

The main limitation referred to the sample size of reports, inspectors and websites considered. However, it does not reject the Hypothesis H2 and the main conclusion of this study. Thus, future studies can expand our procedures to larger or complementary samples of each

population referred (reports, inspectors and websites).

In direction of developing a well-adapted HE for novice evaluators, validating classifications for evaluators' expertise remains necessary. Thus, studies that explore in deeper the differences among evaluators of distinct expertise are a requirement. In addition, this study resulted in a list of ideas for future studies, as showed at section 7.5.

BIBLIOGRAPHY

ALJOHANI, M.; BLUSTEIN, J. Heuristic evaluation of university institutional repositories based on dspace. In: MARCUS, A. (Ed.). **Design, User Experience, and Usability: Interactive Experience Design: 4th International Conference, DUXU 2015, Held as Part of HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015, Proceedings, Part III**. [S.l.]: Springer International Publishing, 2015. (Lecture Notes in Computer Science, v. 9188), p. 119–130. ISBN 978-3-319-20889-3. Citation on page 38.

ANDRE, T. S.; HARTSON, H. R.; BELZ, S. M.; MCCREARY, F. A. The user action framework: a reliable foundation for usability engineering support tools. **International Journal of Human-Computer Studies**, v. 54, n. 1, p. 107 – 136, 2001. ISSN 1071-5819. Available: <http://www.sciencedirect.com/science/article/pii/S1071581900904415>. Citation on page 49.

BABAJO, A.; PETRIE, H. **The effectiveness of Collaborative Heuristic Evaluation**. 2012. Unpublished Msc Thesis, The University of York. Citations on pages 50 e 57.

BARUA, A.; MANI, D.; MUKHERJEE, R. Measuring the business impacts of effective data. **Report accessed at http://www. sybase. com/files/White_Papers on Sep**, v. 15, p. 17, 2012. Citation on page 23.

BEVAN, N.; CARTER, J.; EARTHY, J.; GEIS, T.; HARKER, S. New ISO Standards for Usability, Usability Reports and Usability Measures. In: KUROSU, M. (Ed.). **Human-Computer Interaction. Theory, Design, Development and Practice**. Springer International Publishing, 2016, (Lecture Notes in Computer Science, 9731). p. 268–278. ISBN 978-3-319-39509-8 978-3-319-39510-4. DOI: 10.1007/978-3-319-39510-4_25. Available: <http://link.springer.com/chapter/10.1007/978-3-319-39510-4_25>. Citation on page 32.

BORSCI, S.; MACREDIE, R. D.; BARNETT, J.; MARTIN, J.; KULJIS, J.; YOUNG, T. Reviewing and Extending the Five-User Assumption. **ACM Transactions on Computer-Human Interaction**, v. 20, n. 5, p. 1–23, 2013. ISSN 10730516. Citations on pages 32 e 37.

BORYS, M.; LASKOWSKI, M. Expert vs Novice Evaluators - Comparison of Heuristic Evaluation Assessment. **Proceedings of 16th International Conference on Enterprise Information Systems**, p. 144–149, 2014. Citation on page 38.

BOTELLA, F.; ALARCON, E.; PEÑALVER, A. How to Classify to Experts in Usability Evaluation. In: **Proceedings of the XV International Conference on Human Computer Interaction**. New York, NY, USA: ACM, 2014. (Interaccion'14), p. 25:1–25:4. ISBN 978-1-4503-2880-7. Citations on pages 42 e 88.

BRAJNIK, G.; YESILADA, Y.; HARPER, S. The expertise effect on web accessibility evaluation methods. **Human–Computer Interaction**, v. 26, n. 3, p. 246–283, 2011. Available: <http://www.tandfonline.com/doi/abs/10.1080/07370024.2011.601670>. Citations on pages 24 e 38.

BRUUN, A.; STAGE, J. Barefoot usability evaluations. **Behaviour & Information Technology**, n. February 2015, p. 1–20, 2014. ISSN 0144-929X. Citations on pages 24 e 38.

____. New approaches to usability evaluation in software development: Barefoot and crowdsourcing. **Journal of Systems and Software**, v. 105, p. 40–53, jul 2015. Citations on pages 24 e 38.

BUYKX, L. **Improving heuristic evaluation through collaborative working**. Master's Thesis (Master's Thesis) — The University of York Department of Computer Science, sep 2009. Citations on pages 24, 38, 44, 50, 57 e 91.

COCKTON, G.; LAVERY, D.; WOOLRYCH, A. Inspection based evaluations. In: SEARS, A.; JACKO, J. A. (Ed.). **Human-Computer Interaction: Development process**. [S.l.: s.n.], 2009. p. 273 –276. ISBN 1420088904. Citations on pages 24 e 38.

COCKTON, G.; WOOLRYCH, A. Understanding inspection methods: Lessons from an assessment of heuristic evaluation. In: BLANDFORD, A.; VANDERDONCKT, J.; GRAY, P. (Ed.). **People and Computers XV—Interaction without Frontiers: Joint Proceedings of HCI 2001 and IHM 2001**. London: Springer London, 2001. p. 171–191. ISBN 978-1-85233-515-1. Citation on page 24.

DEMERS, R. A. System design for usability. **Commun. ACM**, ACM, New York, NY, USA, v. 24, n. 8, p. 494–501, Aug. 1981. ISSN 0001-0782. Available: <http://doi.acm.org/10.1145/358722.358730>. Citation on page 29.

DIX, A.; FINLAY, J.; ABOWD, G. D.; BEALE, R. **Human Computer Interaction**. 3rd. ed. [S.l.]: Pearson Education Limited, 2003. Citations on pages 32 e 33.

ERICSSON, K. A.; SIMON, H. A. Verbal reports as data. **Psychological Review**, American Psychological Association, US, v. 87, n. 3, p. 215–251, 1980. Citation on page 33.

FERNANDEZ, A.; ABRAHÃO, S.; INSFRAN, E. A systematic review on the effectiveness of web usability evaluation methods. **IET Conference Proceedings**, Institution of Engineering and Technology, p. 52–56(4), January 2012. Available: <http://digital-library.theiet.org/content/conferences/10.1049/ic.2012.0007>. Citation on page 47.

FERNANDEZ, A.; INSFRAN, E.; ABRAHÃO, S. Usability evaluation methods for the web: A systematic mapping study. **Information and Software Technology**, Elsevier, v. 53, n. 8, p. 789–817, 2011. Citations on pages 23, 24 e 32.

FERREIRA, D. F. **Estatística multivariada**. [S.l.]: Editora UFLA, 2008. Citations on pages 57, 58 e 76.

FØLSTAD, A.; LAW, E.; HORNBÆK, K. Analysis in practical usability evaluation: A survey study. In: **Proceedings of the SIGCHI Conference on Human Factors in Computing Systems**. New York, NY, USA: ACM, 2012. (CHI '12), p. 2127–2136. ISBN 978-1-4503-1015-4. Citations on pages 24 e 36.

FORTES, R. P.; ANTONELLI, H. L.; SALGADO, A. de L. Accessibility and usability evaluation of rich internet applications. In: **Proceedings of the 22Nd Brazilian Symposium on Multimedia and the Web**. New York, NY, USA: ACM, 2016. (Webmedia '16), p. 7–8. ISBN 978-1-4503-4512-5. Available: <http://doi.acm.org/10.1145/2976796.2988221>. Citation on page 27.

_____. Avaliação de acessibilidade e usabilidade em ria. In: **Anais do XXII Simpósio Brasileiro de Sistemas Multimídia e Web: Minicursos, 8 a 11 de novembro, 2016, Teresina, Piauí.** [S.l.]: Serviço de Processamento Técnico – IFPI Biblioteca Dr. Francisco Montojos, 2016. v. 3, p. 37–65. Citation on page 27.

FORTES, R. P. de M.; SALGADO, A. de L.; SANTOS, F. de S.; AMARAL, L. A. do; SILVA, E. A. N. da. Game accessibility evaluation methods: a literature survey. In: **To appear in the Proceedings of HCI International 2017, Vancouver, Canada, July 09–14, 2017**. [S.l.]: Springer International Publishing, 2017. Citation on page 27.

GEORGSSON, M.; WEIR, C.; STAGGERS, N. Revisiting Heuristic Evaluation Methods to Improve the Reliability of Findings. **2014 European Federation for Medical Informatics and IOS Press**, p. 930–934, 2014. ISSN 09269630. Citation on page 24.

GOULD, J. D.; LEWIS, C. Designing for usability: Key principles and what designers think. **Commun. ACM**, ACM, New York, NY, USA, v. 28, n. 3, p. 300–311, Mar. 1985. ISSN 0001-0782. Available: <http://doi.acm.org/10.1145/3166.3170>. Citation on page 30.

GUERRIERO, I. C. Z. Approval of the resolution governing the ethics of research in social sciences, the humanities, and other disciplines that use methodologies characteristic of these areas: challenges and achievements. **Ciência & Saúde Coletiva**, SciELO Public Health, v. 21, n. 8, p. 2619–2629, 2016. Citation on page 52.

HAIR, J. F.; ANDERSON, R. E.; BABIN, B. J.; BLACK, W. C. **Multivariate data analysis: A global perspective**. [S.l.]: Pearson Upper Saddle River, NJ, 2010. Citations on pages 57 e 58.

HARTSON, H. R.; ANDRE, T. S.; WILLIGES, R. C. Criteria for evaluating usability evaluation methods. **International journal of human-computer interaction**, Taylor & Francis, v. 13, n. 4, p. 373–410, 2001. Citations on pages 47, 48, 49 e 50.

HERTZUM, M.; JACOBSEN, N. E. The Evaluator Effect: A Chilling Fact About Usability Evaluation Methods. **International Journal of Human-Computer Interaction**, v. 14, n. 4, p. 421–443, 2001. ISSN 1044-7318. Citations on pages 24 e 38.

HORNBÆK, K. Dogmas in the assessment of usability evaluation methods. **Behaviour & Information Technology**, Taylor & Francis, v. 29, n. 1, p. 97–111, 01 2010. Citations on pages 23, 47 e 58.

HORNBÆK, K.; FRØKJÆR, E. Comparison of techniques for matching of usability problem descriptions. **Interacting with Computers**, v. 20, n. 6, p. 505–514, Dec. 2008. ISSN 0953-5438, 1873-7951. Available: <http://iwc.oxfordjournals.org.ez67.periodicos.capes.gov.br/content/20/6/505>. Citations on pages 49 e 50.

HORNBÆK, K.; STAGE, J. The interplay between usability evaluation and user interaction design. **International Journal of Human-Computer Interaction**, Taylor & Francis, v. 21, n. 2, p. 117–123, 11 2006. Citation on page 32.

HUANG, B. **A Comparison of Remote Collaborative Heuristic Evaluation by Novices and Experts with User-based Evaluation**. Master's Thesis (Master's Thesis) — University of York, 2012. Citations on pages 24, 45, 50, 57 e 61.

ISO 9241-210. **Ergonomics of human-system interaction ¯ Part 210: Human-centred design for interactive systems**. 2010. Available: <https://www.iso.org/obp/ui/#iso:std:iso:9241:-210:ed-1:v1:en>. Citations on pages 29 e 30.

ISO/IEC 25066. **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**. 2016. Citations on pages 23, 29, 30, 32, 34 e 35.

ISO/TR 9241-100. **Ergonomics of human-system interaction — Part 100: Introduction to standards related to software ergonomics**. 2010. Available: <https://www.iso.org/obp/ui/#iso:std:iso:tr:9241:-100:ed-1:v1:en>. Citation on page 23.

JADHAV, D.; BHUTKAR, G.; MEHTA, V. Usability evaluation of messenger applications for android phones using cognitive walkthrough. In: **Proceedings of the 11th Asia Pacific Conference on Computer Human Interaction**. New York, NY, USA: ACM, 2013. (APCHI '13), p. 9–18. ISBN 978-1-4503-2253-9. Available: <http://doi.acm.org/10.1145/2525194.2525202>. Citation on page 34.

JOHANNESSEN, G. H. J.; HORNBÆK, K. Must evaluation methods be about usability? Devising and assessing the utility inspection method. **Behaviour & Information Technology**, v. 33, n. 2, p. 195–206, 2014. ISSN 0144-929X. Citations on pages 24 e 38.

KATZEFF, C.; NYBLOM, ; TUNHEDEN, S.; TORSTENSSON, C. User-centred design and evaluation of EnergyCoach – an interactive energy service for households. **Behaviour & Information Technology**, v. 31, n. 3, p. 305–324, Mar. 2012. ISSN 0144-929X. Available: <http://dx.doi.org/10.1080/0144929X.2011.618778>. Citation on page 30.

KORHONEN, H.; KOIVISTO, E. M. I. Playability heuristics for mobile games. In: **Proceedings of the 8th Conference on Human-computer Interaction with Mobile Devices and Services**. New York, NY, USA: ACM, 2006. (MobileHCI '06), p. 9–16. ISBN 1-59593-390-5. Available: <http://doi.acm.org/10.1145/1152215.1152218>. Citation on page 43.

KOUTSABASIS, P.; SPYROU, T.; DARZENTAS, J. S.; DARZENTAS, J. On the performance of novice evaluators in usability evaluations. In: CITESEER. **11th Panhellenic Conference on Informatics (PCI 2007) Patras, Greece**. [S.l.], 2007. p. 18–20. Citation on page 38.

KRUMM, J. **Ubiquitous computing fundamentals**. [S.l.]: CRC Press, 2016. Citation on page 33.

LAVERY, D.; COCKTON, G.; ATKINSON, M. P. Comparison of evaluation methods using structured usability problem reports. **Behaviour & Information Technology**, Taylor & Francis Ltd, v. 16, n. 4, p. 246–266, 1997. Citation on page 49.

LEWIS, C.; POLSON, P. G.; WHARTON, C.; RIEMAN, J. Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces. In: **Proceedings of the SIGCHI Conference on Human Factors in Computing Systems**. New York, NY, USA: ACM, 1990. (CHI '90), p. 235–242. ISBN 0-201-50932-6. Available: <http://doi.acm.org/10.1145/97243.97279>. Citation on page 34.

LOWRY, P. B.; ROBERTS, T. L.; ROMANO, N. C. What signal is your inspection team sending to each other? Using a shared collaborative interface to improve shared cognition and implicit coordination in error-detection teams. **International Journal of Human Computer Studies**, Elsevier, v. 71, n. 4, p. 455–474, 2013. ISSN 10715819. Citation on page 38.

MACDONALD, C. M.; ATWOOD, M. E. Changing perspectives on evaluation in hci: Past, present, and future. In: **CHI '13 Extended Abstracts on Human Factors in Computing Systems**. New York, NY, USA: ACM, 2013. (CHI EA '13), p. 1969–1978. ISBN 978-1-4503-1952-2. Citation on page 41.

MACFARLANE, S.; PASIALI, A. Adapting the heuristic evaluation method for use with children. In: **Workshop on child computer interaction: methodological research, Interact**. [S.l.: s.n.], 2005. p. 28–31. Citations on pages 24, 43 e 44.

MACFARLANE, S.; SIM, G.; HORTON, M. Assessing usability and fun in educational software. In: ACM. **Proceedings of the 2005 conference on Interaction design and children**. [S.l.], 2005. p. 103–109. Citation on page 24.

MAHATODY, T.; SAGAR, M.; KOLSKI, C. State of the art on the cognitive walkthrough method, its variants and evolutions. **International Journal of Human-Computer Interaction**, v. 26, n. 8, p. 741–785, 2010. Available: <http://dx.doi.org/10.1080/10447311003781409>. Citations on pages 34 e 35.

MANNING, C. D.; SCHÜTZE, H. *et al.* **Foundations of statistical natural language processing**. [S.l.]: MIT Press, 1999. Citation on page 49.

MARTINS, A. I.; QUEIRÓS, A.; SILVA, A. G.; ROCHA, N. P. Usability Evaluation Methods: A Systematic Review. **Human Factors in Software Development and Design**, IGI Global, p. 250, 2014. Citations on pages 24 e 36.

NIELSEN, J. Finding usability problems through heuristic evaluation. In: ACM. **Proceedings of the SIGCHI conference on Human factors in computing systems**. [S.l.], 1992. p. 373–380. Citations on pages 24, 36, 37, 38, 43, 55 e 88.

____. Enhancing the Explanatory Power of Usability Heuristics. In: **Proceedings of the SIGCHI Conference on Human Factors in Computing Systems**. New York, NY, USA: ACM, 1994. (CHI '94), p. 152–158. ISBN 0-89791-650-6. Citations on pages 36 e 57.

NIELSEN, J. Heuristic evaluation. In: NIELSEN, J.; MACK, R. L. (Ed.). **Usability inspection methods**. [S.l.: s.n.], 1994. p. 25 –62. Citations on pages 36, 37, 42, 53 e 58.

NIELSEN, J. **How to Conduct a Heuristic Evaluation**. 1995. Available: <https://www.nngroup.com/articles/how-to-conduct-a-heuristic-evaluation/>. Citations on pages 36 e 38.

____. **Becoming a Usability Professional**. 2002. Available: <https://www.nngroup.com/articles/becoming-a-usability-professional/>. Citations on pages 41 e 42.

____. **Usability 101: Introduction to Usability**. 2012. Available: <https://www.nngroup.com/articles/usability-101-introduction-to-usability/>. Citations on pages 23 e 29.

NIELSEN, J.; MOLICH, R. Heuristic evaluation of user interfaces. In: ACM. **Proceedings of the SIGCHI conference on Human factors in computing systems**. [S.l.], 1990. p. 249–256. Citation on page 36.

NORMAN, D. A. **The design of everyday things: Revised and expanded edition**. [S.l.]: Basic books, 2013. ISBN 0465072992. Citations on pages 9, 30 e 49.

OTHMAN, M. K.; MAHUDIN, F.; AHAGUK, C. H.; Abdul Rahman, M. F. Mobile guide technologies (smartphone apps): Collaborative Heuristic Evaluation (CHE) with expert and novice users. **User Science and Engineering (i-USEr), 2014 3rd International Conference on**, p. 232–236, 2014. Citations on pages 45, 65, 70 e 71.

PAZ, F.; PAZ, F. A.; POW-SANG, J. A. Experimental case study of new usability heuristics. In: **Design, User Experience, and Usability: Design Discourse**. [S.l.]: Springer, 2015. p. 212–223. Citation on page 38.

PAZ, F.; POW-SANG, J. A. A systematic mapping review of usability evaluation methods for software development process. **International Journal of Software Engineering and Its Applications**, v. 10, n. 1, p. 165–178, 2016. Available: <http://www.sersc.org/journals/IJSEIA/vol10_no1_2016/16.pdf>. Citation on page 36.

PETRIE, H.; BUYKX, L. Collaborative Heuristic Evaluation: improving the effectiveness of heuristic evaluation. In: **Proceedings of UPA 2010 International Conference. Omnipress**. [S.l.: s.n.], 2010. Citations on pages 24, 38, 44, 50 e 57.

PETRIE, H.; POWER, C. What do users really care about?: a comparison of usability problems found by users and experts on highly interactive websites. In: ACM. **Proceedings of the SIGCHI Conference on Human Factors in Computing Systems**. [S.l.], 2012. p. 2107–2116. Citations on pages 17, 24, 33, 36, 50, 56, 57, 58 e 73.

POLSON, P. G.; LEWIS, C.; RIEMAN, J.; WHARTON, C. Cognitive walkthroughs: a method for theory-based evaluation of user interfaces. **International Journal of Man-Machine Studies**, v. 36, n. 5, p. 741 – 773, 1992. ISSN 0020-7373. Available: <http://www.sciencedirect.com/science/article/pii/002073739290039N>. Citation on page 34.

PREECE, J.; SHARP, H.; ROGERS, Y. **Interaction design: beyond human-computer interaction**. 4. ed. [S.l.]: John Wiley & Sons, 2015. Citations on pages 24, 30, 32, 33, 34, 36, 37 e 55.

READ, J. Children As Participants in Design and Evaluation. **interactions**, ACM, New York, NY, USA, v. 22, n. 2, p. 64–66, feb 2015. ISSN 1072-5520. Citation on page 24.

RENZI, A. B.; CHAMMAS, A.; AGNER, L.; GREENSHPAN, J. Startup rio: User experience and startups. In: **Design, User Experience, and Usability: Design Discourse**. [S.l.]: Springer, 2015. p. 339–347. Citation on page 24.

SALGADO, A. d. L.; FORTES, R. P. d. M.; LARA, S. M. A. d.; FREIRE, A. P. *et al.* What is hidden in a heuristic evaluation: tactics from the experts. In: FACULDADE DE ECONOMIA, ADMINISTRAÇÃO E CONTABILIDADE-FEA/USP. **International Conference on Information Systems and Technology Management, XIII**. 2016. p. 2931–2946. Available: <http://www.contecsi.fea.usp.br/envio/index.php/contecsi/13CONTECSI/paper/view/4068/2622>. Citations on pages 26 e 44.

SALGADO, A. de L.; AMARAL, L. A.; FREIRE, A. P.; FORTES, R. P. M. Usability and ux practices in small enterprises: Lessons from a survey of the brazilian context. In: **Proceedings of the 34th ACM International Conference on the Design of Communication**. New York, NY, USA: ACM, 2016. (SIGDOC '16), p. 18:1–18:9. ISBN 978-1-4503-4495-1. Available: <http://doi.acm.org/10.1145/2987592.2987616>. Citation on page 26.

SALGADO, A. de L.; AMARAL, L. A. do; CASTRO, P. C.; FORTES, R. P. M. Designing for parental control: Enriching usability and accessibility in the context of smart toys. In: TANG, J. K.; HUNG, P. C. K. (Ed.). **Computing in Smart Toys**. [S.l.]: Springer International Publishing, 2017. p. 170. ISBN 978-3-319-62071-8. Citation on page 27.

SALGADO, A. de L.; AMARAL, L. A. do; FORTES, R. P. de M.; CHAGAS, M. H. N.; JOYCE, G. Addressing mobile usability and elderly users: validating contextualized heuristics. In: **To appear in the Proceedings of HCI International 2017, Vancouver, Canada, July 09–14, 2017**. [S.l.]: Springer International Publishing, 2017. Citation on page 27.

SALGADO, A. de L.; FORTES, R. P. de M. Heuristic evaluation for novice evaluators. In: ____. **Design, User Experience, and Usability: Design Thinking and Methods: 5th International Conference, DUXU 2016, Held as Part of HCI International 2016, Toronto, Canada, July 17–22, 2016, Proceedings, Part I**. Cham: Springer International Publishing, 2016. p. 387–398. ISBN 978-3-319-40409-7. Citations on pages 26 e 44.

SALGADO, A. de L.; RODRIGUES, S. S.; FORTES, R. P. M. Evolving heuristic evaluation for multiple contexts and audiences: Perspectives from a mapping study. In: **Proceedings of the 34th ACM International Conference on the Design of Communication**. New York, NY, USA: ACM, 2016. (SIGDOC '16), p. 19:1–19:8. ISBN 978-1-4503-4495-1. Available: <http://doi.acm.org/10.1145/2987592.2987617>. Citation on page 27.

SALIAN, K.; SIM, G. Simplifying Heuristic Evaluation for Older Children. In: **Proceedings of the India HCI 2014 Conference on Human Computer Interaction**. New York, NY, USA: ACM, 2014. (IndiaHCI '14), p. 26:26—-26:34. ISBN 978-1-4503-3218-7. Citations on pages 24 e 43.

SALIAN, K.; SIM, G.; READ, J. C. Can Children Perform a Heuristic Evaluation? In: **Proceedings of the 11th Asia Pacific Conference on Computer Human Interaction**. New York, NY, USA: ACM, 2013. (APCHI '13), p. 137–141. ISBN 978-1-4503-2253-9. Citations on pages 24 e 43.

SAURO, J. The relationship between problem frequency and problem severity in usability evaluations. **J. Usability Studies**, Usability Professionals' Association, Bloomingdale, IL, v. 10, n. 1, p. 17–25, Nov. 2014. ISSN 1931-3357. Available: <http://dl.acm.org/citation.cfm?id=2817310.2817312>. Citation on page 56.

SCHELLER, T.; KÜHN, E. Automated measurement of {API} usability: The {API} concepts framework. **Information and Software Technology**, v. 61, p. 145 – 162, 2015. ISSN 0950-5849. Citation on page 38.

SHADISH, W. R.; COOK, T. D.; CAMPBELL, D. T. **Experimental and quasi-experimental designs for generalized causal inference**. [S.l.]: Wadsworth Cengage learning, 2002. Citations on pages 23 e 52.

SLAVKOVIC, A.; CROSS, K. Novice Heuristic Evaluations of a Complex Interface. In: **CHI '99 Extended Abstracts on Human Factors in Computing Systems**. New York, NY, USA: ACM, 1999. (CHI EA '99), p. 304–305. ISBN 1-58113-158-5. Citations on pages 38 e 43.

WATZMAN, S.; RE, M. Visual design principles for usable interfaces - everything is designed: Why we should think before doing. In: SEARS, A.; JACKO, J. A. (Ed.). **Human-Computer Interaction: Development process**. [S.l.: s.n.], 2009. p. 329 – 353. ISBN 1420088904. Citation on page 36.

WAZLAWICK, R. **Metodologia de pesquisa para ciência da computação, 2a edição**. [S.l.]: Elsevier Brasil, 2014. Citation on page 23.

WHARTON, C.; RIEMAN, J.; LEWIS, C.; POLSON, P. Usability inspection methods. In: NIELSEN, J.; MACK, R. L. (Ed.). New York, NY, USA: John Wiley & Sons, Inc., 1994. chap. The Cognitive Walkthrough Method: A Practitioner's Guide, p. 105–140. ISBN 0-471-01877-5. Available: <http://dl.acm.org/citation.cfm?id=189200.189214>. Citation on page 35.

WODIKE, O. A.; SIM, G.; HORTON, M. Empowering Teenagers to Perform a Heuristic Evaluation of a Game. In: BCS. **Proceedings of the 28th International BCS Human Computer Interaction Conference on HCI 2014-Sand, Sea and Sky-Holiday HCI**. [S.l.], 2014. p. 353–358. Citations on pages 24, 43 e 91.

Table of Usability Problems by each Report

Report	Usability Problems	Relaxed Criteria	Strict Criteria	CATEGORY (Petrie and Power, 2012)	Severity	url	Heuristic(s)
G1	Home page menu has a high difficulty level of interaction	1	1a	INTERACTIVITY	0	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	2
G1	The menu colors make it difficult to know which tab is selected	1	1	PHYSICAL PRESENTATION	0	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	1,5,9
G1	Problem of data visibility, the menu is very difficult for the user to access.	1	1a	PHYSICAL PRESENTATION	0	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/servicos.html	1
G1	Main menu has information focus problems	1	1b	PHYSICAL PRESENTATION	0	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	1, 6
G1	Bottom menu has many options and no spatial organization, which could help users to filter the contents	4	4a	INFORMATION ARCHITECTURE	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	1,8
G1	Do not prevent the error (go to another page)	43	43i	INTERACTIVITY	1	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	H8

G1	"Suporte" feature forwards users to another system without warning	43	43h	INTERACTIVITY	1	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index60a0.html?option=com_content&view=category&layout=faq&id=285&Itemid=529	9
G2	After users open the pdf file, there is no option to go back to the previous page but the browser`s option	43	43g	INTERACTIVITY	1	http://portalsaude.saude.gov.br/images/pdf/2014/marco/21/cartilha-integra-direitos-2006.pdf	9
G2	The system does not inform that the link directs users to download a pdf file	43	43f	INTERACTIVITY	1	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	3
G2	The system does not inform that the link directs users to download a file	43	43e	INTERACTIVITY	1	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/o-ministerio/principal/secretarias/264-sgep-raiz/cartao-nacional-de-saude/l2-cartao-nacional-de-saude/8760-orientacoes-para-cadastramento.html	2, 5
G2	The link forwards the user to an external website.	43	43d	INTERACTIVITY	1	http://www.aids.gov.br/	2

G2	The link forwards the user to an external website.	43	43c	INTERACTIVITY	1	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	4
G2	The link does not follow link standards	2	2c	INTERACTIVITY	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/o-ministerio/principal/secretarias/264-sgep-raiz/cartao-nacional-de-saude/l2-cartao-nacional-de-saude/8760-orientacoes-para-cadastramento.html	4
G2	The link has the same color of the text	2	2c	INTERACTIVITY	1	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/o-ministerio/principal/secretarias/264-sgep-raiz/cartao-nacional-de-saude/l2-cartao-nacional-de-saude/8760-orientacoes-para-cadastramento.html	4

G2	The link does not look like a link	2	2c	INTERACTIVITY	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/o-ministerio/principal/secretarias/264-sgep-raiz/cartao-nacional-de-saude/l2-cartao-nacional-de-saude/8760-orientacoes-para-cadastramento.html	4
G3	The link is not easily identifiable	2	2d	INTERACTIVITY	1	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/cidadao/entenda-o-sus.html	4
G3	Inappropriate terminology in the main menu	4	4b	CONTENT	1	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	4
G3	The page has an excessive amount of information	3	3a	CONTENT	1	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/cidadao/entenda-o-sus.html	8

G3	Excessive links on the page	3	3a	CONTENT	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/cidadao/entenda-o-sus.html	4,8
G3	The page has an excessive amount of information	3	3b	CONTENT	1	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index5309.html?option=com_content&view=article&id=8757&Itemid=426	8
G3	The page has unnecessary information, polluting the interface.	3	3b	CONTENT	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index5309.html?option=com_content&view=article&id=8757&Itemid=426	8
G3	Excessive links on the page	3	3b	CONTENT	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index5309.html?option=com_content&view=article&id=8757&Itemid=426	4,8

G3	Excessive links on the page	3	3e	CONTENT	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/o-ministerio/principal/secretarias/sgep/doges-departamento-de-ouvidoria-geral-do-sus/ouvidoria-g-sus.html	4,8
G3	Excessive information on the page	3	3f	CONTENT	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	6,8
G3	The PDF has a lot of irrelevant information, which can confuse the user.	3	3h	CONTENT	1	http://portalsaude.saude.gov.br/images/pdf/2014/setembro/24/NOTA-TECNICA-NOME-SOCIAL-18-2014.pdf	8
G3	There is no standard external links (other pages)	43	43b	INTERACTIVITY	1	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index5309.html?option=com_content&view=article&id=8757&Itemid=426	4
G3	Opens a new tab, no undo but to use the browser features	43	43a	INTERACTIVITY	1	http://datasus.saude.gov.br/sistemas-e-aplicativos/suporte-tecnico/capacitacao-service-desk	3

G3	Excessive information on the page	3	3i	CONTENT	1	http://portalsaude.saude.gov.br/index.php/component/search/?searchword=direitos&searchphrase=all&Itemid=242	8
G3	Many irrelevant information on the page	3	3g	CONTENT	1	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index60a0.html?option=com_content&view=category&layout=faq&id=285&Itemid=529	8
G3	Home page has excessive information	3	3f	CONTENT	1	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	8
G3	At the Services page, there is a high amount of links, making it difficult to users to interact	3	3	CONTENT	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	7
G3	Ambiguity about where to find the required information, different paths with the same suggestion	3	3c	INFORMATION ARCHITECTURE	1	https://www.planalto.gov.br/ccivil_03/Constituicao/Constitui%C3%A7ao_Compilado.htm	4, 2
G3	The information is spread along the website, it is difficult to find the desired information	3	3d	INFORMATION ARCHITECTURE	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/cidadao/entenda-o-sus.html	4,5,6

G3	Too many information on the page	3	3b	CONTENT	1	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index5309.html?option=com_content&view=article&id=8757&Itemid=426	8
G3	Small font size	37	37a	PHYSICAL PRESENTATION	1	https://portaldocidadao.saude.gov.br/portalcidadao/primeiroAcesso.htm	8
G3	No breadcrumb (or similar) to help users to undo their actions	36	36g	INTERACTIVITY	1	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index60a0.html?option=com_content&view=category&layout=faq&id=285&Itemid=529	4
G3	The link "ministry" suggest that the video required could be there	4	4b	INFORMATION ARCHITECTURE	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/o-ministerio.html	5
G3	The breadcrumb does not inform the right path, according to the path performed by the user	36	36f	INTERACTIVITY	1	http://portalsaude.saude.gov.br/index.php/cidadao/principal/agencia-saude?filter-search=&filter-chapeu=GRIPE&filter-mesdate=04&filter-anodate=2016&buscar=&3c16febe237012b4ecbc756125eec11e=1	4

G3	The form require too much information	6	6	INTERACTIVITY	1	http://ouvprod01.saude.gov.br/ouvidor/CadastroDemandaPortal.do;jsessionid=52F6799AB8C7638A292347633A727108.server-ouvidorsus-srvjpdf40	5,8
G3	The procedures to register a SUS user is not clear	7	7	CONTENT	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/o-ministerio/principal/secretarias/264-sgep-raiz/cartao-nacional-de-saude/l2-cartaonacional-de-saude/8760-orientacoes-para-cadastramento.html	1,2,4
G3	No breadcrumb (or similar) to help users to undo their actions	36	36e	INTERACTIVITY	1	http://portalsaude.saude.gov.br/index.php/component/search/?searchword=direitos&searchphrase=all&Itemid=242	4
G3	It is difficult for the user to find the information about SUS user rights	8	8	INFORMATION ARCHITECTURE	1	http://portalsaude.saude.gov.br/images/pdf/2014/marco/21/cartilha-integra-direitos-2006.pdf	1

G3	The link ``national health card' should be displayed in the home page'	8	8a	INFORMATION ARCHITECTURE	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/servicos.html	7
G3	The link ``national health card' should be displayed among news content in the home page'	8	8a	INFORMATION ARCHITECTURE	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index5309.html?option=com_content&view=article&id=8757&Itemid=426	6
G3	Breadcrumb does not inform the right path, according to the path performed by the user	36	36d	INTERACTIVITY	1	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index5309.html?option=com_content&view=article&id=8757&Itemid=426	4
G3	The U.K. flag does not translate the whole website	33	33a	CONTENT	1	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	2
G3	Uses a different link standard (bold text as a clickable link)	2	2g	INTERACTIVITY	1	http://portalsaude.saude.gov.br/index.php/component/search/?searchword=direitos&searchphrase=all&Itemid=242	4

G3	Adopts a complex description to the document link	2	2f	INTERACTIVITY	1	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/o-ministerio/principal/secretarias/264-sgep-raiz/cartao-nacional-de-saude/l2-cartao-nacional-de-saude/8760-orientacoes-para-cadastramento.html	2
G3	The links are not similar to clickable links	2	2e	INTERACTIVITY	1	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/servicos/servicos.html	2, 4
G3	Information about H1N1 flu is not clear among search results	12	12a	CONTENT	0	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/cidadao/principal/agencia-saude.html	8
G3	The news should be highlighted in the home page	12	12	INFORMATION ARCHITECTURE	0	http://portalsaude.saude.gov.br/index.php/cidadao/principal/agencia-saude?filter-search=H1n1&filter-chapeu=&filter-mesdate=04&filter-anodate=2016&buscar=&a1efdf0b24f299518cb9d1b06184c5be=1	2

G3	The link for the news page is difficult to find in the home page	12	12	PHYSICAL PRESENTATION	0	http://portalsaude.saude.gov.br/index.php/component/search/?searchword=h1n1&searchphrase=all&Itemid=242	1
G3	The links are not similar to clickable links	2	2c	INTERACTIVITY	1	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/o-ministerio/principal/secretarias/264-sgep-raiz/cartao-nacional-de-saude/l2-cartao-nacional-de-saude/8760-orientacoes-para-cadastramento.html	2, 4
G3	The link "send us a question" has a poor contrast	2	2b	PHYSICAL PRESENTATION	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index60a0.html?option=com_content&view=category&layout=faq&id=285&Itemid=529	1
G3	The link for read more news is not clear	2	2a	PHYSICAL PRESENTATION	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	6
G3	The video is not accessible for users with wider auditory characteristics	21	21b	CONTENT	1	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	1, 7

G3	The search does not allow to filter by date	13	13a	INFORMATION ARCHITECTURE	0	http://portalsaude.saude.gov.br/index.php/component/search/?searchword=h1n1&searchphrase=all&Itemid=242	7
G3	It is difficult to find the right path for user registration	13	13	INFORMATION ARCHITECTURE	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	2
G3	The search for user registration does not return any result, users might expect the search to help in such task	13	13b	INTERACTIVITY	1	http://portalsaude.saude.gov.br/index.php/component/search/?searchword=cadastramento%20SUS&searchphrase=all&Itemid=242	10,7
G3	Inconsistency of standards, part of the information is in a pdf file, while the other part is in HTML text	14	14b	INTERACTIVITY	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index5309.html?option=com_content&view=article&id=8757&Itemid=426	4
G4	The content "public health history" and "public poll" have both the same design, but only one is a clickable link	14	14a	INTERACTIVITY	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/o-ministerio.html	4

G4	The website has many pages with distinct layouts	14	14	PHYSICAL PRESENTATION	1	http://portalsaude.saude.gov.br/index.php/o-ministerio/principal/secretarias/268-sgep-raiz/sgep/l2-mais-sobre-sgep/9587-videos	4
G4	The page presents a complex language to the user	15	15a	CONTENT	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/cidadao/legislacao.html	2
G4	Lack of compatibility among terms employed	15	15b	CONTENT	1	http://portalsaude.saude.gov.br/index.php/o-ministerio/principal/secretarias/505-sgep-raiz/doges-raiz/doges/fale-com-o-ministerio-da-saude/l2-fale-com-o-ministerio-da-saude/9905-como-registrar-manifestacoes-em-ouvidorias-do-sus-pela-internet	2
G4	The informative video does not have signal language option	21	21a	CONTENT	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	2
G4	The informative video does not have audio description	20	20b	CONTENT	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	1

G4	The page for sending a question employ difficult terms	15	15c	INTERACTIVITY	1	http://portalsaude.saude.gov.br/index.php/o-ministerio/principal/secretarias/505-sgep-raiz/doges-raiz/doges/fale-com-o-ministerio-da-saude/l2-fale-com-o-ministerio-da-saude/9905-como-registrar-manifestacoes-em-ouvidorias-do-sus-pela-internet	2,4
G4	The user might face difficulties to control the form (e.g. undo/redo) for sending questions	16	16	INTERACTIVITY	1	http://portalsaude.saude.gov.br/index.php/o-ministerio/principal/secretarias/505-sgep-raiz/doges-raiz/doges/fale-com-o-ministerio-da-saude/l2-fale-com-o-ministerio-da-saude/9905-como-registrar-manifestacoes-em-ouvidorias-do-sus-pela-internet	3
G4	Video player has no textual description for users that need screen readers	20	20a	CONTENT	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	H3

G4	Wrong use of the scroll inside the page	17	17	PHYSICAL PRESENTATION	1	http://portalsaude.saude.gov.br/index.php/o-ministerio/principal/secretarias/505-sgep-raiz/doges-raiz/doges/fale-com-o-ministerio-da-saude/l2-fale-com-o-ministerio-da-saude/9905-como-registrar-manifestacoes-em-ouvidorias-do-sus-pela-internet	5,8
G4	The website does not guide the user to the objective of the task (watch the video) because it is too confuse	18	18e	INFORMATION ARCHITECTURE	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/cidadao/principal/videos/video/0.html	3
G4	The video is not accessible for users with wider visual characteristics	20	20a	CONTENT	1	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	1, 7
G4	Important information is not highlighted	18	18f	INFORMATION ARCHITECTURE	0	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/o-ministerio/principal/secretarias/264-sgep-raiz/cartao-nacional-de-saude/l2-cartao-nacional-de-saude/8760-orientacoes-para-cadastramento.html	8

G4	Important information is out of user's focus	18	18c	PHYSICAL PRESENTATION	1	https://portaldocidadao.saude.gov.br/portalcidadao/index.htm	1
G4	Inadequated spaces among news list	18	18b	PHYSICAL PRESENTATION	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/cidadao/principal/agencia-saude795a.html?start=30	8
G4	The read more links (“leia mais”) are not grouped in an appropriate alignment (proximity, Gestalt)	18	18a	PHYSICAL PRESENTATION	0	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	8
G4	The news are not organized	18	18b	PHYSICAL PRESENTATION	0	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/cidadao/principal/agencia-saude.html	7,8
G5	Lack of alignment among information about SUS, guidance/prevention and ``health for you''	18	18d	PHYSICAL PRESENTATION	0	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	4

G5	The page does not inform clearly about the information wanted	15	15d	CONTENT	0	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/o-ministerio/principal/secretarias/264-sgep-raiz/cartao-nacional-de-saude/l2-cartao-nacional-de-saude/8760-orientacoes-para-cadastramento.html	1
G5	The unique way yo access information about fighting Dengue fever is through the banner	20	20c	CONTENT	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/o-ministerio/principal/secretarias/svs/dengue.html	4,6,7
G5	The search for flu (``GRIPE'') and ``h1n1'' does not provide any result	13	13f	INTERACTIVITY	1	http://portalsaude.saude.gov.br/index.php/cidadao/principal/agencia-saude?filter-search=&filter-chapeu=GRIPE&filter-mesdate=04&filter-anodate=2016&buscar=&3c16febe237012b4ecbc756125eec11e=1	4
G5	The page does not show relevance of results	13	13e	INFORMATION ARCHITECTURE	0	http://portalsaude.saude.gov.br/index.php/component/search/?searchword=direitos&searchphrase=all&Itemid=242	1

G5	After the search, the page still show information about search settings	13	13d	PHYSICAL PRESENTATION	0	http://portalsaude.saude.gov.br/index.php/component/search/?searchword=direitos&searchphrase=all&Itemid=242	1, 8
G5	The page does not inform clearly about the information searched	13	13c	INTERACTIVITY	1	http://portalsaude.saude.gov.br/index.php/component/search/?searchword=direitos&searchphrase=all&Itemid=242	2
G5	The website has poor support to TAB key	11	11c	INTERACTIVITY	0	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	1,7
G5	Although the webpage supports TAB navigation, it should have other accelerator keys	11	11b	INTERACTIVITY	0	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index2b2d.html?option=com_content&view=article&id=9950&Itemid=536	7
G5	Although the webpage supports TAB navigation, it should have other accelerator keys	11	11a	INTERACTIVITY	0	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/o-ministerio/principal/secretarias/svs/dengue.html	7

G5	The ``health for you" page inform about every kind of user, not only ``you" (the current user)	47	47	CONTENT	0	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/cidadao/saude-para-voce.html	2
G6	FAQ title is contradictory in comparison to the questions itself	46	46	CONTENT	0	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index60a0.html?option=com_content&view=category&layout=faq&id=285&Itemid=529	4
G6	No message for a blank page (probably an error)	45	45	INTERACTIVITY	1	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/o-ministerio/principal/secretarias/505-sgep-raiz/doges-raiz/doges/fale-com-o-ministerio-da-saude/l2-fale-com-o-ministerio-da-saude/9905-como-registrar-manifestacoes-em-ouvidor	9
G6	The organization of the site does not help users to achieve their goal of finding a video	24	24a	INFORMATION ARCHITECTURE	1		1, 4
G6	The informative video has no highlight, it is on the home page but users might face difficulties to find it	24	24	INFORMATION ARCHITECTURE	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	1,2

G6	The flags are difficult to recognize (circular flags with no text indication)	44	44	CONTENT	0	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	6
G6	Zooming distorts content	25	25	PHYSICAL PRESENTATION	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	3,7
G6	Users will loss content when they change the contrast	26	26	PHYSICAL PRESENTATION	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	5,7
G6	Lack of user control when users need to return to the default font	27	27	INTERACTIVITY	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	4,5
G6	To show the document, a new tab is opened without warning the user previously	43	43	INTERACTIVITY	1	http://portalsaude.saude.gov.br/images/pdf/2014/marco/21/cartilha-integra-direitos-2006.pdf	1
G6	The website responsiveness is not working properly, users may think that it is responsible because a bit of responsiveness is working, but it is not responsible	28	28	PHYSICAL PRESENTATION	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	5,7
G6	Dynamic content does not fit in responsive sizes	28	28a	PHYSICAL PRESENTATION	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/cidadao/principal/videos/video/0.html	5,8

G6	Static content does not adapt to responsive sizes	28	28b	PHYSICAL PRESENTATION	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index60a0.html?option=com_content&view=category&layout=faq&id=285&Itemid=529	5,8
G6	There is no help or documentation	29	29	CONTENT	1		10
G6	The local of help content is not clear	29	29	INFORMATION ARCHITECTURE	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	10
G6	The webpage has ambiguity, users may think that they must create a card	42	42	CONTENT	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	1
G6	The button "campanha publicitárias" is not working as the other buttons in the right side menu	30	30	INTERACTIVITY	0	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	4
G6	There is no contact information at the common place (footer)	41	41	PHYSICAL PRESENTATION	1		2, 4 ,6, 7
G6	There are identical terms that guide users to distinct pages	32	32	INFORMATION ARCHITECTURE	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/cidadao/acoes-e-programas/saude-sem-limite.html	2,4,8

G6	The video takes too much time to play, without any feedback informing users about its loading time	40	40	INTERACTIVITY	0	www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/cidadao/principal/videos/video/270.html	7
G6	Only a part of the website is translated after using the translate feature (the flags)	33	33b	CONTENT	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/cidadao/principal/english.html	4,2
G6	The right side form is out of users' focus	39	39	PHYSICAL PRESENTATION	1	https://portaldocidadao.saude.gov.br/portalcidadao/primeiroAcesso.htm	4,6
G6	The error message ``campo obrigatório'' does not help users to recover from the error	38	38	INTERACTIVITY	1	https://portaldocidadao.saude.gov.br/portalcidadao/primeiroAcesso.htm	9
G7	The font is too small	37	37	PHYSICAL PRESENTATION	1	https://portaldocidadao.saude.gov.br/portalcidadao/index.htm	1,8
G7	Inadequate font size (small)	37	37	PHYSICAL PRESENTATION	1	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	1, 8

G7	The website does not give proper feedback for users about their progress in direction to their objective	36	36	INTERACTIVITY	0	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/o-ministerio/principal/secretarias/264-sgep-raiz/cartao-nacional-de-saude/l2-cartao-nacional-de-saude/8760-orientacoes-para-cadastramento.html	1
G7	No option for undo, but using the browser option to reload the previous page	36	36a	INTERACTIVITY	0	https://www.planalto.gov.br/ccivil_03/Constituicao/Constitui%C3%A7ao_Compilado.htm	5
G7	The user may have difficulty to remember where they found a specific information, as h1n1	36	36b	INTERACTIVITY	0		6
G7	The webpage has no title, it is difficult for users to know where they are	36	36c	INTERACTIVITY	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index5309.html?option=com_content&view=article&id=8757&Itemid=426	1,4
G7	Inconsistent breadcrumb (does not show the user's steps properly)	36	36d	INTERACTIVITY	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index5309.html?option=com_content&view=article&id=8757&Itemid=426	1,6

G7	There is no breadcrumb	36	36	INTERACTIVITY	1	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/o-ministerio/principal/secretarias/264-sgep-raiz/cartao-nacional-de-saude/l2-cartao-nacional-de-saude/8760-orientacoes-para-cadastramento.html	4
G7	The group ``Sobre o SUS" suggest information about registration, but it does not have such information	35	35	INFORMATION ARCHITECTURE	0	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/o-ministerio/principal/secretarias/264-sgep-raiz/cartao-nacional-de-saude/l2-cartao-nacional-de-saude/8760-orientacoes-para-cadastramento.html	5
G7	The website has no tooltip, or other complementary information, teaching about how to change the size of video player window	34	34	CONTENT	0	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/cidadao/principal/videos/video/0.html	5
G7	The Spanish flag does not translate the whole website	33	33	CONTENT	1	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	2

G7	The information groups ``SUS'' and ``Ministérios'' in more than one place at the same page might make users think that the information they want does not exist, but it can be in another page	32	32	INFORMATION ARCHITECTURE	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html#	8
G7	The icons ``Redes e Programas'' at the middle of the page might make users understand that place as the end of the page	31	31	PHYSICAL PRESENTATION	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html#	8
G7	The help is not easy to access	29	29	INFORMATION ARCHITECTURE	1	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/o-ministerio/principal/secretarias/264-sgep-raiz/cartao-nacional-de-saude/l2-cartao-nacional-de-saude/8760-orientacoes-para-cadastramento.html	10
G7	The responsiveness is not working properly	28	28	PHYSICAL PRESENTATION	1	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	1
G7	The A+ feature overlaps content	25	25	PHYSICAL PRESENTATION	1	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	7, 8

G7	The option ``fale conosco'' is at a difficult place to find, and lacks highlight	23	23	INFORMATION ARCHITECTURE	0	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	6
G7	There is no option to stop the news carousel	22	22	PHYSICAL PRESENTATION	0	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	3
G7	The video has no legend for those users who need it	21	21	CONTENT	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	2
G7	There is no alternative description for images	20	20	CONTENT	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	1,3
G7	The link description is wrong, because it does not help users to achieve their goal	43	43b	INTERACTIVITY	1	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/o-ministerio/principal/secretarias/264-sgep-raiz/cartao-nacional-de-saude/l2-cartao-nacional-de-saude/8760-orientacoes-para-cadastramento.html	2

G7	The pdf document does not speak the users language (using technical terms)	43	43b	INTERACTIVITY	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/o-ministerio/principal/secretarias/264-sgep-raiz/cartao-nacional-de-saude/l2-cartao-nacional-de-saude/8760-orientacoes-para-cadastramento.html	2
G7	The text showed on the link does not inform properly about the content users will find after clicking on it	43	43c	INFORMATION ARCHITECTURE	0	https://portaldocidadao.saude.gov.br/portalcidadao/index.htm	2,4
G7	The text showed on the link does not inform properly about the content users will find after clicking on it	43	43	INTERACTIVITY	0	http://www.conselho.saude.gov.br/biblioteca/livros/AF_Carta_Usuarios_Saude_site.pdf	2,4
G7	The unique link for the information is in a image without description	20	20	CONTENT	1	http://adelimasalgadao.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	1
G7	It is not possible to pause the banner about campaigns	19	19	INTERACTIVITY	0	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	3
G7	The content alignment is poor, inappropriate for skimming content	18	18	PHYSICAL PRESENTATION	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	6,8

G7	The system creates a subpage inside the original page, instead of showing the form directly)	17	17	PHYSICAL PRESENTATION	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/o-ministerio/principal/secretarias/505-sgep-raiz/doges-raiz/doges/fale-com-o-ministerio-da-saude/l2-fale-com-o-ministerio-da-saude/9905-como-registrar-manifestacoes-em-ouvidor ou http://ouvprod01.saude.gov.br/ouvidor/CadastroDemandaPortal.do	4
G7	It is not clear to users that they must let all terms without selection in case they want to select in all terms	15	15	INFORMATION ARCHITECTURE	1	http://portalsaude.saude.gov.br/index.php/cidadao/principal/agencia-saude?filter-search=H1N1&filter-chapeu=&filter-mesdate=04&filter-anodate=2016&buscar=&47a919fb7d53c1f42cc134073ad79eb1=1	6,5
G7	Although the webpage supports TAB navigation, it should have other accelerator keys	11	11	INTERACTIVITY	0	http://portalsaude.saude.gov.br/index.php?option=com_content&view=article&id=11365&Itemid=695	7

G7	There is no option to skip to content (for those users who need it)	10	10	INTERACTIVITY	0	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/cidadao/saude-para-voce.html	7
G7	Instructions about how to get the SUS card are not clear	7	7	CONTENT	1	https://portaldocidadao.saude.gov.br/portalcidadao/primeiroAcesso.htm	10
G7	The page about fighting the Dengue fever is in English	5	5	CONTENT	1	http://combateaedes.saude.gov.br/en/	2,7
G7	The second level of the main menu is not logical for users	4	4	INTERACTIVITY	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	6
G7	Too many information	3	3	CONTENT	1	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.php/servicos/servicos.html	8
G7	No highlight for the link `Em destaque" (Highlights)	2	2	INTERACTIVITY	1	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	1
G7	Poor visibility of the link of the requested information	2	2	INTERACTIVITY	1	http://adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	1

G7	Poor contrast (grey and light grey) in the main menu	1	1	PHYSICAL PRESENTATION	0	http://www.adelimasalgado.com.br/experimentos/portalsaude/portalsaude.saude.gov.br/index.html	4,8