UNIVERSIDADE FEDERAL DO PAMPA

Fabienne Charles

A Study on the Perception to Usability of the Thoth Tool

Alegrete
2022

Fabienne Charles

A Study on the Perception to Usability of the Thoth Tool

Master thesis presented as partial requirement for obtaining the degree of Masters of Software Engineering at Universidade Federal do Pampa.

Supervisor: Prof. PhD. Elder de Macedo Rodrigues

Co-supervisor: PhD. Ildevana Poltronieri

Alegrete
2022

Ficha catalográfica elaborada automaticamente com os dados fornecidos
pelo(a) autor(a) através do Módulo de Biblioteca do
Sistema GURI (Gestão Unificada de Recursos Institucionais) .

C475s Charles, Fabienne
A study on the perception to usability of the Thoth tool / Fabienne Charles.
137 p.

Dissertação(Mestrado)-- Universidade Federal do Pampa, MESTRADO EM ENGENHARIA DE SOFTWARE, 2022.
"Orientação: Elder de Macedo Rodrigues".

1. Usability Evaluation. 2. Heuristic Evaluation. 3. Usability Testing. 4. Thoth tool. I. Título.

Este trabalho é dedicado a todas às crianças adultas nascidas
de um pai irresponsável que, quando pequenas,
sonharam em se tornar cientistas.

Fabienne Charles

A Study on the Perception to Usability of the Thoth Tool

Dissertação apresentada ao Programa de Pós-graduação em Engenharia de Software da Universidade Federal do Pampa, como requisito parcial para obtenção do Título de Mestre em Engenharia de Software.

Dissertação defendida e aprovada em: 29 de julho de 2022.

.

Banca examinadora:

_____

Prof. Dr.. Elder de Macedo Rodrigues - Orientador
UNIPAMPA

_____

Prof. Dra. Ildevana Plotronieri Rodrigues - Coorientadora
2Day IT Solution

_____

Prof. Dr. Fábio Paulo Basso
UNIPAMPA

_____

Prof. Dra. Raquel Mainardi Pillat
UFRJ

Assinado eletronicamente por **ELDER DE MACEDO RODRIGUES**, **PROFESSOR DO MAGISTERIO SUPERIOR**, em 24/08/2022, às 18:34, conforme horário oficial de Brasília, de acordo com as normativas legais aplicáveis.

Assinado eletronicamente por **FABIO PAULO BASSO**, **PROFESSOR DO MAGISTERIO SUPERIOR**, em 24/08/2022, às 18:37, conforme horário oficial de Brasília, de acordo com as normativas legais aplicáveis.

Assinado eletronicamente por **Raquel Mainardi Pillat**, **Usuário Externo**, em 24/08/2022, às 18:43, conforme horário oficial de Brasília, de acordo com as normativas legais aplicáveis.

Assinado eletronicamente por **Ildevana Poltronieri Rodrigues**, **Usuário Externo**, em 24/08/2022, às 19:10, conforme horário oficial de Brasília, de acordo com as normativas legais aplicáveis.

A autenticidade deste documento pode ser conferida no site https://sei.unipampa.edu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **0908989** e o código CRC **0EBF8805**.

ACKNOWLEDGEMENTS

Philippians 4:13 I can do all this through him who gives me strength.

Through these lines I want to express my gratitude to all the people (friends and family) who with their support have collaborated in the realization of this work.

I thank God for giving me health and guiding my decisions. My most sincere gratitude to all the institutions (educational and financial) that gave me the opportunity to be a Magister, especially to the Universidade Federal do Pampa (UNIPAMPA).

In a very special way, I want to thank my advisors PhD. Elder de Macedo Rodrigues and PhD. Ildevana Poltronieri for the appropriate guidance, support and careful discussion that allowed me to make a good use for the realization of this thesis.

I thank all the researchers who contributed to making this work possible, my colleague Adriana Charpe Pimenta dos Santos for the inestimable help and patience she has given me since my first class and Flavia Amin Barbosa, my research colleague. Thanks to Alexandre Osorio Roballo Guedes who helped me and guided me since my arrival in Alegrete and was always available to collaborate with my research and Daniel Francisco de Luca who was the first person who welcomed me in Brazil. I thank my friend Sergeline Louis for her unconditional support.

"If you don't go after what you want, you'll never have it.
If you don't ask, the answer is always no.
If you don't step forward, you're always in the same place."
(Nora Roberts)

ABSTRACT

Usability is an extremely important factor for a software, so many studies have been done in relation to the evaluation of usability in web tools. This study, through a Systematic Literature Review, seeks to know which methods, tools for empirical evaluation, and metrics are being used in evaluations, and the limitations found by evaluators. The results of the SRL were support for the methodology conducted in this study, which is of exploratory qualitative nature, where a heuristic evaluation and a usability test were applied to users of Thoth, a web tool for automation of Systematic Literature Reviews. The main purpose was to map the users' limitations and find heuristic flaws in the tool, in order to catalog them, aiming at a better usability of this system. As a result, we identified that the system violated nine of ten heuristics. According to the severity degrees assigned, error prevention is the most frequent. Concerning the usability test, the opinions were very diverse between the graduate and post-graduate users in terms of ease of adaptation but at the beginning they had almost the same preoccupations.

Key-words:Usability, Usability Evaluation, Usability Evaluation Method, Web System, Heuristic Evaluation, Usability Testing.

RESUMO

A usabilidade é um fator de extrema importância para um software, devido a isto, muito se estuda em relação a avaliação de usabilidade em ferramentas web. Este estudo, através de uma Revisão Sistemática de Literatura, busca saber quais métodos, ferramentas para avaliação empírica e métricas estão sendo utilizadas nas avaliações, deseja-se saber também as limitações encontradas pelos avaliadores. Os resultados da Revisão Sistematica Literatura (RSL) foram suporte para a metodologia conduzida neste estudo, que é de caráter qualitativo exploratório, onde aplicou-se uma Avaliação Heurística e Teste de Usabilidade a usuários da Thoth, uma ferramenta web para automação de Revisões Sistemáticas de Literatura. O intuito principal foi mapear as limitações dos usuários e encontrar as falhas heurísticas na ferramenta, para assim catalogá-las visando a melhor usabilidade deste sistema. Como resultado, identificamos que o sistema violou nove de dez heurísticas. De acordo com os graus de severidade atribuídos, a prevenção de erros é a mais frequente. Referente ao teste de usabilidade, as opiniões são variadas entre usuários graduandos e pós-graduandos quanto à facilidade de aprender, embora que no início possuem as mesmas dificuldades.

Palavras-chave: Usabilidade, Avaliação de Usabilidade, Método de Avaliação de Usabilidade, Sistema Web, Avaliação Heurística, Teste de Usabilidade.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ACM Association for Computing Machinery

AEC Ambulatory Emergency Care

CSWR Client-Side Web Refactorings

DE Data Extraction

DOM Document Object Model

EC Exclusion Criteria

EWEB Experimentation in the WEB

H Heuristic

HCI Human-Computer Interaction

HE Heuristic Evaluation

HEUA Heuristic Evaluation with Usability and Accessibility

IC Inclusion Criteria

ICT Informed Consent Term

IEEE Institute of Electrical and Electronics Engineers

ISO International Organization for Standardization

IT Information Technology

MDWD Model-Driven Web Development

MDWE Model Driven Web

MPA Methodology of Academic Research

NASA TLX NASA Task Load Index

PCs Personal Computers

PU Perceived Usefulness

QA Quality Assessment

RQ Research questions

RQs Research questions

RSL Revisão Sistematica Literatura

RUM Real-time Usage Mining

S Study

SAGAT Situation Awareness Global Assessment Techniqu

SART Situation Awareness Rating Technique

SLR Systematic Literature Review

SMS Systematic Mapping of Studies

SUS System Usability Scale

TA Think-Aloud

UX User Experience

WaPPU "Was that Page Pleasant to Use"

WCAG Web Content Accessibility Guidelines

Web DUE Web Design Usability Evaluation

WUEP Web Usability Evaluation Process

TABLE OF CONTENTS

A INTRODUCTION

Web systems are now mainstream and represent the majority of new projects in system development (ESCALONA; KOCH et al., 2006). The reason for this is none other than the multiple benefits of web systems (LOWE, 2003) and (KIENLE; DISTANTE, 2014), since they can be merged with sophisticated enterprise architecture, a complex information architecture and a highly component-based technical architecture for a better change in the business model (RUSSELL, 2000).

Web systems are services and systems that can be used over the Internet. Basically, they can be used on devices equipped with a web browser, such as Personal Computers (PCs), smartphones and tablets, regardless of the operating system, such as Windows, Android and iOS. The web system storages data on the server, and allows access to the site and use of the service from anywhere, as long as an Internet environment is available (JAZAYERI, 2007). Simultaneous access by an unspecified number of users is possible within the capacity of the server.

In the modern context, it is important that the interfaces of interactive systems are easy to use, with high usability. While usability is a really important metric in the development process of a system to ensure its success, many developers do not take into consideration these concepts to develop an adequate and practical design (YAN; GUO, 2010).

In general, the interfaces are done at the final stage of software development (HOLZINGER, 2005). In fact, analysts and programmers agree on that point because they do not see the necessity of spending time and money to seriously consider and involve users in the design. However, experience has shown that poorly designed interfaces can have serious consequences (BAROLLI et al., 2006). If the interfaces are not efficient, a risk of wasting time, money and resources can occur, a disaster that generates loss of users (SUDUC; BIZOI; FILIP, 2010).

The usability in software has recently become a focus of attention due to the challenges of making web systems (graphical user interface) easy to use. Usability is a complex word that means ease of use; in the field of information technology, however, this term has much more meaning (BEVAN, 1995). In fact, according to International Organization for Standardization (ISO) 9241, For the ISO, usability is when a product is used by a specific user to achieve a specific purpose under specific conditions of use (GEORGSSON; STAGGERS, 2016). It is defined as "effectiveness, efficiency and degree of user satisfaction". In other words, usability is a characteristic that determines how understandable and convenient the user interface is. This term focuses more on functionality and ergonomics, on its simplification and improvement, than on the aesthetic and visual content.

Users are selective and do not waste time going through a system with interfaces

that have a complex structure or looking for the information they need in poorly ordered content. It is important to understand that usability is not only about understandable form, but also about quality content (HASSENZAHL, 2008). Interface design is not only how you see, nor how you feel, but also how it works. To mitigate the ease-of-use inconveniences that may present the interfaces of the systems, enhanced usability is key.

Jakob Nielsen is one of the most recognized experts in the usability field. He began working on the issues of user interface usability, design simplicity, and website structure in the 1980s (NIELSEN, 1994). Nielsen formulated the results of his research on the subject of heuristics in the form of usability principles. The Nielsen heuristics and the ISO metrics agree that usability is about user satisfaction and a performance that does not overload the user. In addition, it is an indicator that users and conditions of use achieve their objectives under specific conditions.

According to Nielsen's principles, there are five main components of usability from the user's point of view: 1) Learn-ability: how easy it is for the user to perform basic tasks if they come across the interface for the first time; 2) Efficiency: how fast a user completes a task when they are already familiar with the interface; 3) Memorability: how easy it is for the user to restore the skills of working with the interface after a long period of time, 4) Errors: how often errors occur in use, their severity, and how easy it is for the user to recover from them, and 5) Satisfaction: how friendly the interface is to use (NIELSEN, 1994). The usability of a system is based on principles or slogans, among which "the simpler it is, the better" is one of the most famous (NASR et al., 2016). The more familiar the structure and interface the user sees, the easier it is to navigate to find the desired section.

This research aims to make a study of the usability of the Thoth tool. Thoth is a web system developed by the LESSE research group of Universidade Federal do Pampa. With the perspective of mitigating the challenges to perform an SLR, it is a solution of great potential that helps the automation process of systematic reviews. With the tool, it is possible to map and execute the protocol of the literature review, with shared access among the researchers involved (MARCHEZAN et al., 2019). Due to the irregularities of the system, it causes inconveniences to its users, and makes them look for alternatives to conduct their reviews. All web tools with expertise in systematic literature review also need to be subjected to a usability evaluation, for better results for the end user. For more information you can access to the tool through the link:<http://200.132.136.13/Thoth/>.

A.1 Motivation

The technological development that has been observed for some decades is quickly consolidating in recent years in people's daily lives, creating a constant need for interaction with a variety of digital media. However, since the early stages of computer development, the need to study human communication with computers has become evident. This need

has brought to the forefront a new research topic called Human-Computer Interaction (HCI) (GRUDIN, 2017).

It is important that a web system is appropriate for its target audience. Usability can not be determined simply by looking at a web system. Usability depends on whether the target user can meet their needs and achieve their goal with the information on the site. In other words, a solid understanding of the target user's purpose is an important point when considering usability.

A user-friendly web system allows you to significantly increase the number of users, both temporarily and over a longer period of time. Users always derive maximum satisfaction from interacting with properly configured and organized web system content.

A.2 Objectives

A.2.1 Main objective

Evaluate the usability of the Thoth tool in order to contribute to the improvement of the design of the graphical interface of the system, suggesting changes according to how end users perceive its usability.

A.2.2 Specific objectives

- Identify and analyze the usability evaluation methods used to assess web tools.
- Verify how these methods have been used.
- Identify what are the problems reported when evaluating the usability of web tools.
- Apply a heuristic evaluation and a usability test for users of the Thoth tool.
- Catalog the errors and violated heuristics in the Thoth tool.
- Identify possible solutions to correct the usability problems in the tool.

A.3 Methodology

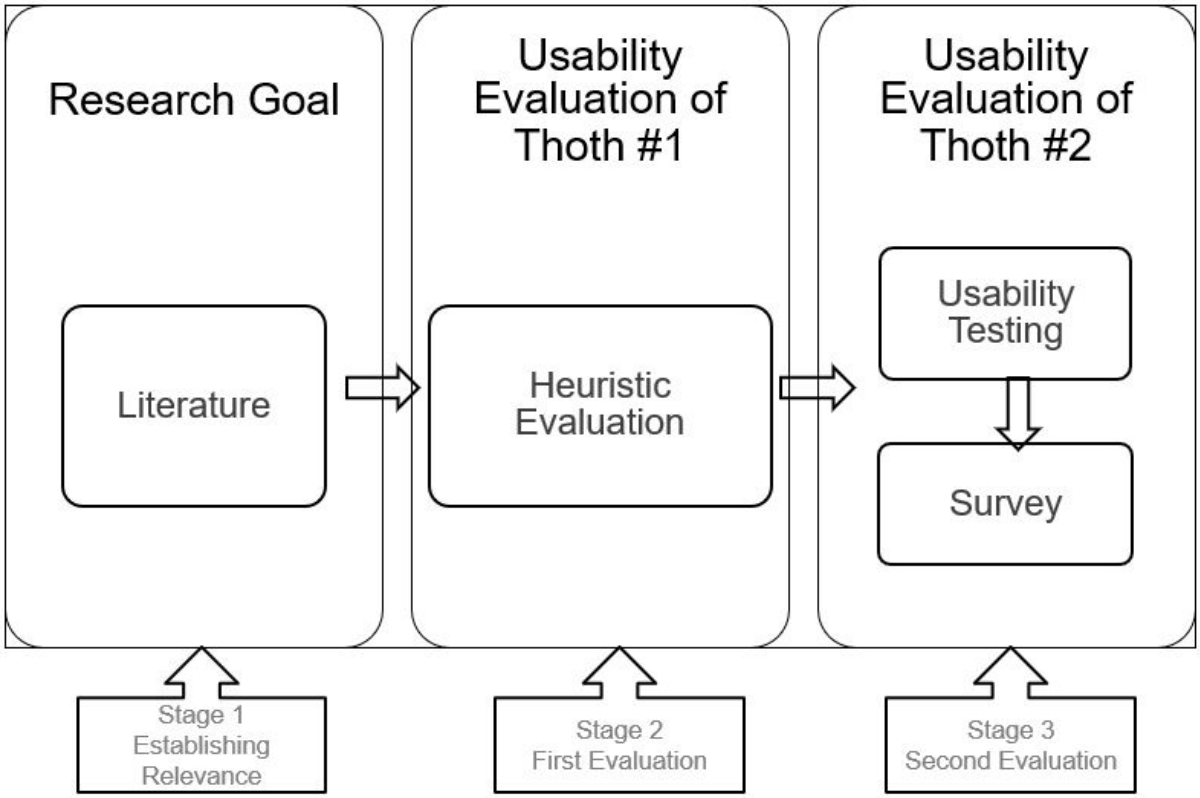
In the perspective of evaluating the usability of the Thoth tool, and to facilitate and optimize its use, a three (3) stage methodology was used. First, a Systematic Literature Review (SLR) was performed. The main objective of this SLR was to detect a large number of possible usability evaluation methods for web tools along with their metrics and tools used, their main limitations and their modes of use. This review was conducted by more than one evaluator, in collaboration with an undergraduate student. After performing the SLR, we considered the threats and validations that are inconvenient and may affect the final result. The possibility that some methodologies were not mapped in other digital libraries not used in the search process of the studies cannot be

excluded. From the perspective of researching primary sources, an important part of the SLR is based on the definition of key and relevant terms together with their respective synonyms, related to the objective and Research questions (RQs).

In the second phase, an inspection of the tool was conducted through a Heuristic evaluation with six (6) evaluators in the usability area. The main objective of this evaluation was to discover the heuristic violations of all the interfaces of the tool in order to improve it.

In the last stage of this methodology, a usability test was conducted through a survey with twenty-five (25) end users of Thoth.

Figure 1 – Methodology

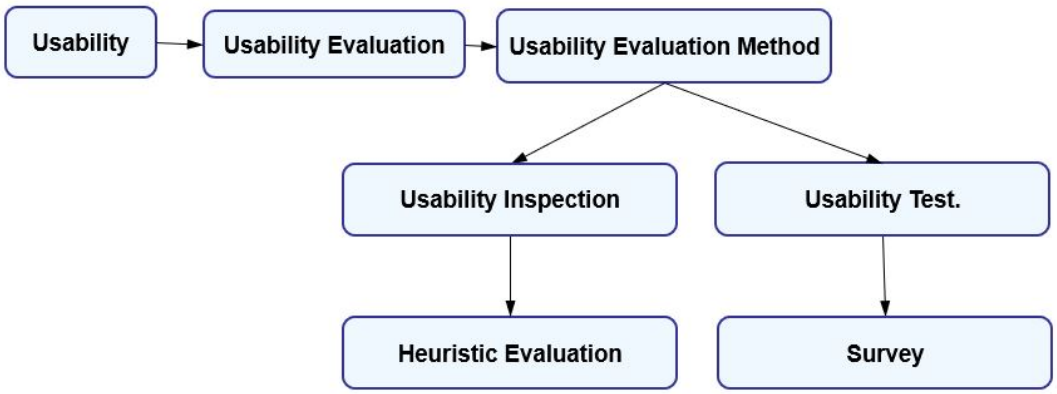


Source: Author

B BACKGROUND

In the chapter we present the definitions of the basic concepts that we consider essential for the development of our work. We introduce the terminology and describe the main concepts addressed throughout this work such as. We chose, as well as key words in our topic: Usability, Usability Evaluation, Usability Evaluation Method: Heuristic Evaluation, Usability Test. (See Figure 2)

Figure 2 – Background plan



Source: Author

B.1 Usability Evaluation

"Usability refers to the grade of effectiveness, efficiency and satisfaction provided by a product when used for a specific purpose by a specific user under a specific usage scenario (STANDARD, 1998). The according to (KERZAZI; LAVALLéE, 2011) and (BEVAN et al., 2016)" ISO/IEC 9126-1 defines usability as "the ability of a software product to be understood, learned, used and able to attract users in specific usage scenarios" (ALKILIDAR; COX; KITCHENHAM, 2005) . ISO/IEC 9241 affirms the "Usability is the efficiency and satisfaction with which a product enables specific objectives to be achieved by specific users in a specific context of use".

In his book Principles of Usability Engineering, Dr. Jakob Nielsen, a leading usability researcher, defines usability as follows:

"Ease of learning: The system must be easy to learn so that users can start using it immediately. Efficiency: Once learned, it must be able to be used efficiently so that it can be highly productive. Memorability: Make it ready to use, even if the user has not used it for a while, and it should not cause a fatal error. Errors: reduce the error rate, allow recovery if an error occurs, and should not cause a fatal error. Satisfaction: It must be pleasant for users to be personally satisfied and like it." (NIELSEN, 1994)

The usability evaluation is to evaluate the "usability" of the software to verify whether it conforms to the usability standard. Currently there are more than 20 usability evaluation methods, which can be divided according to (RIIHIAHO et al., 2000): into expert evaluation and user evaluation according to the personnel involved in usability evaluation; according to the software development stage of evaluation, usability evaluation can be divided into formative evaluation and summative evaluation. Formative evaluation refers to the process of software development or improvement, asking users to test the product or prototype and improve the product or design through the data collected after testing until the required usability goals are achieved. The purpose of formative evaluation is to find as many usability problems as possible and to improve the usability of the software by solving the usability problems. The purpose of summative evaluation is to evaluate multiple versions or products horizontally and generate evaluation data for comparison.

For many developers a usability evaluation is an optional tool because it incurs additional costs. If usability issues arise during an evaluation at a later stage of the development process, the following difficulties often arise: certain adjustments and improvements to the product are required based on the results of the usability evaluation. To make such adjustments is very costly. Therefore, the improvements may reduce or delay its use or commercialization. To mitigate that problem is to give a high priority to usability evaluation in product development. Ideally, usability evaluation is not a separate phase in product development (ABELEIN; PAECH, 2015) , but interspersed throughout all phases. Users should be involved in product development at an early stage.

B.1.1 Heuristic Evaluation

Heuristic evaluation is a detection method based on user interface design principles. Typically, evaluation personnel are design experts with a wealth of professional knowledge, so heuristic evaluation is also called expert evaluation. Dr. Nielsen believes that one person usually finds only 35% of the problem, so it takes about five people to fully identify the problem (NIELSEN; MOLICH, 1990).

During the evaluation, assessors are required to evaluate individually to avoid the common assessors being influenced by "opinion leaders". The evaluators record the prob-

lems they find. After the evaluation, they hold an evaluation meeting and summarize the results of the evaluation. The recorder finds summarizes the most frequently encountered problems, summarizes the results of the evaluation and makes recommendations for revision (DYKSTRA, 1993). The advantage of heuristic evaluation is that it can quickly find the problem of product usability, but at the same time, this method is too subjective and the cost of hiring some experts to evaluate the start-up product is not small (MANZARI; TRINIDAD-CHRISTENSEN, 2006).

The interface is evaluated using a predefined list of features or aspects of the user interface that are generally considered useful. Heuristic evaluation is generally faster and cheaper than usability testing, although it has drawbacks and should be used early in development (MANKOFF et al., 2003).

In the following is the list of the ten (10) heuristics based on Nielsen's principles(NIELSEN, 1994).

1-Visibility of system status: check if you are consistently and adequately feeding back the status of your system to your users. We also verify that the feedback is fast and adequate.

2- Harmony between the system and the real world: test that the system is made to fit the real-world environment. Verify that you are using terms, phrases and concepts that are familiar to you, rather than jargon or internal terms.

3- User control and degrees of freedom: make sure you create a situation where the user can exit the system at any time. Also check if the user can cancel or redo the operation if he makes a mistake or uses the function.

4- Consistency and standardization: make sure you don't use different terms to describe the same thing. Also, make sure you can get the same result by doing the same thing.

5- Error prevention: make sure the design is designed to prevent the error itself from occurring, rather than improving the countermeasures after the error (error message, page transition, etc.).

6- As you can see without remembering: make sure you don't create a situation where you have to remember information as the user navigates the page. Make sure that instructions for using the system are visible or easy to retrieve at any time.

7- Flexibility and efficiency: be sure to provide shortcuts and customizations for advanced users.

8- Aesthetic and minimalist design: check for unnecessary elements that are not related to the interface.

9- Aesthetic and minimalist design: check for unnecessary elements that are not related to the interface.

10- Help and manual: after designing it so that it can work without a manual, prepare the content for support.

One of the main disadvantages of heuristic evaluation is that it applies common standards to different types of systems (HVANNBERG; LAW; LÉRUSDÓTTIR, 2007). A feature that may be necessary in one software may be unnecessary in another; while some features that may be considered bad design for some programs may be useful for others. However, many companies still use experts to perform heuristic evaluation because the process is faster and cheaper than long-term usability testing with large groups of users (MASIP et al., 2012) and (MURILLO; SANG; PAZ, 2018).

B.1.2 Usability Testing

Usability testing is a method of discovering problems related to efficiency and satisfaction in a product by observing users using a product to complete typical tasks. Observe actual user operation during the process of actual use, record and analyze the problems encountered by the user in using the product in detail, to discover existing usability problems in the product, collect qualitative and quantitative data to help improve the product and identify target users Satisfaction with the product (DUMAS; SALZMAN, 2006).

Usability is defined as: "according to human functional characteristics, the system can be easily and effectively used by a specific group of users, after specific training and user support, in a specific environmental situation, to complete a specific set of tasks". And it is divided into four factors: effectiveness, learnability, flexibility and attitude (SHACKEL; RICHARDSON, 1991).

Usability testing is measuring the usability of a product's interface by users (HAN et al., 2001). Usability testing shows how the product performs to the user's perspective, identifies problem areas in the interface and allows the product to be seen through the eyes of the users (KUNIAVSKY, 2003). During usability testing, the user performs typical tasks with the product in the presence of the test leader (TARKKANEN et al., 2013).

Usability testing can be performed at different phases of product development (AU et al., 2008). However, it is most recommendable to start performing it already at the initial stages of interface development, even before it is implemented in the program code. This allows to immediately make the appropriate and necessary adjustments and make the interface convenient. The earlier to make the changes to the interface, the easier, faster and, therefore, the cheaper it will be to do it (SNYDER, 2003).

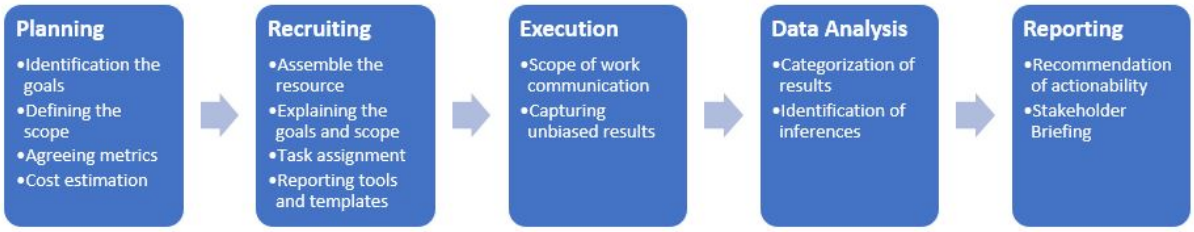
B.1.3 Survey

Surveys: as an indicator of user experience, surveys can be used in conjunction with usability testing as a method of monitoring or collecting feedback (ALBERT; TULLIS, 2013).

The usability of a system is an important factor in ensuring user satisfaction and increasing the probability of executing a user scenario (FLAVIAN; GUINALíU; GURREA,

2006). Usability testing is useful for any web tool or company interested in guiding the customer through the complete user journey on the site. It allow to immediately see the weak and inconvenient points for the user, remove their missing elements and fix them. In the following, an diagram of the protocol to be followed for conducting a usability test is presented. (See Figure 3)

Figure 3 – Steps of usability testing.



Source: link:<https://tede2.pucrs.br/tede2/handle/tede/9950>

B.1.4 Metrics

The metrics are quantitative measures of usability (HUSSAIN; KUTAR, 2009). As a result of testing, it always finds a set of problems in the interface (BANGOR; KORTUM; MILLER, 2009). Metrics, on the other side, allow to understand how good or bad the overall. To conduct a usability evaluation, you must first define the metrics to be evaluated and define them appropriately (FINSTAD, 2010).

Ease of learning: the system must be easy to l earn so that users can start using it quickly.

Efficiency: the system should be efficient to use so that, once the user is familiar with it, it can be highly productive. Resources expended in relation to the results obtained.

Easy to remember: it should be easy to remember so that the user can use it immediately upon use, even if it is not used for a while.

Error rate: the system should have a low error rate, be less error prone during system testing, and be easily recoverable if an error occurs. In addition, no fatal errors should occur.

Subjective satisfaction: the system must be pleasant for users to be personally satisfied and like it. A user's physical, cognitive and emotional perception of the extent to which the user's needs and expectations resulting from the use of a system, product or service are satisfied.

B.1.5 Usability vs. User Experience (UX)

The usability is more narrow than UX, it focuses on task completion. Rather, UX is the combined result of factors such as appearance, functional composition, system performance and interaction behavior (BARGAS-AVILA; HORNBÆK, 2011). The concepts related to UX include design, ergonomics, HCI, accessibility, marketing and usability.

According to the ISO 9241-11 definition: Usability refers to "the effectiveness, efficiency and satisfaction of users in achieving specific objectives in a specific environment" (ISO 9241-11) (BEVAN; MACLEOD, 1994). Usability is an essential quality measure for interactive systems, representing whether the product is effective, easy to learn, easy to use, high performance, with fewer errors and high user satisfaction (LIU; ZHU, 2012). Usability is mainly about the functional part of the product (BARBIERI et al., 2013).The According to ISO 9241-210 UX therefore refers to "the user's interaction with a product, experience and feelings in all aspects"(YOGASARA et al., 2011). User experience also refers to the subjective emotions and attitudes of people when using a particular product. It includes functional scope, product branding, psychological expectations and real-time emotional feelings.

The typical usability issues in user-centered design include: comprehensively evaluate product efficiency and effectiveness and focus on these as design goals; evaluate user comfort and satisfaction levels as design goals; design a product that is easy to use and can be evaluated for suitability issues. While typical UX issues in user-centered design include: design and evaluate what users do in the end-to-end interactive process while using the product; maximize motivation (product use), awareness (product) and emotional resonance (JOKELA et al., 2003).

Thus, although there is no fundamental difference between usability measures and UX measures at a given point in time, the difference in emphasis between task performance and enjoyment leads to different concerns during development (BEVAN, 2009) .

B.1.6 Thoth Tool

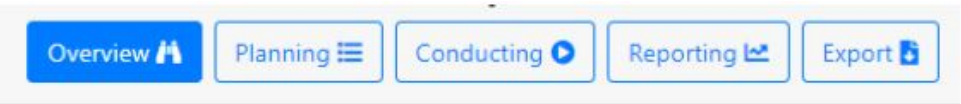
Thoth is a web tool for shared access between researchers involved in the performing RSL. was developed by the LESSE group (MARCHEZAN et al., 2019). In the following is an overview of its organization link:<http://200.132.136.13/Thoth/>. Like (FERNANDEZ; ABRAHÃO; INSFRAN, 2012), the tool follows a four-step standard to automate and support bibliographic reviews. It is worth mentioning that Thoth offers the opportunity to its users to conduct their revisions with more than one reviewer.

B.1.6.1 How the tool works?

After accessing the main page, you log in, if you do not have a registered user, you must make a new one in order to register or access an existing project or create a

new project. The system is practically in English. Basically, a project is divided into 5 fundamental stages, in which each stage has a set of information that must be filled in (see the Figure 4)

Figure 4 – Stages of the systematic literature review in the Thoth tool.



Source: (MARCHEZAN et al., 2019)

Step1. Overview

The project overview presents the following information: description, objectives, information, and the roles of each of the evaluators.

Step2. Planning

The planning stage is divided into 8 sub-steps:

Overall information: In this sub-stage the domains are registered, the keywords are defined, the language is set, the type of study is selected, and the time frame for conducting the review is established.

Research Questions: In this sub-step the research questions that will serve as the basis for the study must be defined.

Databases: In this sub-step the databases that will be consulted during the RSL are defined.

Search String: Stage in which the keywords and their respective synonyms are recorded in order to generate the generic search strings to be applied to each of the defined databases.

Search Strategy: In this step the strategy with which the RSL will be conducted is defined.

Criteria: The selection criteria section is the moment in which the criteria that will serve for inclusion and exclusion of a study in its selected work base are identified.

Quality Assessment: In this step the scores of one of the studies are established, i.e. this score that organizes the priority of reading the selected work.

Data Extraction: In this step it is defined the type of data that is sought in the studies, as well as the description of the information to be raised and the research question that the evaluator intends to answer from the analyzed work.

Step3. Conducting

Conducting, in turn, is divided into 4 sub-steps:

Importing studies In this sub-stage the studies are added according to their respective bases, this is an automatic process from the file generated by the database that will be included in Thoth.

Study Selection In study selection all the studies to be filtered are displayed, taking into account their relevance based on the inclusion or exclusion criteria.

Study Evaluation In this sub-stage the selected studies should be evaluated according to the research questions defined in the RSL protocol.

Data extraction In this sub-stage the data extraction of the evaluated studies is performed as defined in the previous steps.

Step4. Reporting

The Study Report presents the information in an organized way following each of the steps performed in the RSL. This Thoth feature provides a summary of the selection of studies from import into the tool to data extraction, as well as the daily activities performed by the researchers. An overview of the study selection is also provided, and a quantitative analysis presented through explanatory graphs.

Step5. Export

The export step offers the possibility to export all steps of the review in Latex format.

B.1.7 Related work

Some papers in the past have presented and discussed the results of a systematic review of the literature on web tool usability evaluation. (PAZ; POW-SANG, 2016) Despite this large number of methods, he said, the most appropriate technique for a particular scenario has yet to be determined. The most appropriate technique for a particular scenario. He also stated that the emergence of hybrid categories has forced academics to propose particular evaluation tools, such as usability questionnaires for specific domains, heuristics for a particular type of software, variants of a usability method. A deeper analysis is also needed in each category, especially in the methodology, establishing the differences and how they affect the final result.

The methodologies used to evaluate usability and UX, at least in separate phases, are often not specific. When they are specific, they are designed only for the task at hand, with no standardized aspects to provide replicability (GUERINO; VALENTIM, 2020).

In the (FERNANDEZ; INSFRAN; ABRAHãO, 2011) demonstrate in the result of his work results obtained demonstrate the need for usability evaluation methods developed specifically for the web domain that can be better integrated into the life cycle of web applications, particularly during the early stages of the web development process.

The multitude of empirical methodologies is a strength of the HCI research community, and one of the most powerful of the methodologies is the experimental method as affirmed in (GRAY; SALZMAN, 1998) . We believe that "the sole purpose of experiments is to provide stronger tests of causal hypotheses than other forms of research allow" . your review has uncovered no inherent obstacles to the application of the experimental method to HCI topics. Interest in interface design has been a persistent theme of HCI; interest in the design of experiments has not. The experimental method is a powerful vehicle that can be to address these and other fundamental IHC issues.

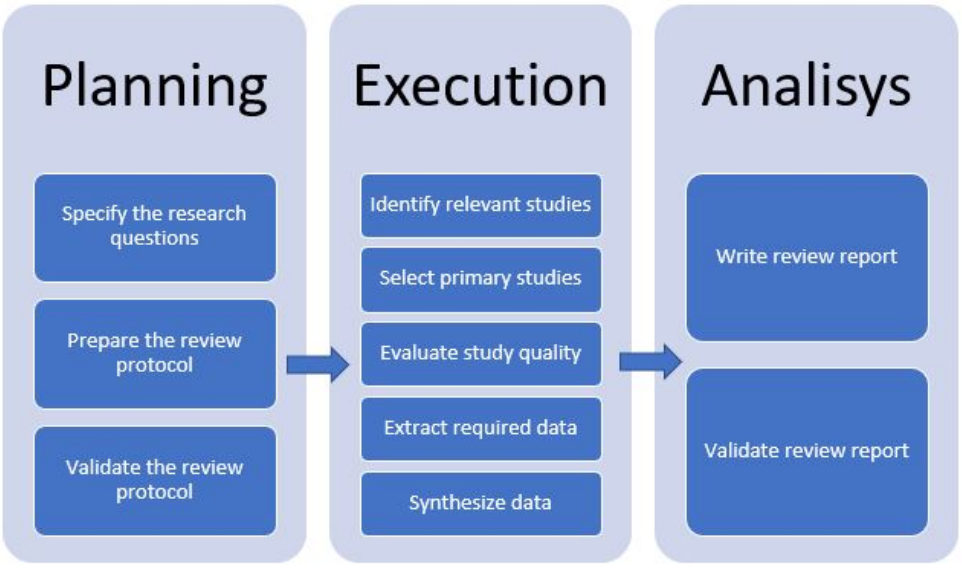
C SYSTEMATIC LITERATURE REVIEW

A systematic literature review is a type of bibliography mapping, an academic type of work that is performed specifically to filter, classify, analyze and extract exhaustively a large number of literature on a specific topic. It is a highly concentrated review work. According to the topic and scope of the content involved, it can be classified into: comprehensive review or thematic review. In the case of our usability thematic review it is only for Web tools. The literature review consists of analyzing and describing what work has been done by predecessors in a given field of research and to what extent it has progressed. Our review follows the stages established according to the . The SLR is divided into three stages, each of which involves further sub-stages, as you can observe in the Figure 5.

In order to map related studies that have been published in the area, a Systematic Mapping of Studies (SMS) was performed. The mapping was based on identifying the different methods used to evaluate web tools and their respective limitations, how they were used, what metrics are considered when evaluating them, and what types of studies have been conducted with them. Various types of threats of this study were considered. There is a risk that some methods were not mapped; some authors did not make clear the methods used.

The Systematic Literature Reviews, search, identify, select, evaluate and integrate evidence on research questions using a clear, reproducible and minimalist methodology. And evaluate previous studies that answer well-organized questions.

Figure 5 – Stages of systematic literature review.



Source: (BUDGEN et al., 2006)

C.0.1 Planning

In the planning step, we performed the following activities to establish the protocol for the systematic literature review: we established the objectives and research question, the search strategy, the selection of primary studies, the quality assessment, the definition of the data extraction strategy, and the selection of the synthesis methods.

C.0.1.1 Research Question

Based on the Research questions (RQ) structure, the detailed research questions are as follows:

RQ1: What are the usability methods used to evaluate web tools?

RQ2: How have these methods been used to evaluate these tools?

RQ3: What are the problems of limitations they encountered using these evaluated methods?

RQ4: What the empirical validation of the Usability Evaluation Methods?

RQ5: What are the usability metrics used?

C.0.1.2 Search string construction

The search string was constructed from the terms Usability evaluation methods for web tool and User experience, and their synonyms (see Table 1). Table 2 presents the search string.

Table 1 – Terms, Synonyms and the Search String

Terms	**Synonyms**
Usability	User Centered Design, User Experience
Evaluation	evaluating, assessment, validation
Web tools	web application, web system,
Search String	(Usability OR "User Centered Design " OR "User Experience ") AND (Evaluation OR "evaluating " OR assessment OR validation) AND ("web tools " OR "web application " OR "web system ")

Source: Author

Table 2 – Digital Libraries and Search Strings

Digital Library	Search String
IEEE	(Usability OR "User Centered Design " OR "User Experience ") AND (Evaluation OR "evaluating " OR assessment OR validation) AND ("web tools " OR "web application " OR "web system ")
SCOPUS	TITLE-ABS-KEY (Usability OR "User Centered Design " OR "User Experience ") AND (Evaluation OR "evaluating " OR assessment OR validation) AND ("web tools " OR "web application " OR "web system ")
SPRINGER LINK	(Usability OR "User Centered Design " OR "User Experience ") AND (Evaluation OR "evaluating " OR assessment OR validation) AND ("web tools " OR "web application " OR "web system")
ACM	(Usability "User Centered Design" "User Experience") AND (Evaluation "evaluating " assessment validation) AND ("web tools" "web application " "web system ")
SCIENCE DIRECT	(Usability OR "User Centered Design " OR "User Experience") AND (Evaluation OR "evaluating" OR assessment OR validation) AND ("web tools " OR "web application" OR "web system")

Source: Author

In the systematic review protocol, the next step to follow is the quality assessment of the selected studies. For this task, we formulate a series of questions by which we filter them. A score is attributed to them, which will be summarized at the completion of the answers and will allow them to be classified. The possible answers for each question are: "*Yes*", when the study has information to answer the Quality Assessment (QA), then the question scores 1 point; "*partial*", when we can use the study only to partially answer the QA, the score, in this case, is 0.5; and "*none*" when the study does not provide any information to answer the question, having a score equal to 0. The QAs and their respective score rules are described next.

Quality Assessment Questions:

QA1. Does the paper present usability evaluation methods to evaluate web tools?

Yes: The study presents a clearly identified evaluation method;

Partial:The study partially presents at least one method, but it's not clearly defined;

None: The study did not present any method.

QA2. Does the paper mention how these methods have been used to evaluate these tools?

Yes: The study explained how the method has been used;

Partial: It does not provide a clear indication of how the method has been used;

None: Did not provide any explanation of how the method has been used.

QA3. Does the paper mention the limitation problems encountered using these evaluated methods?

Yes: The study presented the limitations found using the method(s) presented;

Partial: It mentioned the limitations, but didn't provide a clear explanation;

None: The study didn't mention the limitations.

QA4. Does the paper present the usability metrics used?

Yes: The study presented the metrics that have been considered for the evaluation;

Partial: The study didn't clearly identify the metrics used;

None: The study didn't present metrics.

QA5. Does the paper present an empirical usability evaluation method?

Yes: The study reported an empirical evaluation;

Partial: The study didn't present an empirical usability evaluation method;

None: The study didn't address limitations.

C.0.2 Data Extraction (DE)

The next step is data extraction, which is a very important phase. It determines what data will be extracted from each study to meet the purpose of the SLR. A further detailed analysis of each publication included in the review in terms of the specific elements of the research questions and the purpose of the review, and enter the data in the form. In addition to the extraction elements, the form contains fields for standard information as follows:

DE1. Title;

DE2. Author;

DE3. Year of publication;

DE4. Methods;

DE5. Tools;

DE6. Metrics;

DE7. Limitation of the methods;

DE8. How the methods have been used;

DE9. Types of Evaluation;

C.0.3 Search Strategy

Search conditions are determined, search resources (database, specific journals or conference materials), search terms are formulated; independent reviewers are identified to verify the search results. After searching for studies in the different databases mentioned above, the number of studies remained at 1633. Of that number there were 58 duplicate studies. After excluding the 58 duplicate studies, 1575 studies remained for classification. The number per base is presented in the following (Table 3).

Table 3 – Numbers of the SLR on Scoping

Source	#
IEEE	115
SCOPUS	627
SPRINGER LINK	456
ACM	255
SCIENCE DIRECT	180
Retrieved Studies	1633
Duplicates Removed	58

Source: Author

C.0.4 Selections Criteria

This review was conducted by two reviewers, respectively a post-graduate student and a graduate student. The results of the classification of the studies by each reviewer are presented below. As the studies were classified by two evaluators, the intervention of one reviewer was necessary to balance the studies that were rejected by one evaluator and those that were rejected by the other. one reviewer was necessary to disentangle the studies that were rejected by one evaluator and accepted by the other and vice versa. Conflict resolution was performed by an Expert.

The selections criteria are a set of predefined criteria (e.g., topic, time period, language, etc.) to identify potentially relevant publications. The selection criteria were evaluated on a subset of primary studies. During the screening stage, the Inclusion Criteria (IC)/Exclusion Criteria (EC) were as follows (see Table 4):

Table 4 – Inclusion/Exclusion Criteria

Inclusion Criteria	**Exclusion Criteria**
IC1. The article presents a usability evaluation method for web tools.	EC1. Only complete articles will be accepted.
IC2. Studies reporting usability evaluations in the Web using.	EC2. Articles that present poorly designed experiments.
	EC3. Studies written in languages other than English.
	EC4.Works that are about usability but not about web tools.
	EC5. Papers that are not focused on the Web domain.
	EC6. Papers presenting only recommendations for Web design.
	EC7. Introductory papers for special issues, books and workshops.

Continued on the next page

Inclusion Criteria	Exclusion Criteria
	EC8. Papers presenting only accessibility studies.
	EC9. Duplicate reports of the same study in different sources.
	EC10. Paper not about the usability.

End of the table

Source: Author

C.0.5 Selection of Primary Studies

After the classification phase of the studies by the Inclusions/Exclusions criteria, where only the titles, abstracts and key words are read to filter them, it is the moment to filter the classified studies according to whether they comply with the requirements of the questions to assign them scores. According to the result of the sum of the weight of the questions of each study, it will pass to the next stage if they comply with the minimum weight to validate. The Table 5 presents the classification of the studies by each evaluator.

Table 5 – Quality studies scores

			Evaluator I						Evaluator II					
ID	Years	Studies Reference	Q1	Q2	Q3	Q4	Q5	Score	Q1	Q2	Q3	Q4	Q5	Score
S01	2018	(KAUR; SHARMA, 2018)	Y	N	P	N	P	2.5	Y	N	P	N	Y	2.5
S02	2010	(LEW; OLSINA; ZHANG, 2010)	P	P	P	Y	P	3	P	P	P	Y	P	3
S03	2017	(KUMAR; HASTEER, 2017)	Y	Y	N	Y	N	3	Y	Y	N	Y	N	3
S04	2015	(HARRATI et al., 2015)	Y	Y	N	P	N	2.5	Y	Y	N	P	P	3
S05	2012	(RIVERO; CONTE, 2012)	Y	P	N	Y	N	2.5	Y	P	N	Y	N	2.5
S06	1999	(PAOLINI, 1999)	P	Y	Y	N	N	2.5	P	Y	Y	N	N	2.5
S07	2012	(FERNANDES; CONTE; BONIFÁCIO, 2012)	Y	Y	N	Y	N	3	Y	Y	N	Y	N	3
S08	2016	(CHYNAł; SOBECKI, 2016)	Y	Y	N	Y	Y	4	Y	Y	N	Y	Y	4
S09	2014	(SHAMSUDDIN; SYED-MOHAMAD; SULAIMAN, 2014)	Y	Y	P	Y	N	3.5	Y	Y	P	Y	N	3.5
S10	2010	(OREHOVACKI, 2010)	Y	Y	N	Y	Y	4	Y	Y	N	Y	Y	4
S11	2011	(VARGAS; WEFFERS; ROCHA, 2011)	Y	Y	N	Y	N	3	Y	Y	N	Y	N	3
S12	2014	(BELE et al., 2014)	Y	Y	N	Y	P	3.5	Y	Y	N	Y	P	3.5
S13	2017	(WICHIENNIT et al., 2017)	Y	Y	N	Y	N	3	Y	Y	N	Y	N	3
S14	2014	(DIAS; FORTES; MASIERO, 2014)	Y	P	Y	N	N	2.5	Y	P	Y	N	N	2.5
S15	2006	(ATTERER; WNUK; SCHMIDT, 2006)	Y	Y	N	Y	Y	4	Y	Y	N	Y	Y	4
S16	2020	(MAGYAR; XU; MAHER, 2020)	Y	Y	Y	P	Y	4,5	Y	Y	Y	P	P	4
S17	2013	(OGNJANOVIC; RALLS, 2013)	Y	Y	N	N	P	2.5	Y	Y	Y	N	P	3.5
S18	2020	(RESKI et al., 2020)	Y	Y	Y	Y	Y	5	Y	Y	Y	Y	Y	5
S19	2005	(JOHNSON; MARSHALL, 2005)	Y	P	N	P	P	2.5	Y	P	N	P	P	2.5
S20	2019	(RAMLI et al., 2019)	Y	P	N	Y	N	2.5	Y	P	N	Y	N	2.5
S21	2019	(RIBEIRO et al., 2019)	Y	Y	N	N	Y	3	Y	Y	N	N	Y	3
S22	2010	(SARRAJ; TROYER, 2010)	P	P	N	Y	Y	3	P	P	N	Y	Y	3
S23	2017	(SOUTH et al., 2017a)	Y	Y	Y	P	N	3.5	Y	P	N	Y	Y	3.5
S24	2019	(XEXAKIS; TRUTNEVYTE, 2019)	Y	P	N	P	Y	3	Y	P	N	P	Y	3
S25	2018	(CAYOLA; MACíAS, 2018)	P	Y	Y	Y	Y	4.5	P	Y	Y	Y	Y	4.5
S26	2007	(HVANNBERG; LAW; LÉRUSDÓTTIR, 2007)	Y	Y	Y	Y	Y	5	Y	Y	Y	Y	Y	5

Continued on the next page

			Evaluator I						Evaluator II					
ID	Years	Studies Reference	Q1	Q2	Q3	Q4	Q5	Score	Q1	Q2	Q3	Q4	Q5	Score
S27	2011	(MALIZIA et al., 2011)	Y	P	N	P	P	2.5	P	Y	N	Y	Y	3.5
S28	2014	(BECCHI et al., 2014)	Y	Y	Y	Y	N	4	Y	Y	Y	Y	N	4
S29	2013	(TORRENTE et al., 2013)	Y	Y	N	Y	N	3	Y	Y	N	Y	N	3
S30	2013	(CLEMMENSEN et al., 2013a)	Y	Y	Y	Y	N	4	Y	P	Y	Y	P	4
S31	2015	(CLEMMENSEN et al., 2013b)	Y	Y	N	P	N	2.5	Y	Y	N	P	N	2.5
S32	2014	(SPEICHER; BOTH; GAEDKE, 2014)	Y	P	P	Y	N	3	Y	P	P	Y	N	3
S33	2011	(FIRMENICH; WINCKLER; ROSSI, 2011)	Y	Y	N	Y	N	3	Y	Y	N	Y	N	3
S34	2009	(BOLCHINI; GARZOTTO; SORCE, 2009)	Y	Y	P	P	Y	4	Y	Y	P	P	Y	4
S35	2010	(MALY; MIKOVEC, 2010)	Y	Y	P	N	N	2.5	Y	Y	P	N	N	2.5
S36	2011	(CARTA; PATERNÒ; SANTANA, 2011)	Y	Y	P	P	N	3	Y	Y	P	P	N	3
S37	2007	(BOSENICK et al., 2007)	Y	P	P	N	Y	3	Y	P	P	N	Y	3
S38	2014	(MORI; PATERNÒ; FURCI, 2014)	Y	P	N	Y	Y	3.5	Y	P	N	Y	Y	3.5
S39	2016	(MÄRTIN; RASHID; HERDIN, 2016)	Y	Y	N	Y	Y	4	Y	Y	N	Y	Y	4
S40	2015	(VALENCIA et al., 2015)	P	P	N	Y	Y	3	P	P	N	Y	Y	3
S41	2020	(DONATI; MORI; PATERNÒ, 2020)	Y	P	P	Y	N	3	Y	P	P	Y	N	3
S42	2010	(SCHREPP, 2010)	Y	P	Y	P	N	3	Y	P	Y	P	N	3
S43	2007	(PATERNÒ; PIRUZZA; SANTORO, 2007)	P	Y	N	Y	N	2.5	P	Y	N	Y	N	2.5
S44	2007	(LÓPEZ; FAJARDO; ABASCAL, 2007)	P	P	N	N	Y	2.5	P	P	N	N	Y	2.5
S45	2019	(GARCÍA-PEÑALVO et al., 2019)	Y	Y	P	Y	Y	4.5	Y	Y	Y	P	Y	4.5
S46	2015	(HVANNBERG, 2015)	Y	N	N	Y	P	2.5	Y	N	N	Y	P	2.5
S47	2009	(YEOW, 2009)	Y	Y	P	Y	N	3.5	Y	Y	P	Y	N	3.5
S48	2020	(SALAU et al., 2020)	Y	Y	N	Y	N	3	Y	Y	N	Y	N	3
S49	2020	(KAVVADIAS; DROSATOS; KALDOUDI, 2020)	P	Y	N	Y	N	2.5	P	Y	N	Y	N	2.5
S50	2020	(HARUN; ABDULLAH; GUNARATNAM, 2020)	Y	P	N	Y	Y	3.5	Y	P	N	Y	Y	3.5
S51	2020	(MAGYAR; MAHER; XU, 2020)	Y	P	N	N	Y	2.5	Y	P	N	N	Y	2.5
S52	2020	(VASCONCELOS; BALDOCHI; SANTOS, 2020)	Y	Y	N	Y	Y	4	Y	Y	N	Y	Y	4
S53	2020	(FATTO et al., 2020)	Y	Y	N	Y	N	3	Y	Y	N	Y	N	3
S54	2020	(BUITRAGO-CASTRO et al., 2020)	Y	P	N	Y	N	2.5	Y	P	N	Y	N	2.5

Continued on the next page

			Evaluator I						Evaluator II					
ID	Years	Studies Reference	Q1	Q2	Q3	Q4	Q5	Score	Q1	Q2	Q3	Q4	Q5	Score
S55	2020	(AKAYAMA et al., 2020)	Y	Y	N	Y	N	3	Y	Y	N	Y	N	3
S56	2019	(BILJON; PRETORIUS, 2019)	Y	Y	N	Y	P	3.5	Y	Y	N	Y	P	3.5
S57	2007	(BOLCHINI; GARZOTTO, 2007)	Y	P	N	Y	Y	3.5	Y	P	N	Y	Y	3.5
S58	2007	(GARCÍA et al., 2007)	Y	Y	N	Y	Y	4	Y	Y	N	Y	Y	4
S59	2012	(MASIP et al., 2012)	Y	Y	N	Y	N	3	Y	Y	N	Y	N	3
S60	2012	(KRIEKE et al., 2012)	Y	Y	Y	Y	P	4.5	Y	Y	Y	Y	P	4.5
S61	2012	(VASCONCELOS; JR, 2012)	Y	Y	Y	Y	N	4	Y	Y	Y	Y	N	4
S62	2012	(VASCONCELOS; JR, 2012)	Y	Y	P	Y	N	3.5	Y	Y	P	Y	N	3.5
S63	2010	(FERNANDEZ; ABRAHÃO; INSFRAN, 2010)	Y	Y	P	Y	Y	4.5	Y	Y	P	Y	Y	4.5
S64	2010	(OTAIZA; RUSU; RONCAGLIOLO, 2010)	Y	Y	Y	Y	Y	5	Y	Y	Y	Y	Y	5
S65	2009	(CONTE et al., 2009)	Y	Y	N	Y	N	3	Y	Y	N	Y	N	3
S66	2008	(PANACH et al., 2008)	Y	Y	N	Y	N	3	Y	P	N	Y	P	3
S67	2007	(OLSINA et al., 2007)	P	Y	N	N	P	2.5	P	Y	N	N	P	2.5
S68	2019	(HE et al., 2019)	Y	Y	N	P	Y	3.5	Y	Y	N	P	Y	3.5
S69	2019	(MISTRY; RAJAN, 2019)	P	Y	N	P	Y	3	P	Y	N	P	Y	3
S70	2019	(FIRMENICH et al., 2019)	Y	P	P	Y	P	3.5	Y	P	P	Y	P	3.5
S71	2016	(DEVI; SHARMA, 2016)	P	N	N	Y	Y	2.5	P	N	N	Y	Y	2.5
S72	2015	(QADOUMI; AL-SHURUFAT, 2015)	Y	Y	N	Y	N	3	Y	Y	N	Y	N	3
S73	2015	(RODRÍGUEZ; ACUÑA; JURISTO, 2015)	P	P	P	P	P	2.5	P	P	P	P	P	2.5
S74	2012	(OLSINA et al., 2012)	Y	N	N	Y	P	2.5	Y	N	N	Y	P	2.5
S75	2009	(FERNANDEZ et al., 2009)	Y	P	N	P	P	2.5	Y	P	N	P	P	2.5
S76	2009	(LILIENTHAL, 2009)	Y	P	N	Y	N	2.5	Y	P	N	Y	N	2.5
S77	2006	(MATERA; RIZZO; CARUGHI, 2006)	Y	N	P	P	P	2.5	Y	N	P	P	P	2.
S78	2005	(HORNBAEK; FROKJAER, 2005)	Y	Y	N	P	Y	3.5	Y	Y	N	P	Y	3.5
S79	2013	(OREHOVACKI; HRUSTEK, 2013)	Y	N	N	Y	Y	3	Y	N	N	Y	Y	3
S80	2014	(SCHMIDT-KRAEPELIN; SUNYAEV, 2014)	Y	Y	N	Y	N	3	Y	Y	N	Y	N	3
S81	2019	(ANI; NOPRISSON; ALI, 2019)	P	Y	Y	Y	Y	4.5	P	Y	Y	Y	Y	4.5
S82	2018	(MARENKOV; ROBAL; KALJA, 2018)	P	P	P	P	P	2.5	P	P	P	P	P	2.5

Continued on the next page

			Evaluator I						Evaluator II					
ID	Years	Studies Reference	Q1	Q2	Q3	Q4	Q5	Score	Q1	Q2	Q3	Q4	Q5	Score
S83	2018	(JURCAU; STOICU-TIVADAR, 2018)	Y	Y	N	Y	N	3	Y	Y	N	Y	N	3
S84	2018	(MURILLO; SANG; PAZ, 2018)	Y	Y	N	Y	N	3	Y	Y	N	Y	N	3
S85	2018	(WAKIL; JAWAWI, 2018)	P	P	P	P	P	2.5	P	P	P	P	P	2.5
S86	2018	(TUNÇ; KÜLCÜ, 2018)	Y	P	Y	N	N	2.5	Y	P	Y	N	N	2.5
S87	2018	(LIAPIS; KATSANOS; XENOS, 2018)	Y	Y	P	P	N	3	Y	Y	P	P	N	3
S88	2018	(RYBARCZYK et al., 2018)	P	P	N	P	Y	2.5	P	P	N	P	Y	2.5
S89	2017	(GRIGERA ALEJANDRA GARRIDO, 2017)	Y	Y	N	Y	Y	4	Y	Y	P	Y	Y	4.5
S90	2017	(SOUTH et al., 2017b)	Y	P	Y	Y	N	3.5	Y	P	P	Y	N	3
S91	2017	(SHIGA; TAKAMI, 2017)	Y	Y	Y	Y	Y	5	Y	Y	Y	Y	Y	5
S92	2017	(SWEDBERG; PEUQUET, 2017)	Y	Y	Y	P	P	4	Y	Y	Y	P	P	4
S93	2017	(MURILLO et al., 2017)	Y	Y	N	P	N	2.5	Y	Y	N	P	N	2.5
S94	2017	(PAZ; POW-SANG, 2017)	P	P	N	Y	P	2.5	P	P	N	Y	P	2.5
S95	2017	(LESTARI; AKNURANDA; RAMDANI, 2017)	Y	P	N	Y	N	2.5	Y	P	N	Y	N	2.5
S96	2017	(IBRAHIM et al., 2017)	P	P	P	P	P	2.5	P	P	P	P	P	2.5
S97	2017	(HUSSAIN; MKPOJIOGU, 2017)	Y	Y	N	Y	N	3	Y	Y	N	Y	N	3
S98	2016	(CASTILLA et al., 2016)	Y	Y	N	P	N	2.5	Y	Y	N	P	N	2.5
S99	2016	(FALKOWSKA; SOBECKI; PIETRZAK, 2016)	Y	Y	Y	Y	P	4.5	Y	Y	Y	Y	P	4.5
S100	2015	(PAZ et al., 2015)	P	P	P	P	P	2.5	P	P	P	P	P	2.5
S101	2015	(RIVERO et al., 2015)	Y	Y	P	P	Y	4	Y	Y	P	P	Y	4
S102	2015	(CHYNA; SOBECKI, 2015)	Y	Y	N	N	P	2.5	Y	Y	P	N	P	3
S103	2015	(HUSTAK et al., 2015)	Y	Y	P	N	P	3	Y	Y	P	N	P	3
S104	2015	(SPIELER et al., 2015)	Y	P	N	Y	N	2.5	Y	P	N	Y	N	2.5
S105	2014	(ESTEVEZ SEAN RANKIN; INDRATMO, 2014)	Y	Y	N	Y	N	3	Y	Y	N	Y	N	3
S106	2014	(RIVERO; KALINOWSKI; CONTE, 2014)	Y	P	N	P	P	2.5	Y	P	N	P	P	2.5
S107	2014	(NETO ANDRé P. FREIRE; ABíLIO, 2014)	Y	Y	N	P	N	2.5	Y	Y	N	P	N	2.5
S108	2014	(VENKATESH; ALJAFARI, 2014)	Y	Y	N	P	P	3	Y	Y	N	P	P	3
S109	2014	(ISLAM; TÉTARD, 2014)	Y	Y	Y	Y	Y	5	Y	Y	Y	Y	Y	5
S110	2014	(CHYNAL, 2014)	Y	P	N	Y	N	2.5	Y	P	N	Y	N	2.5

Continued on the next page

			Evaluator I						Evaluator II					
ID	Years	Studies Reference	Q1	Q2	Q3	Q4	Q5	Score	Q1	Q2	Q3	Q4	Q5	Score
S111	2016	(MARENKOV; ROBAL; KALJA, 2016)	Y	Y	P	Y	Y	4.5	Y	Y	P	Y	Y	4.5
S112	2017	(GRIGERAA et al., 2017)	Y	P	P	P	P	3	Y	P	P	P	P	3
S113	2018	(PUUSKA et al., 2018)	Y	Y	N	P	N	2.5	Y	Y	N	P	N	2.5
S114	2011	(FERRACIOLI; OLIVEIRA, 2011)	Y	Y	N	P	N	2.5	Y	Y	N	P	N	2.5
S115	2011	(CONTE; SILVA, 2011)	Y	P	N	Y	N	2.5	Y	P	N	Y	N	2.5
S116	2013	(FERNANDEZ; ABRAHÃO; INSFRAN, 2013)	Y	Y	P	Y	N	3.5	Y	Y	P	Y	N	3.5
S117	2009	(BABU; SINGH, 2009)	Y	Y	N	Y	N	3	Y	Y	N	Y	N	3
S118	2009	(AL-WABIL; AL-KHALIFA, 2009)	Y	Y	P	P	P	3.5	Y	Y	P	P	P	3.5
S119	2011	(ISLAM, 2011)	Y	Y	N	P	P	3	Y	Y	N	P	P	3
S120	2012	(FERNANDES; CONTE; BONIF'CIO, 2012)	Y	Y	N	P	Y	3.5	Y	Y	N	P	Y	3.5
S121	2011	(FERNANDEZ; INSFRAN, 2011)	Y	Y	P	P	Y	4	Y	Y	P	P	Y	4
S122	2013	(RIVERO; CONTE, 2013)	Y	Y	N	Y	Y	4	Y	Y	N	Y	Y	4
S123	2013	(DÍAZ et al., 2013)	Y	Y	N	Y	N	3	Y	Y	N	Y	N	3

End of the table

Source: Author

To use the usability evaluation methods, the evaluators used tools. The following table shows the tools per study(See Table 6).

Table 6 – Support tools of evaluation methods by studies

Tools	**Studies**
Questionnaire	[S01], [S03], [S9], [S10], [S24], [S25], [S28], [S31], [S32], [S34], [S37], [S38], [S41], [S44], [S46], [S47], [S48], [S50], [S56], [S66], [S84], [S85], [S93], [S94], [S97], [S101], [S104], [S105], [S108], [S109], [S79], [S80]
Checklists	[S01], [S10], [S65]
Not described	[S02], [S04], [S06], [S08], [S15], [S26], [S27], [S59], [S60], [S63], [S64], [S86], [S87],[S89], [S100], [S110], [S114], [S115], [S116],[S117], [S118], [S119], [S121], [S123]
crowdsourcing techniques	[S05]
WE-QT technique	[S07]
System Usability Scale (SUS)	[S12], [S18], [S23], [S33], [S45], [S49],[S68], [S81], [S88], [S90], [S91], [S92],[S96], [S103], [S113]
Usability Evaluation (Web DUE)	[S5],[S101],[S106]
Mockup DUE	[S106], [S122]
Data Collect	[S09], [S17], [S34]
Software Usability Measurement Inventory (SUMI)	[S12], [S92]
Web Analysis and Measurement Inventory (WAMMI)	[S92]
WebHint Method	[S11]
NASA Task Load Index (NASA TLX)	[S23], [S90]
Not clearly defined	[S20], [S22], [S29], [S30], [S83]
Prototype Evaluation	[S16], [S51]
Metaphors of human thinking	[S19]
Refactoring technique	[S21]
Unmoderated Online Tests	[S31], [S102]
MiLE+	[S32], [S57]
Controlled experiments	[S21], [S32], [S34]
Eye tracker and FaceReader	[S39]
Goals, Operators, Methods and Selection rules (GOMS)	[S42]
Task Model	[S35]
Web Usability Probe (WUP)	[S36]
Remote Testing	[S40], [S102]
Fuzzy Tsukamoto	[S95]
WebRemUsine	[S43]

Continued on the next page

Tools	Studies
EWEB	[S44]
User Behavior Analysis	[S52]
Screen Recorder	[S53], [S107]
MUSiC (Metrics for Usability Standards in Computing)	[S54]
Web Usability Scale (WUS)	[S55]
Laboratory testing	[S21], [S58], [S109]
USABILICS	[S61], [S62]
WMR and WebQEM	[S67]
Ultimate Dependable and Native Usability System (URANUS)	[S69]
Screen Recorder Scenario recorder((ScRec))	[S70], [S93], [S107]
EMOTIV EPOC	[S99]
Clicktracking	[S102]
voice recorder and webcams	[S107]
Usability Model Based	[S109]
WCAG, MEGANTA, OCAWA	[S111]
USF, Situation Awareness Global Assessment Technique (SAGAT) , Situation Awareness Rating Technique (SART)	[S113]
Guideliner, MAUVE	[S82]

End of the table

Source: Author

C.1 Threats to validity

Internal validity refers to the influence of the research results on the factors manipulated during the conduct of the research. In a systematic review, the discussion of internal validity is often omitted because the results are not influenced by anything other than the article under study. The research returns sometimes include papers of low quality because they answer questions from poorly constructed studies or experiments. To avoid this threat, two reviewers independently evaluated the studies by title and abstract to verify whether the articles really presented research on the proposed topic, in addition to this selection, a quality assessment stage was also carried out, in which studies that did not attain an average of 2.5 points were excluded.

External validity refers to the extent to which the results of a study can be generalized to various situations or conditions. Examples of possible causes that may affect this validity of the research are: the selection of the chain; the selection of the bases, the most representative articles of the area to serve as a baseline, among others. Thus, to mitigate with such threat, we sought to establish a restricted and replicable

research process, for this study we formulated strings, applied to five (5) search bases, being these Scopus; Science direct; Association for Computing Machinery (ACM) Digital Library; Institute of Electrical and Electronics Engineers (IEEE) Xplore and Springerlink. We verified whether the reference studies were contained in at least one of these databases. The papers studied for the review are selected for quality assessment by limiting the source to international journal articles or conference proceedings. If the methods of quality assessment differ, the papers surveyed may differ and the results obtained may differ.

The external validity of the journal results is high for international journals or conferences that have an A rating or higher because they are journals or conferences of major importance in the research field. The surveyed papers were articles dealing with usability evaluation published between 2005 and 2020. Different search years may lead to different papers being studied and different results being obtained. In fact, to increase the external validity of the results, the search period should be increased to 2022 and another review should be performed.

Construct validity refers to the degree of standardization of the operations performed to obtain the results. Given that the collection and selection of the articles studied and the extraction of the information manually, the possibility of omissions cannot be ruled out. In other words this threat refers to the exclusion of powerful and relevant studies that may occur. To reduce this threat, a strategy has been defined for the selection process, in which studies are selected on the basis of exclusion and inclusion criteria, and studies are rated for quality. However, the construct validity of the results is high because the review was performed with great care.

Validity of the conclusions Although we obtain important data, these results cannot be generalized. We cannot completely ignore the possibility that some terms defined in the search strings may have synonyms that we have not identified.

C.2 Results and RQ Answers

RQ1 What are the usability methods used to evaluate web tools?

During this mapping, a total of 18 usability evaluation methods were identified. Among the most used methods of the studies found were: Questionnaire was reported 24 times as the Heuristic Evaluation was used 24 times, Think-Aloud (TA) 11 times and interview 8 times. It is possible to observe that in 24 studies the evaluation methods are not clearly defined and 11 studies do not have methods.

The study [S03] Using split-testing or A/B testing, a comparative study of the data generated can be used, and metrics are analyzed to conclude which version is more usable. The performance of the online tools was also evaluated to determine which tool is more useful. He also reported that surveys consisting of questionnaires among a target audience or performing a series of tasks to a group of users, which are used by a large group of researchers, are effective but time consuming. the authors of study [S32] mentioned

A/B testing is a tool that allows testing by comparing the hedonic and pragmatic quality of products. "Was that Page Pleasant to Use" (WaPPU) is an A/B testing tool that covers the entire process, from tracking interactions to obtaining correlations and learning usability models.

The author of study [S70], mentioned A/B testing, applying refactorings to create alternative solutions without modifying the application server code. A/B testing in the context of an iterative and incremental method of usability improvement makes the method feasible and compatible with an agile development process. Provide tools to assist usability experts in each step of the usability improvement method, so that they can apply the method in parallel and independently of the development cycle. In the studies [S23][S04], A/B testing is generally applied in large organizations to measure the market performance of different solutions with statistical significance, although the cost of A/B testing can be prohibitive for small companies.

In the study [S30], the author reported that the advantages of using an interview is that local interviewers are available. While the authors of the study [S109] used lab-based methods, questionnaires and interviews, but the authors do not explain why they chose. The researchers also used a TA method to observe how participants performed certain tasks and to find any usability problems. In the study [S113], System Usability Scale (SUS) method to evaluate the ease of use of the system implementation, Situation Awareness Rating Technique (SART) is a situational awareness assessment technique to evaluate the subjective level of situational awareness, and Situation Awareness Global Assessment Techniqu (SAGAT) uses queries designed to assess the participant's actual situational awareness.

The author of studies [S38][S40] mentioned With the questionnaire method, a large number of users can be evaluated. [S28] He considered that a questionnaire is optimal for measuring frequency and quality attitudes.In the studies [S09], [S41], [S48], [S56], [S60], [S90], [S97], [S80] and [S105] They used the questionnaire method without giving a detailed explanation of why they used it. In the studies [S13] [S20], the authors used the SUS assessment method, but did not provide any information as they used. In the study [S110] The questionnaire method, where users complete a survey about their experience with the system and also click tracking: tracks and records the user's action while browsing the system (e.g. their clicks) using some dedicated tools and applications. It allows collecting information on how users worked with the systems, where they clicked, which parts they did not notice and from where they entered the system.

In the studies [S12][S17][S23][S49][S88][S90][S68] The authors presented SUS as the method that most quickly and easily collects users' subjective ratings of product usability. They consider their results to be more robust and versatile. The SUS can, at best, provide a measure of usability and ease of learning, but none of the individual SUS elements can really provide a solution on how to fix the interface. To successfully

identify areas that affect the SUS score, participants must perform and evaluate realistic tasks within the system. According to the study [S81] The experiment confirmed that SUS is an easy method to measure usability. However, SUS is not suitable for revealing system deficiencies and should be carried out in conjunction with other evaluations. It also uses SUS method to avoid the threat of learning effect, each subject interacts with one version of each application.According to the study [S90] In addition to SUS, it also uses NASA Task Load Index (NASA TLX) which is a subjective workload assessment tool that measures the workload of application usage. NASA TLX is task oriented, while SUS is system oriented.

As the study [S45] reported, the SUS has become one of the most popular standardized post-study questionnaires. This questionnaire is also one of the fastest to converge on the correct conclusion, which means that the SUS is a good choice if the sample size is limited or if you suspect that it might be challenging to have a significant sample size of testers.

According to the study [S14] Heuristic Evaluation with Usability and Accessibility (HEUA) was conceived as a questionnaire that can be used to evaluate the usability and accessibility of existing web systems, it is considered as a user-friendly registry to be accompanied in the quick version construction.[29][103][58] Heuristic evaluation is to find usability problems in the user interface design in order to correct them in the iterative design process. The heuristics are in fact specified as concrete elements, and it is possible to verify the compliance of the interface with these elements by inspecting the interface. There is no standard guide with reference guidelines or criteria for determining the level of usability of a web site. Proposals for heuristic usability evaluation do not take into account the type of site that is evaluated when the evaluator finds non-compliance with a heuristic.

Conform to the study [S59] Heuristic evaluation is an inspection method that allows usability experts to obtain improvements faster and cheaper than other evaluation methodologies, such as user testing. Therefore, heuristic evaluation also helps to obtain the fault list. [S84] Heuristic evaluation is executed in less time and cost. [45][94][79][4] Heuristic evaluation is an inspection method that requires experts in interactive technologies and usability. It is based on ten heuristics that are applied as the user interface is examined, often while performing a predefined set of tasks. In the heuristic evaluation, an inspector examines the user interface and records any usability problems and labels them according to one or more of the heuristics. If more than one inspector performs the evaluation, the problems are consolidated into a single problem list.

The evaluators of the study [100] emphasize Heuristic Evaluation is defined as the most efficient technique based on a comparison of seven methods. A heuristic evaluation can be performed by a minimum of three specialists. This method does not require a representative number of people. According to Nielsen, a maximum number of five experts

is sufficient to identify most of the usability problems of a user interface. This method can be applied during any phase of the software development process. It is not necessary to release a functional component to perform a usability study. These evaluation procedures can also be performed on prototypes. Conform to the study [S50] The authors used the usability evaluation heuristic, but do not provide any complementary information about the choice.

According to studies [S63] [S116] Heuristic Evaluation (HE) requires a group of evaluators to examine the user interface according to recognized usability principles called heuristics. The Web Usability Evaluation Process (WUEP) extends and adapts the quality evaluation process proposed in ISO 25000, WUEP employs a web usability model that decomposes the concept of usability into sub-characteristic and measurable attributes. WUEP is widely used in industrial domains, as it can also be applied to intermediate artifacts produced during the early stages of the web development process. There is no other method based on the Model-Driven Web Development (MDWD) process with which to compare WUEP. The WUEP is more effective and efficient than HE in the detection of usability problems in artifacts obtained from a model-driven Web development process. The evaluators' perceived satisfaction of using WUEP is different from the evaluators' perceived satisfaction of using HE.

Conform to the studies [S64][S93] Heuristic evaluation is well suited to identify usability problems through inspection for transactional web applications. Usability testing methods that do not allow direct user interaction with the transactional web application are not the best candidates for evaluating transactional web applications. Usability test methods that allow direct interaction with the transactional web application are able to confirm the main usability issues identified during previous heuristic evaluations. If the objective of the usability evaluation methodology is precisely the evaluation of some specific scenarios or functions of the transactional web application, formal experiments would be much more suitable than co-discovery, because of their focus on specific tasks. The usability testing methods, performed after the heuristic evaluation, allowed to confirm the most critical issues.

In the study [S53], the evaluators used the TA method to get ideas on how to evolve the prototype into a new, more usable version. Based on the study [S87], the TA protocol is a qualitative tool used to understand the behavior of users as they interact with a system in the context of a usability evaluation study. The TA protocol was originally developed to help researchers and practitioners in the domain of cognitive psychology to better understand people's mental processes. During a session, participants are required to verbalize their thoughts about their interaction experience while performing tasks on the evaluated system. This method allows evaluators to identify usability issues that need to be resolved in the next version of the system.

The evaluators of the studies [S107][S117] mentioned task-based user evaluation

using a protocol is one of the best methods used for usability evaluation with the visually impaired. To perform it is necessary to consider supporting materials such as: a computer, stylus and finger touch-controlled tablet, webcams, a set of speakers and a SimpleScreenRecorder to record the screen and audio. You do not need a big name participant.

The authors of the studies [S39][S118] affirm the eye-tracking method is simple to use, no manual operator intervention or special configuration servers are required. An instance of the EyeTracker Browser class is created and starts considering the eye tracker connected in the network. The study [S56] also used the Eye tracking Method, but did not explain why he used them. according to the study [S99] Eye tracking is one of the most advanced methods used in usability testing and provides much more information about user behavior than standard user testing.

The study [S44] reports the evaluation is conducted with Experimentation in the WEB (EWEB) which is a tool for automatic empirical evaluation of web browsing, it supports naive evaluators to create experiments containing types of experiments, web logs to be captured, task models and surveys to be performed by experimental participants. It can be used for both laboratory evaluation and remote evaluation in various browsers that require minimal installation on the client computer. It can be used for both laboratory evaluation and remote evaluation in various browsers that require minimal installation on the client computer.

According to the study [S118] Card sorting is a usability evaluation method for examining disorientation problems that suggested that users were not able to understand the structure of the portal, or that they were confused with the labeling of portal sections, and the organization of information within the portal.

Conforms the study [S11] the WebHint method is a method that has three main activities: In the Task Definition the tasks to be analyzed in the evaluation are determined. A task consists of a sequence of actions performed by users on the application interface to achieve a certain goal. In the User Interaction Capture, the interaction of users with the application interface was monitored. All actions performed by users, such as mouse movements, keystrokes, links accessed, pages loaded, etc., were captured. To capture user interaction, a proxy server was placed between users and applications. In the Data Analysis, all the data obtained in the previous stages were analyzed. They were basically composed of two types of data 1 - the sequences of actions representing the expected behavior for the tasks to be analyzed; 2 - the captured user interaction.

The studies [S122][S106] mentioned, the Web Design Usability Evaluation (Web DUE) technique aims to reduce the cost of correcting usability problems by evaluating Web artifacts in early stages of the development process .The Web DUE technique guides inspectors through the evaluation process of mockups by dividing Web pages into Web page zones. The Web DUE allows the identification of more usability problems of paper based

prototypes of Web applications in reasonable time. Therefore they decided to focus on low-fidelity prototypes (or mockups) which are images of what the software would look like, and which can be evaluated before writing the source code. The Web DUE technique allows inspectors to identify usability problems early in the development process by evaluating low-fidelity Web prototypes or mockups. The Mockup DUE tool helps inspectors by allowing them to (a) interact with the mockups as if they were a real application, and (b) use the Web DUE technique to find usability problems.

In the study [S111], they used empirical testing methods to involve the target users in the tests which means that the tests can expose more serious, more recurring and more global problems. The authors comment, to understand user preferences, interviews are a good and effective approach to measure user satisfaction with the system. system. Empirical evaluation methods are more effective in finding workflow problems and inefficient solutions in user interfaces. These methods also find highly used or unused features, helping to track changes in user requirements and views. In the case of user feedback analysis, special organizations may be needed to reach a proper conclusion.

Importantly, there are studies where the authors report that the evaluators used more than one method. As can be seen: [S19] has an Interview and Remote Usability Testing; [S37] has Interview and Questionnaire; [S46] has Interview, Questionnaire and Heuristic Evaluation; [S85] has Interview and Questionnaire; [S24] has Questionnaire and Survey; [S44] has Questionnaire, Survey and Card sorting; [S50] has Questionnaire and Heuristic Evaluation; [S63][116] has Heuristic Evaluation and Web Usability Evaluation Process (WUEP); [S64] [S114] has Heuristic Evaluation and TA; [S89] has Heuristic Evaluation and Controlled Experiment; [S93] has Heuristic Evaluation and User testing; [S112] has Heuristic Evaluation, Remote Usability Testing and User testing; [118] has Heuristic Evaluation, Eyetracking and Card sorting (show the Table 7).

Table 7 – Usability evaluation methods

Methods	**Studies**
Not clearly defined	[S01], [S07], [S09], [S10], [S12], [S20], [S21],[S35], [S36], [S51],[S54], [S57], [S61], [S65],[S66], [S67], [S82], [S86], [S96],[S97],[S101], [S106], [S115], [S120], [S79]
Not described	[S02], [S6], [S8], [S13], [S16], [S42], [S52], [S62],[S69], [S95]
A/B testing	[S03], [S32], [S70], [S98]
Interview	[S04], [S19], [S30], [S37], [S46], [S85],[S109], [S113], [S119]
Questionnaire	[S24], [S18], [S23], [S28], [S31], [S33], [S37],[S38], [S41], [S44],[S45], [S46], [S48], [S49],[S50], [S55], [S81], [S85], [S88], [S90]

Continued on the next page

Methods	**Studies**
Heuristic Evaluation	[S14], [S26], [S27], [S29], [S34], [S46], [S50], [S59], [S60], [S63], [S64], [S68], [S84], [S89],[S93], [S94],[S100], [S103], [S104], [S112],[S114], [S116],[S118], [S123]
Think Aloud	[S18], [S25], [S53], [S64], [S87], [S105], [S107],[S109], [S114], [S117], [S119]
Eyetracking	[S39], [S43], [S56], [S99], [S118]
Survey	[S24], [S44], [S50], [S91], [S108]
Card sorting	[S44], [S118]
Focus groups	[S118], [S80]
Remote Usability Testing	[S11], [S15], [S19], [S112]
Web DUE	[S122]
Web Usability Evaluation Process (WUEP)	[S63], [S116], [S121]
User testing	[S47], [S87], [S93], [S112]
GOMS method	[S83]
Peer Tutoring and Active Intervention	[S17]
Cognitive dimensions	[S22]
Hybrid Method	[S102], [S110]
Controlled Experiment	[S40], [S58], [S64], [S89], [S91], [S111]

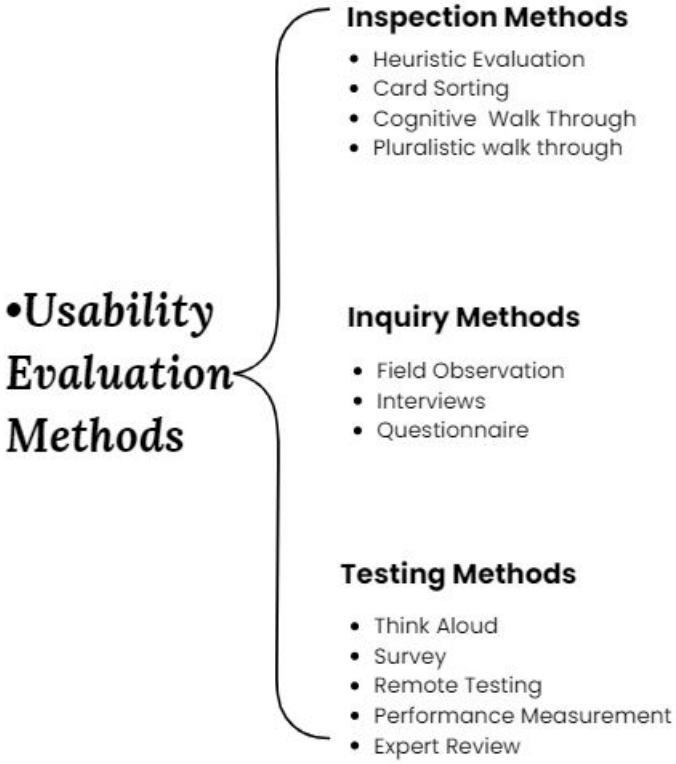
End of the table

Source: Author

Throughout this review we can observe the methods are classified in three categories: inspection, investigation and testing (see Figure 6).

In this review, data were extracted from one hundred and twenty-three (123) studies. Of the total number of studies, thirty-five thirty-five (28.45%) of them did not clearly define the evaluation methods used. Of the other eighty-eight (88) studies a total of one hundred and ten (110) assessments were mapped using eighteen (18) different methods. It is important to mention that few studies contain only one method of assessment. This means that in a large number of the studies more than one evaluation method was found. The methods were distributed as follows: 21.82% Heuristic Evaluation; 18.18% Questionnaire; 10% Think Aloud; 8.18% Interview; Controlled Experience 5.45%; Survey and Eye Tracking 4.55% Each; A/B Testing, Remote Usability Testing and User Testing 3.64% on each; Web Usability Evaluation Process (WUEP) 2.73%; Focus groups, Hybrid Method and Card sorting 1.82% each; GOMS method, Peer Tutoring and Active Intervention and Cognitive dimensions 0.91% for each.

Figure 6 – Method by categories



Source: Author

RQ2 How have these methods been used to evaluate these tools?

In the Study (S) [S04], to evaluate the usability aspect of a given web application, the proposed system consists of three main phases: i) Task Modeling, ii) Usage Tracking, iii) Data Analysis. The experiments are conducted using a newly developed web, in which users are invited to use the system remotely.

Based on the study [S05], to evaluate the usability aspect using A/B testing, they focus on identifying UX problems that users experience, i.e. refactoring opportunities; Repairing UX problems in terms of usability, created by and for the community; and Validation through controlled experiments that will ultimately guide the entire process.

According to the study [S08], To evaluate the usability aspect using eye tracking that allows tracking how the participants' gaze moved through the evaluated application while performing the tasks. An eye-tracking recording is made from this recording, the professionals perform the analysis.

As reported by the reviewers of the study [S15], to evaluate the usability aspect, a proxy needs to be run on the site server. With the appropriate server configuration (for example, using the Apache server's proxy mod module), the proxy can be used to track a user's actions on the site, once that user has agreed to the user test.

Conforming to the study [S16], to evaluate the usability aspect, 10 students were recruited, after filtering for students who were supposed to have taken an online course

before, they randomly selected 12 names, contacted them by email and scheduled 10 participants who were available within the project schedule to participate in the protocol and answer the questionnaire.

Complying with the study [S19], to evaluate the usability aspect, a compilation of evaluation methods was used that produced results that compare favorably with more formal methods.

Based on the evaluators of the study [S25], to evaluate the usability aspect, the STRUM (scheduling tool for recommending usability methods), i.e. loud thinking technique, was used.

In the study [S35], to evaluate the usability aspect, the user was shown a step by step procedure for performing a task and the user's development and time spent was observed.

As indicated in the study [S36], this usability evaluation involved 10 users. All were part of the target user set. The task of adding a book by a particular author to the cart was determined, and the development of the participants was observed.

According to the study [S38], the questionnaire was filled in two different sections during two hours by 57 users. One section was carried out in the presence of the assessors and the other by the users alone, without help. The questionnaire was divided into 3 parts: Personal information, classification of emotions in web interaction, 3 opinions on emotion-based web design. The first part was intended to collect personal information from users and their experiences with the Web. In the second part, users had to propose some emotions they considered relevant during Web interaction and, for each of them, they had to freely associate colors and some attributes characterizing the user's activity. The third part was more oriented to Web design, in which users had to give their opinion by associating each emotion with different characteristics of the Web interface.

In accordance with the study [S39], it is implemented as a client-side application in AngularJS prototype web application, which handles the state of the user experience. AngularJS allows dynamic changes to the page without having to load the whole page again. Through the API, Angular requests data and evaluates it in the directive. A directive is essentially a function that is executed, when the Angular compiler finds it in the Document Object Model (DOM).

Based on the evaluators of the study [S40], to conduct this experiment the following steps are established: 1) Specify the type of experiment; 2) Determine the tasks and the stimuli of the experimental sessions; 3) Define the procedure of the experimental sessions; 4) Specify the interaction data to be collected.

In the study [S41] the evaluators conducted a user test where they evaluated the experience and the of the three types of transitions and to better understand them. They did it in 3 phases during which the users had to execute a task before and after each transition and then they had to fill in a questionnaire. Forty people (23 women and 17

men) with an average age of 36 years were evaluated. All participants were accustomed to surfing the internet.

The authors of the study [S45] report the usability aspects of the site were evaluated in this order: usability was checked for the number of broken links and accessibility using online tools. At the same time, each system was checked for the availability of different language versions of the content. The authors of the study [S52] considered the evaluate their approach, they perform four different experiments. The first one is to make sure that the usability of the tasks can be evaluated as users execute them. It's to demonstrate their efficacy. The second is to demonstrate how to use the usability evaluation service to allow applications to adapt to help users who have difficulty performing tasks. The third is to improve the user experience based on usage patterns. Finally, the fourth experiment is to show that the Real-time Usage Mining (RUM) approach is effective in supporting the development of adaptive web applications.

Conforms the study [S54] the tools to be evaluated are specified, considering their careful potentials. Then a typical scenario is selected in which the most important characteristics are taken into consideration in the evaluation. Then the metrics to be evaluated are defined, for which the tasks to be developed have clear criteria, instructions, and measurement form, in order to be consistent with the established evaluation metrics. The evaluators used an instrument of 26 questions divided into two categories: task execution (4 questions) and level of satisfaction (22 questions). The assessment was subcategorized as follows: task execution, ease of use of the application, information provided, graphical interface and functionality. They chose 43 volunteers, including university students and professionals (teachers) aged between 19 and 35 years (38 percent female, 62 percent male). All volunteers read and signed the informed consent form.

In the study [S62], to validate their usability index calculation methods, they select tasks from different applications and evaluate them with USABILICS. They then run a lab test on the same tasks in order to observe the agreement between the lab results and the usability index. Then they do a validation test to verify the effectiveness of the recommendations. For the UsaTasker method, the evaluator makes a task definition, then the UsaTasker presents the captured events graphically. It is a way to verify if each captured event was recorded correctly. The Evaluator can delete an event in case it is judged irrelevant.

Based on the evaluators of the study [S69], to conduct the evaluation, the first step is to identify the parameters. This phase aims to identify the set of parameters used to evaluate the Web application. Then the calculation of the parameters is performed. In this phase, the set of technical parameters identified in the first phase is calculated.

According to the study [S11] to evaluate with the A/B method, first the metrics are identified, and the site whose usability will be evaluated, and then two different versions of the site were created with the same content but different interfaces. The test

tools were identified and their results were studied to evaluate the scores of the attributes that constituted the metrics. In the study [S32], the evaluation was conducted as an asynchronous user study. The participants were recruited through an internal mailing list of the collaborating company. The semi-structured task was then defined to simulate the common intention of all participants. One of the two interfaces was randomly presented to complete the task. At the end of the task, the usability was assessed by means of a unit questionnaire that displayed WaPPU.

The WebDUE application according to the study [S5], the first step consists in identifying the usability problems by dividing the prototype into zones of the web page. The system state zone that corresponds to the actual state of the user while using the application. The data entry zone that shows the form that the user can use to edit his account data. The navigation zone that shows the navigation links within the application. Then the inspectors proceed to the evaluation of the usability verification elements. The second stage consists in analyzing the data which is a qualitative analysis. In study

As mentioned in the study [S70], the A/B test follows this sequence: 1) the specification and execution of usability tests; 2) the analysis of results, design and assembly of alternative versions; the specification of scenarios and execution of tests for each version of the application; the analysis of results and identification of the best version and the refinement of the application of the best version. In more explicit terms First the expert designs the user test, i.e., the tasks to be exercised, the test scenario and the metrics to be calculated during each test run. Second the expert analyzes the test results to identify usability problems, which can be done using a test analysis tool that displays the results in different graphs and diagrams. Third the user tests each new version of the tasks, dividing the subjects into as many groups as versions, similar to A/B testing. Fourth the UX expert compares the test results of each version with each other and with the results of the first stage to determine the best solution. And finally, the developers receive the specification of the best solution (the winning combination of CSWRs) and implement it on the back-end; that is, the best combination of Client-Side Web Refactorings (CSWR) is coded into the main application. In the next iteration, the main application can be retested for usability with another use case or group of tasks.

The evaluators of the study [S98],report before starting the evaluation, a pilot test was conducted with two 79-year-old participants who had the same characteristics as the target sample. The objective of this evaluation was to adjust the parameters of the eye-tracking software and the elements and duration of each phase of the test. The data from these users were not included in the final sample. The final sample was composed of 34 participants who met the inclusion criteria: being 60 years of age or older, cognitive ability to hold a conversation, auditory, visual and motor skills to interact with the system, and not having undergone cataract surgery. Once the experimental protocol was established, they called the participants by telephone to make an appointment to

perform the experiment. An examiner, who was unaware of the objectives and methods of the study and was blind to the experimental conditions, was in charge of randomization. Seventeen participants were assigned to the A/B sequence and 17 to the B/A sequence. The software used for randomization was the free Random Allocation Software.

Each of these methods has advantages and disadvantages. Therefore, the ideal is to use them appropriately for each phase: Planning phase and Proposal phase, as it is difficult to invest time and money to perform evaluation. In most cases cognitive walkthroughs and user testing are performed with one (1) or two (2) user approaches in order to identify the main problems first. In the project initiation phase, when time and cost can be invested, a cognitive walkthrough is performed, where the amount of testing and user testing is increased, and further investigation is carried out in order not to filter out the main issues. In the phase of formulation of the improvement policies, in case the improvement plan is really problematic, it is revised after finding the problem in user tests or confirmed quantitatively by questionnaire evaluation. Finally, in the product creation phase, when developing the final product a checklist of important issues is created based on the data up to that point, and the final verification is done mostly by heuristics.

RQ3- What are the problems of limitations they encountered using these evaluated methods?

A considerable number of studies report limitations encountered, but a total of 87 of the 123 studies extracted do not report any limitations on the methods used to evaluate their tools.[S02], [S03], [S04], [S05], [S07], [S08], [S09], [S10], [S11], [S12], [S13], [S15], [S17], [S19], [S20], [S21], [S22], [S24], [S25], [S27], [S29], [S34], [S31], [S32], [S33], [S36], [S38], [S39], [S40], [S41], [S42], [S43], [S44], [S45], [S46], [S48], [S49], [S50], [S51], [S52], [S53], [S54], [S55], [S56], [S57], [S58], [S59], [S62], [S65], [S66], [S67], [S71], [S72], [S74], [S75], [S76], [S80], [S94], [S68], [S69], [S83], [S84], [S88], [S89], [S91], [S93], [S95], [S96], [S97], [S98], [S99], [S100], [S102], [S103], [S104], [S105], [S106], [S107], [S108], [S113], [S114], [S115], [S117],[S118], [S120], [S121], [S122], [S123].

In the study [S23], the author reports that SUS does not show an answer, it is purely a classification tool to indicate whether the application in question is usable and not used for diagnostic purposes. There are limitations to this, such as the familiarity that end users have with the Ambulatory Emergency Care (AEC) proforma, which would only be addressed if end users became familiar with the application in question.aboratory setting, the quality of the guidelines and the experience of the evaluator. In the study [S90] SUS is purely a classification tool to indicate whether the application in question is usable, and is not used for diagnostic purposes. [S92] Although SUS stimulate spatio-temporal vision, this vision is limited because it reduces the viewer's field of view by

placing small multiples directly on the map.

According to the study [S30] As this is an interview involving 72 participants from 3 different nationalities, different age categories and with different experiences, the limitations were reported as follows: the age differential related to the level of education could affect the result. Concerning developers and usability professionals, there may be a breach between their personal constructs and their professional knowledge.

The authors of study [S85] define a new Framework for usability evaluation of Model Driven Web (MDWE) methods. The limitations pointed out are related to expert people and web designers that are not available anywhere, and most of the designers worked on a specific method. This is a reason for the evaluators not using their framework. [S109] They do not report limitations directly about the methods used, but about the lack of knowledge on how the consideration of semiotic aspects in user interface design and usability evaluation affects the level of usability of an application. There are very few semiotic theories related to interface design and evaluation in the literature. Few EMU take in consideration semiotic issues in the usability evaluation of web applications.

The study [S18] notes that the inherent limitations of the chosen method and method of data collection must be considered. The nature of tasks in the CSCW environment requires participants to cooperate with each other and present their needs for social processes, which are dynamic in nature and can change themselves.

As can be observed in the study [S37], the remote method is compared with the laboratory method. He affirms that the laboratory method is less effective than the remote method. Laboratory evaluation detects fewer problems, it detects only usability problems generated mainly by the observation of participants in an artificial situation. It is less effective because it provides few investigation gains. Concerning speed, it has less potential to reveal problems in a short time. It is also more difficult to recruit participants.

The evaluators of the study [S14], the author mentions two problems with evaluation methods based on automatic verification only: One is that all check points can be verified automatically, and the other is that guidelines such as Web Content Accessibility Guidelines (WCAG) do not allow the evaluator to differentiate serious problems from trivial ones, regardless of the existence of well-defined priority levels. So The authors of the study [S64] related the following: Heuristics may overlook domain-specific problems; Card Sorting: does not allow evaluation of transactions and navigability scenarios.

As indicated in the study [S26] It is possible that difficulties may occur in finding the correct heuristic to which it refers. That is, the heuristic is not always explicitly guiding the evaluators to discover problems or the evaluators are finding problems for which no heuristic exists in the respective set of heuristics. The levels of knowledge and experience of individual evaluators have observable influences on the results. in the study [S60], the Heuristic evaluation is normally conducted by more than one evaluator (expert) because it is difficult for a unique person to detect all usability problems. Finding

an expert to conduct the evaluation is not an easy task.

According to the study [S63], the design of the assessment could have affected the results due to the selection of the attributes to be assessed during the design phase of the WUEP. The exchange of information could have affected the results, in the case that the experiment extends over several days and it is difficult to know if the participants exchange information with each other.

In the study [S112] Thus, inspection methods are limited in terms of the type of problems that can be encountered in a laboratory environment, the quality of the heuristics, and the expertise of the tester. The limitations of inspection methods have led to the popularity of empirical methods, particularly user testing, which captures and analyzes actual usage data. However, this approach has some limitations. There are some usability issues that require human reasoning, so the automated solution cannot detect them. In addition, the number of usability odors that can be detected is limited to those that users encounter repeatedly.

For the TA method: According to the study [S64] When speaking their mind, users can change their problem-solving behavior; the author of study [S118] considered When a problem of distraction appears with the users, the think-aloud protocol is not able to provide sufficient explanations about it. According to the study [S87] However, one of the most important disadvantages of the RTA method is that valuable segments of information may be lost due to participants' memory problems, as has been confirmed. Furthermore, RTA requires additional time, on top of the user testing session, for both the participant and the facilitator.

Based on the study [S110] one of the disadvantages of the eye tracking method is the immobility of the head during eye tracking, the use of several invasive devices, the relatively high price of commercially available eye trackers, and the difficulty of calibration.

On the study [S121], mentioned WUEP can detect various usability problems of a wide range of types in various artifacts employed during the early stages of a MDWD process.

The controlled experiment is explained in the study [S111], it isn't always possible to evaluate all aspects of the user interface and to increase the coverage of the evaluated functions due to time, cost, and resource constraints. In the case of user feedback analysis, special organizations may be necessary to reach a proper conclusion. Finding the required number of users belonging to a focus group is a problem. Empirical methods are time-consuming and human resource-intensive. It is difficult to analyze and compare results. The authors of the study [S102] highlight that the main assumptions for a hybrid method are the following: it must have the ability to perform complex usability testing much faster than with other methods and the ability to gather all kinds of data related to user interaction with the web-based system under evaluation. In addition, it should be low-cost, with no moderation required, and should allow testing a large group of users at

once.

RQ4- What the empirical validation of the Usability Evaluation Methods?

In this SRL were extracted data from 123 studies, were found basically three types of empirical usability evaluation methods. 47.15% of the studies did not present an empirical evaluation method. Of the 65 studies in which at least one empirical method was found, 70 evaluations were found with empirical methods, 35.71% of them are Case Studies, 55.71% of them are Controlled Experiences and 8.57% of them are Surveys. Some studies have more than one empirical evaluation. That is the case of studies [S24][S92] in which the evaluators conducted a case study and a survey. In studies [S81][S85][S88] there is both a case study and a controlled experience. See Table 8 for a more extended view of them.

Table 8 – Empirical validation of the Usability Evaluation Methods

Empirical Validation of the Usability Evaluation Methods	**Studies**	**Percentage of Participation in the Research**
Case study	[S15], [S16], [S17], [S18], [S19], [21], [S26], [S44], [S46], [S64], [S67], [S68], [S73], [S75], [S77], [S81], [S82], [S85], [S88], [S94], [S96], [S100], [S101], [S118], [S121]	19,53%
Controlled experiment	[S08], [S10], [S12], [S13], [S22], [S24], [S25], [S27], [S34][S39], [S40], [S45], [S50], [S51], [S52], [S55], [S56], [S57], [S58], [S60], [S63], [S69], [S71], [S78], [S79], [S81], [S85], [S88], [S89], [S81], [S92], [S99], [S102], [S109], [S111], [S112], [S119], [S120], [S122]	30,47%
Survey	[S24], [S30], [S38], [S92], [S106], [S108]	4,69%
N/A	[S01], [S02], [S03], [S04], [S05], [S06], [S07], [S09], [S11], [S14], [S20], [S23], [S28], [S29], [S31], [S32], [S33], [S35], [S36], [S37], [S41], [S42], [S43], [S47], [S48], [S49], [S53], [S54], [S59], [S61], [S62], [S65], [S68], [S70], [S72], [S74], [S76], [S80], [S83], [S84], [S86], [S87], [S90], [S93], [S95], [S97], [S98], [S103], [S104], [S105], [S107], [S110], [S113], [S114], [S115], [S116], [S117], [S123]	45,31%

End of the table

Source: Author

RQ5- What are the usability metrics used?

Usability metrics is a list of specific features that the web application must comply with in terms of usability were described in terms of the requirements necessary for the adequate provision of the services for which it was designed (FINSTAD, 2010). Table 9 contains in full detailed form all the studies with their respective metrics used to conduct their evaluations.

Seventeen (17) of these studies do not present or haven't clearly indicated the metrics they have used to evaluate their tools. [S06], [S11],[S19], [S21], [S31],[S34], [S35], [S37], [S38], [S40] [S44], [S47], [S51], [S67], [S83], [S102], [S107].

Very few studies use a unique metric to evaluate their tools. Studies [S13] and [S103] considered Satisfaction as the only metric to conduct their evaluation, but didn't justify their election. Studies [S41], [S46], [S68] used Effectiveness. The Study [S49] that used Efficiency. Study [S19] considered only Mental models as a metric. The Study [S30] considered only Attractiveness. The Study [S11] considered behavior, [S24] considered Understanding - tested and [S45] used only Measure compliance.

The studies [S10][S48] [S110] and [S112] mentioned according to the ISO 9241-11 standard that is related to quality and usability evaluation, the evaluation focus is centered on the users. For that it considers these three attributes of measurements: Efficiency, Effectiveness and Satisfaction. They provide the advantages of making a simple and direct measurement of usability or comparison of software products. However, the above mentioned features are not sufficient for a good evaluation. ISO/IEC 9126-1 represents a binary framework for software quality assessment with four characteristics: effectiveness, productivity, security and satisfaction. They allow the development of a software quality assessment model that can bridge the gap between developers and users.

The authors of the studies [S10][S53][S70][S100] define their metrics as the following: Efficiency measures the amount of resources expended during the performance of activities. Effectiveness measures the level at which the activity can be performed accurately and completely. Utility determines the level of user confidence to increase the quality of activity performance. The effort determines the amount of effort to perform an activity. Memorability measures how easy it is to memorize the use of the application and remember the functionalities. The satisfaction is the ability for a tool to meet the expectations of its users. In the study [S95], effectiveness as users attain the objectives set.

Resource efficiency is used in relation to the accuracy and completeness with which users achieve the objectives. The study [S25] mentions that effectiveness is measured by the similarity between the user's actions and the actions performed by the specialist. In study [S94] user satisfaction is usually the result of a successful interaction. In the studies [S95][S96] It's noted that satisfaction is the absence of discomfort and positive attitudes towards the use of the product.

According to study [S96] the effectiveness and efficiency are directly related to the implementation and design of the system; therefore, they are frequently evaluated during usability testing. In the study [S52] The effectiveness reports the task completion, i.e., how many steps of the task have been performed by the user. The study [S25] defines their metrics as the following: Effectiveness: average percentage of tasks successfully completed by users, Efficiency: average time spent (seconds) by users to complete each task. utility, satisfaction, ease of use and ease of learning: average normalized percentage obtained from the USE questionnaire. The study [S32] considered that efficiency is a new approach to guarantee that evaluations are performed with a minimum of effort for both developers and users. He also considered that effectiveness A new approach should be more effective than conversion-based split tests in determining the usability of an interface. the precision is a new approach should provide accurate but easy-to-understand metrics to compete with conversion-based split tests.

It has been reported in studies [S54][S56][S58][S64] and [S65] that Affectivity: The number of tasks performed (Quantity) and the quality with which they have been solved (Quality). Effectiveness: The amount of time a user is able to complete a task. Gives information about the intuitiveness of the tool. The Satisfaction: Measure of the acceptance of the tool, varying on a scale from 0 to 100. Ease of use: Ratio between the user's feeling of learning and the time it takes to complete a task. Memorability: The ease with which the user navigates the tool by remembering the icons versus memorizing them. Errors: The number of errors made during the test.

Studies [S46][S53][S79][S86][S89][S101] and [S122] used effectiveness, efficiency and satisfaction as metrics to conduct their evaluations but did not provide any supplementary explanation about their choice. In study [S62], the efficiency and effectiveness are considered as an index of the usability of a task. The study [S63] used effectiveness, efficiency, perceived ease of use and satisfaction of use and satisfaction with WUEP in comparison to the famous inspection method that is widely used: Heuristic Evaluation (HE). The study [S61] did not make clear the metrics used for the evaluation of HE.

Table 9 – The metrics for each evaluation by studies

Metrics	**Studies**
Not described	[S06],[S11], [S19], [S21],[S31],[S34], [S35], [S37], [S38], [S40] [S44], [S47], [S51], [S67], [S83], [S102], [S107]
Efficacy	[S19]
Efficacy, Efficiency Satisfaction	[S71], [S118]
Effectiveness , Efficiency indicators, satisfaction, learnability number of tasks performed successfully	[S02], [S54]

Continued on the next page

Metrics	Studies
Navigability, Readability, Loading speed, Accessibility, Functional performance	[S03]
Time spent per task, Completion rate duration, Mouse Clicks and Movement, Error	[S04]
Effectiveness, Efficiency indicators, performance	[S07]
Emotion Recognition, Temperature of all participants, Frustration	[S08]
Effectiveness , learnability, Accessibility and Help frequency	[S09]
Effectiveness, Efficiency, satisfaction	[S10], [S46],[S70], [S86], [S89], [S94], [S95],[S96], [S100], [S101], [S110], [S112], [S122], [S79]
Efficiency, Errors, Affect, Helpfulness, Control	[S12]]
Satisfaction	[S13], [S103]
Navigability, Mouse Clicks and Movement, Navigation behaviour metrics, Time-based metrics	[S15]
Mental models	[S19]
Behavior, Time-based metrics, Mental models	[S20]
Effectiveness, Efficiency, satisfaction, learnability	[S22], [S63], [S116], [S71], [S123]
Performance, Frustration, Mental models, physical demand, effort	[S23]
Understanding – tested	[S24]
Effectiveness, Efficiency	[S25], [S11], [S32], [S48], [S52], [S53], [S54], [S56], [S58], [S61], [S62],[S63],[S64], [S65]
Effectiveness, Reliability	[S26]
learnability, Flexibility, minimal memory load, minimal action	[S27]

Continued on the next page

Metrics	Studies
Effectiveness, learnability, Understanding – tested, Operability, Attractiveness,	[S28]
Efficiency, Functional, reliability,	[S29]
Attractiveness	[S30]
satisfaction, number of tasks performed successfully	[S33]
Behavior, Optimal sequence of actions	[S36]
Behavior, Mental models	[S39]
Effectiveness	[S41], [S46], [S68]
Efficiency, performance	[S42]
Errors, Behavior	[S43]
Measure compliance	[S45]
Efficiency	[S49]
Effectiveness, Efficiency, Error, Memorability mental demand	[S54]
The time required for completing each task.	[S55]
Effectiveness, Efficiency indicators, learnability, Cost Effectiveness	[S57]
Flexibility, minimal memory load, Consistency Shortcuts Help Search	[S59]
Efficiency indicators, learnability, Errors, Memorability mental demand	[S61]
Efficiency indicators, learnability, Functional, Understanding – tested, Operability, Attractiveness, reliability	[S71]
Efficiency , the time required for completing each task	[S76]
Navigability, Brevity, Message concision, Labelling significance	[S66]
Efficiency , data delivery and improved server response time	[S69]

Continued on the next page

Metrics	Studies
Effectiveness, Efficiency, satisfaction, Memorability mental demand	[S81]
Efficiency, satisfaction, learnability, Errors, Memorability mental demand	[S82], [S85], [S91], [S117]
Efficiency, Flexibility	[S84]
Effectiveness, Emotion Recognition, Behavior	[S87]
Time spent per task	[S88]
Effectiveness, performance, Frustration, Time-based metrics, Memorability mental demand, demand, Understanding – tested	[90]
Effectiveness, Efficiency, satisfaction, performance	[S92], [S115]
Efficiency, satisfaction, Flexibility, Control	[S93]
Efficiency, satisfaction, learnability, heat map, the time required for completing each task, Precision	[S97]
Efficiency, satisfaction	[S98]
Efficiency	[S99]
Effectiveness, Efficiency, satisfaction, Navigability, Flexibility, Accessibility, Time spent per task, Errors, Time-based metrics, Memorability mental demand, Understanding-tested	[S04]
Effectiveness, Efficiency	[S104]
Effectiveness, Flexibility, Understanding – tested,	[S105]
Effectiveness, Efficiency, learnability	[S106]
Effectiveness, satisfaction, measured reflectively	[S108]

Continued on the next page

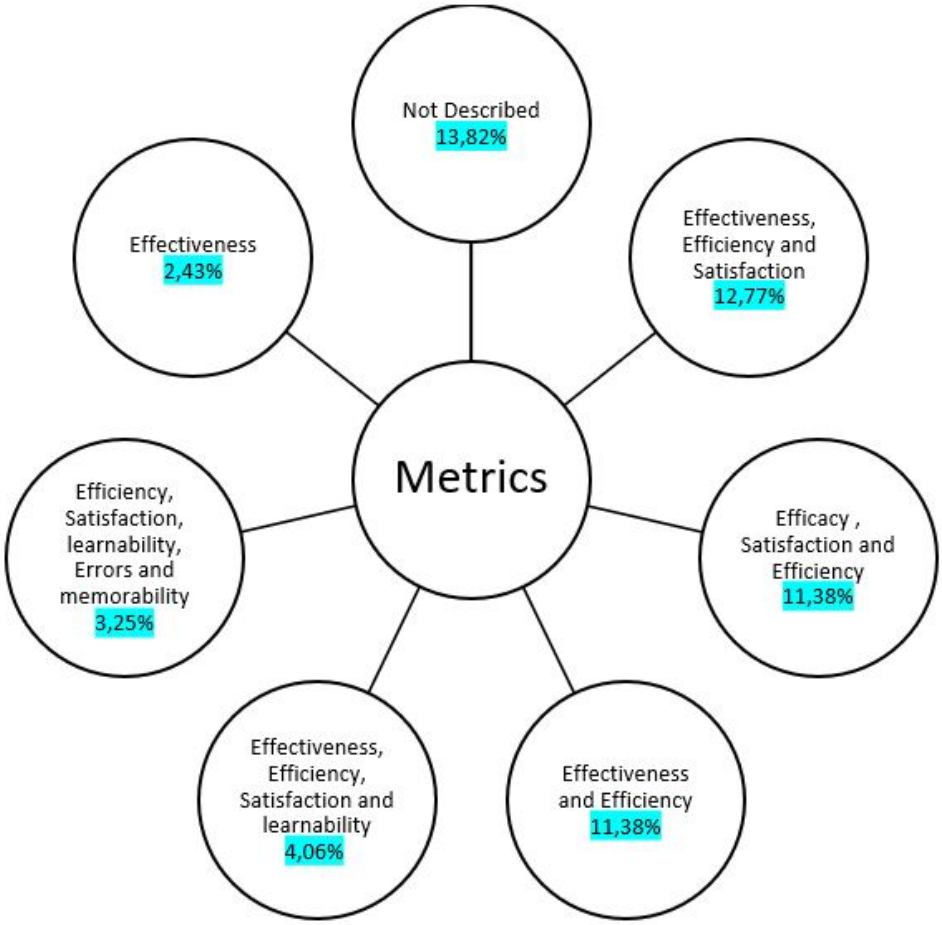
Metrics	**Studies**
Effectiveness, Efficiency , Time-based metrics, Frustration, happiness, users attention, Easiness	[S109], [S119]
Efficiency, satisfaction, learnability, Understanding – tested	[S111]
Efficiency, scalable	[S113]
Effectiveness, satisfaction, performance	[S114]
Effectiveness, Efficiency, the time required for completing each tas	[S120]
Effectiveness, learnability, Easiness	[S121]
Effectiveness, Precision, satisfaction, Efficiency	[S80]

End of the table

Source: Author

In the Figure 7 the most used metrics in this review are presented. As can be seen, of the one hundred and twenty-three (123) studies, 13.82% of them did not clearly indicate the metrics used to conduct their evaluations. 12.77% use these three metrics: Effectiveness, Efficiency and Satisfaction; 11.38% use Efficacy, Satisfaction and Efficiency; 11.38% use Effectiveness and Efficiency; 4.06% use Effectiveness, Efficiency, Satisfaction and learnability; 3.25% use Efficiency, Satisfaction, learnability, Errors and memorability; 2.43% use only the Effectiveness.

Figure 7 – The most used metrics.



Source: Author

C.2.1 Discussion and research contribution

The bibliographic research of methods and tools to evaluate the usability of web tools demonstrated that, in despite of their abundance, there are no consistent techniques for this purpose. This is because there are neither methods nor tools that are clearly superior to the rest, as they all have their advantages and disadvantages. Therefore, they must be selected according to the specific development needs and requirements of each product. Research on this must continue.

The impossibility of finding and accepting a clear and commonly accepted term for user experience from the various scientific fields is reflected in the different methods and dimensions involved in its evaluation. Many aspects of user experience have been defined, formulated, analyzed and used in specific contexts. The concept of user experience and usability is not yet satisfactorily described at a theoretical level, although there are already several definitions that attempt to describe it from many perspectives and authors.

The development of the theoretical basis for user experience will allow the design of evaluation methods that will collect data, which can be interpreted in a theoretical

context, as well as empirical measurements. Improving the involvement of future users in the design process, in-depth evaluation and avoidance of complex evaluation tools for their applications in simple projects will lead to a desirable and positive user experience. Positive user experience combined with quality of experience can be important "tools" in the development of smart products, where the user and the technology are at the center.

D HEURISTIC EVALUATION

D.1 Heuristic Evaluation: Inspection

Heuristic evaluation is a method used to verify the usability of an interface for an application, system, program, etc., and to help identify usability problems with that interface. In other words, heuristic evaluation is a method that systematically tests an interface or a method to assess the quality of a redesign. A heuristic is a recommendation, in a sense, a usability rule that helps to find the answer to a question. It indicates in what you should pay attention, but it does not indicate the exact solution. In general, heuristic evaluation is a very laborious process for a single person, because one can hardly detect all the usability problems in an interface.

During heuristic evaluation, each evaluator performs their work individually. It is very important to follow this procedure in order to obtain independent and unbiased evaluations from each evaluator. During the evaluation session, the expert goes through the interface several times and examines it if necessary. Since heuristic evaluation is about explaining each problem encountered in terms of certain usability principles, this method makes it very easy to make a new design. In addition, many usability problems, once discovered, can be solved quite easily and quickly.

Our heuristic evaluation protocol is composed of three steps. The first one consists of presentation of the system, and the second one consists of conducting a systematic literature review with ten articles that we have left ready for the evaluators. The idea of the review is for the evaluators to explore and get knowledge of the system, to have a better vision of what its purpose is. The last step is the conduct of the evaluation process. In conducting the evaluation, we provide a document containing a brief review of the concept of Neilsen's Heuristics, a table with the legend of the degrees of severity (see Table 10) and a table with the ten heuristics. The heuristic evaluation protocol is available in this link: <https://zenodo.org/record/7013296#.YwG7x3bMLDc>.

The table reserved for the evaluators' observations is composed of 4 columns, respectively one for the heuristics, one to review the objectives of each heuristic placed, one to place the degrees of severity they consider that the violated heuristics deserve and one to describe the violations of each violated heuristic.

Table 10 – Prominent Severity Scales

Degrees of severity	Definition
0	I don't agree that this is a usability problem at all
1	Cosmetic problem only: need not be fixed unless extra time is available
2	Minor usability problem: fixing this should be given low priority
3	Major usability problem: important to fix, so should be given high priority
4	Usability catastrophe: imperative to fix this before product can be released

Source: (NIELSEN, 1992)

This evaluation was conducted by six (6) evaluators from the Human-Computer Interaction (HCI) area, with different degrees of experience: a professor, a graduate student, a master, and two students of incomplete graduation. Before this evaluation was sent by mail to the evaluators, a pilot test was performed by a postgraduate student to check that it was correct and understandable.

D.2 Context of the inspection

To evaluate the heuristics of the Thoth tool, we performed an online inspection. The purpose of this evaluation was to identify the maximum possible violations of Heuristics. We invited people (professors as well as students) knowing in the area from the university. We invited a member of a study group on a social network (Twitter), where we recruited two reviewers who are people working in the area.

We sent nine invitations by mail, to people known and referred by professors in the area. Two of them, due to personal engagements, could not collaborate, but one of them checked the protocol and made recommendations to improve the comprehensibility and visibility of the protocol. He recommended that the protocol be divided in three mini-documents to avoid the collaborators to abandon it, because it was too long and too much information in one single document. Four people did not reply, and three of those invited agreed to collaborate. Through a call that we opened on Twitter, we were able to find three more participants, one of whom withdrew while the other two collaborated.

Using Google Drive, we created a folder for each evaluator with their names. Inside the folder, there were three documents and another folder that contained two documents with ".txt" extensions with the scripts of the two article search bases that would be used in the review. The three documents found inside this folder were the description of the tool in pdf, the using scenario in pdf, and the Heuristic evaluation, which was an editable document. After creating the cover, the link was sent by mail to each evaluator with the terms of use and consent, attached with a profile mapping of the evaluators.

D.3 Pilot Test

A pilot test helps to prepare for the evaluation and makes the results more reliable. This test allows to approve the task wording, estimate the time of the sessions and, if all goes well, shows additional benchmarks for future work.

We did a first pilot test with a postgraduate student who has already studied the Human-Computer Interaction discipline. According to the results obtained, we concluded that the instructions and tasks were not clear and understandable for the evaluators. We had to make many changes to make them simpler. After the changes we performed another pilot test with another postgraduate. The evaluator was then able to complete the tasks without any doubts. This time everything was understandable. The results of the pilot test are presented in the Table 11.

D.3.1 Pilot Test Result

Table 11 – Pilot Test Result

P	Reported	Severity Levels
H1	On the "Data Extraction" page, when inserting options from a multiple choice list the inserted options do not appear on the interface, thus you have to refresh the page.	3
	It was necessary to change the minimum score for approval, and the articles already evaluated in the quality criteria did not update, since they are within the new minimum score. It is necessary to change some fields so that they update the status to be approved.	4
H2	The language selection field has few options and does not support the insertion of unregistered languages.	2
H4	There is no email confirmation field.	2
	The email entered is not validated.	2
	There are no password security elements.	2
H6	The option to exclude duplicate files is not intuitive.	3
	In the "Study Selection" part there is a menu of icons, without any text or explanation of the action of each button.	3

Continued on the next page

P	Reported	Severity Levels
H7	When you get the wrong password, the e-mail field is deactivated.	3
	The Enter key does not work to confirm the entry of the fields.	3
	In the last tab of planning there is no "next" button to go to the Conducting part, being necessary to go back to the top menu.	2
	In the data extraction part, when opening the pop-up with the data of an article, there could be the option "next" at the end to open the information of the next article, without having to close the pop-up and select another article.	2
	In the data extraction part, in multiple choice lists, when clicking on the option name the checkbox should be selected, which does not occur.	3
	In the data extraction, there is no option to download all the extracted data into a spreadsheet or table.	4
H8	The interface is a little confusing and could better delimit the areas of each field.	1
	The Search String screen is a little confusing, where it is necessary to insert a term, then select the term to insert its synonyms. The most ideal would be to register the term and the synonyms at the same time.	2
H9	The "Invalid email or password" message does not go away, even after setting the password.	3
	Even if the "Invalid email or password" message is closed, it reappears when you change page.	4
H10	The "Overview" screen does not provide help on the elements present, especially in the "Progress of Systematic Review" section which can be confusing for new users.	3
	The fields on the "Planning" tab offer help icons, but the content is empty.	3
	The general system help has no content.	3

End of the table

Source: Author

D.4 Profile

To conduct this heuristic evaluation, six participants were recruited as described in the section on the context of the inspection. One of them was declined because she had never conducted a systematic literature review. The evaluation involved a short review so that participants could explore and familiarize themselves with the purpose of the tool. This participant profile mapping questionnaire had two sections: the first was the term

of consent and the second collected the participants' profiles.

The profile questionnaire was used to identify the participant's experience and other relevant data. It consisted of the following questions: Q1 - What is your level of education (incomplete undergraduate, complete graduate, incomplete master, complete master, incomplete doctorate, complete doctorate); Q2 - What is your occupation (teacher, student, computer industry); Q3 - If you work in the Human-Computer area, what is your position? Q4 - Have you already performed a Heuristic Evaluation (yes, no); Q5 - Have you already performed a systematic literature review (yes, no); Q5 - Have you already used or heard about the Thoth tool (yes, no); Q6 - If you know Thoth, tell us how you discovered it; Q7 - If you have already used the Thoth Tool, how was your experience?

In a distributed way, the answers of each participant of this profile mapping are shown in Table 12.

Table 12 – Participant profiles

P	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
P1	Complete PhD	Teacher	Not Indicated	Yes	Yes	Yes	I just observed my students use it by screen sharing (I didn't get to see the whole process)	Not Indicated
P2	Complete undergraduate degree	Student	Designer UX/UI	No	No	No	I discovered it through participating in Heuristic Analysis, where I got information through Twitter	Not Indicated
P3	Incomplete undergraduate degree	IT Industry	Not Indicated	Yes	Yes	Yes	Not Indicated	Not Indicated
P4	Incomplete undergraduate degree	Student, IT Industry	Not Indicated	Yes	Yes	Yes	I discovered it when I needed a tool to support me in the RSL	I think the usability is very bad
P5	Incomplete PhD	Teacher, Student, and IT Industry	Not Indicated	Yes	Yes	Yes	I am one of their developers	The experience was good
P61	Incomplete undergraduate degree	Student, IT Industry	Not Indicated	Yes	Yes	Yes	I am in academic research	I found some problems with bugs

Source: Author

D.5 The inspection

The inspection was conducted during a period of approximately 5 months, from January 2022 to May 2022. It was initiated after the second pilot test, when we knew that the understanding of the tasks was clear. This inspection was performed online. In the evaluation protocol, the link to the Thoth tool was inserted for the participants to have access to it.

This inspection was basically designed to be conducted in three stages. Each evaluator received a folder with three documents, one for each stage, and all were carefully described. The first was a theoretical description of the system. The second document was the Use Scenario, a Systematic Literature Review (SLR), which, mounted with ten articles, had the purpose of simulating the use and operation of the tool so that the evaluator could explore it. This stage of the tasks was one of the best ways for the evaluator to discover gaps and violations of Thoth. The third document was the Heuristic Evaluation itself.

The document set up to receive the observations of Heuristic (H) violations along with their respective degrees of severity was mainly constructed of tables. There was a first table with the link of the system and the tasks: definition of the scope of the evaluation and screens to be evaluated. It contained another table with the different degrees of severity and a brief explanation of each of them. Finally, there was a table with the ten Neilsen heuristics, a field to place the affected screen, the degree of severity of that violation and a field to place observations.

Once the inspections were completed, the opinions and observations of the evaluators were transcribed. The following section presents the analysis of the observations.

Table 13 – Heuristic 1 - Visibility of system status

P	**Reported**	**Severity Levels**
P1	Absence of description of the texts that appear on the home page, i.e., it is difficult to know what each session is about.	2
	In the data extraction section, the data do not appear after insertion.	2

Continued on the next page

P	Reported	Severity Levels
P2	In the data extraction section, the data do not appear after insertion.	2
	After an error message is shown (e.g. invalid email or password!), even if the user clicks on the X to close it, it appears again.	3
	In the new project registration section, it is not possible to make a copy of the planning. It is not possible to know if the user is a researcher or a reviewer.	4
	In the Data Extraction section at the planning stage, when adding options to a question of the type "Multiple choice list", these options did not appear next to the list of questions. We had to press F5 for the browser to update it.	4
P3	Analyzing the first heuristic, the flow presents a sequence of information that the user needs to perform in the project registration and advance in the form of progress bars, however, we do not see this progress bar at the beginning as something to follow. When we talk about data to enter, the user does not know the next steps based on the information presented. There is a status bar, but it is not presented for a flow sequence, but as an indication of the activities on the platform.	4
P4	The system could improve the interface concerning the progress of the results collection.	1
P5	In a lot of the functionalities it is necessary to reload the page so that the previous action appears for the user, as an example: in the quality evaluation.	3

End of the table

Source: Author

Table 14 – Heuristic 2 - Correspondence between system and real world

P	Reported	Severity Levels
P2	The system is in English, but it is used by people whose first language is not English.	3
P5	As I have knowledge about the RSL, I had no difficulty in understanding the terms.	2

Source: Author

Table 15 – Heuristic 3 - User control and freedom

P	Reported	Severity Levels
P1	In the steps: planning; conducting; reporting and exporting, the system does not give me permission to view the fields of a next step if I have not completed all the requirements of a previous step. I think it is correct to tell me the error and show me the fields I want to view from one step forward.	3
P2	In Conducting » Import Studies, it wasn't possible to import the ACM.txt file (a bug was observed - see e-mail). Unable to proceed from this point with the ACM base.	3
P4	The tool causes some problems when we try to edit the protocol after starting a review.	2
P5	It is possible to go back and correct the errors in many functionalities of the system, but some are difficult, as an example: in the Quality Analysis, if we vote to change one rating for another, the score does not update automatically.	3

Source: Author

Table 16 – Heuristic 4 - Consistency and standards

P	Reported	Severity Levels
P2	"Data Bases" in the planning interface vs. "Database" in the conduction interface.	1
P3	Commands and actions keeping the same effect, it is clear that the colors and buttons have highlights for editing, adding and the neutral buttons above.	1

Source: Author

Table 17 – Heuristic 5 - Error prevention

P	Reported	Severity Levels
P1	The new user registration interface does not have a button to show my password while I'm typing it so I can see if I typed it correctly.	3
	On the user login screen there is no "I forgot my password" button, which, in this case, is what happened because I typed it wrong in the registration part.	3
	In the Planning step, in the search sequence part, though there is text in the box, an error appears saying that there is nothing inputted.	3
	In the planning interface, when typing the deadline, the ideal is to have the division by Bars in the date automatically, besides the possibility for the user to choose the date by a calendar modal.	2
P2	Blank (About) or incomplete (Help) page, accessible from the Home Page (My Projects).	3
	No password confirmation for sign up.	4
	No password recovery possible.	4
	In Planning » Data Bases, the list of databases is not in alphabetical order. This can confuse the user (e.g. I didn't see ACM and so I tried to add it as another database).	2
	In Planning » Search String, when adding an item from the drop down list, it leaves the list, but when deleting it, it does not return to the list.	2
P3	The registration error, there is a warning above error, the registration in the head informed that there was something wrong in the email or password, however even when the error is solved the error message continues, there is a movement of the user in other screens and flows.	2
P4	In some steps of the review, the tool does not check if the user has input invalid data.	1

Continued on the next page

P	Reported	Severity Levels
P5	I didn't identify the functionality of "forgot password", there is no information in case I lose the login access.	4
	In case the user finds a problem or bug in the system, I did not find a support or help contact to recover my data.	4
	I can't extract the data in the tool, the functionality has a problem.	4
	In the Conducting step, when trying to insert the bases (ACM and Scopus) the system identified an error, but did not explain what this error was.	4
	When the tool is with too much information, the system fails to load, it is necessary to give many "F5" to be able to open the tool and continue the review.	4
	The session expires without the user knowing, there is no marking warning that the access time is ending.	4
P6	It was not possible to import the bib from the chosen database, so it was necessary to use a bib and inform that it was from another database.	4

End of the table

Source: Author

Table 18 – Heuristic 6 - Recognition rather than recall

P	Reported	Severity Levels
P2	The system tries to keep the user informed of the history of their actions. A clear example is the Overview section.	0
P3	The icons follow a pattern and do not change the form in the presentation of the interface, thus the presentation of the elements suggests that the user does not need to memorize a lot of information.	0

Source: Author

Table 19 – Heuristic 7 - Flexibility and efficiency of use

P	Reported	Severity Levels
P1	For efficiency and flexibility, I think a place in the system where the user can put some predefined settings and select them quickly when filling out the form would be of good advantage.	1
P2	No shortcut keys are offered for frequent actions, e.g. create criteria, create questions etc.	2
P3	When you return to the overview, there's no indication of the bases of information, only status in progress, which also occurs with the planning, to which the lay user has no indications, information or comments that signal the main phases and screens of service.	3
P4	I believe that beginner users may feel lost due to the large number of features.	1

Source: Author

Table 20 – Heuristic 8 - Aesthetic and minimalist design

P	Reported	Severity Levels
P2	What is the role of the "string improver" link?	1
P5	The "duplicates" icon looks like the "copy" icon, I would never think that the objective was to identify duplicate studies.	3
	In the "Magic Search", when searching for some term, it returns all reviews with that name along with an "eye" symbol, when clicking on the symbol nothing happens.	3
	When clicking on "Help", the first terms (Sign In, Sign Up , Profile) appear without text and without anything to help the user.	3

Source: Author

Table 21 – Heuristic 9 - Help users recognize, diagnose, and recover from errors

P	Reported	Severity Levels
P1	On the new user registration page, persistence of error message as an existing email, even when advancing to other pages.	2
P5	The system identifies errors, but the messages are not specific.	4

Source: Author

Table 22 – Heuristic 10 - Help and documentation

P	Reported	Severity Levels
P1	Though there is a documentation icon in the first part, clicking on it does not show any information	2
P2	In Planning » Quality Assessment, when assigning "General Score Interval", there are predefined values, but they are not valid. The User may want to use these values, but they are equivalent to empty fields. Although there is a help page, it is incomplete.	2
P3	We can't find help or documentation on the help page, which is still under construction.	2
P4	It should have a clearer documentation.	2
P5	I didn't identify manuals or texts that help to use the system, I believe that it is very complicated for a lay person to use all the functionalities of the tool alone.	4

Source: Author

D.6 Qualitative Analysis

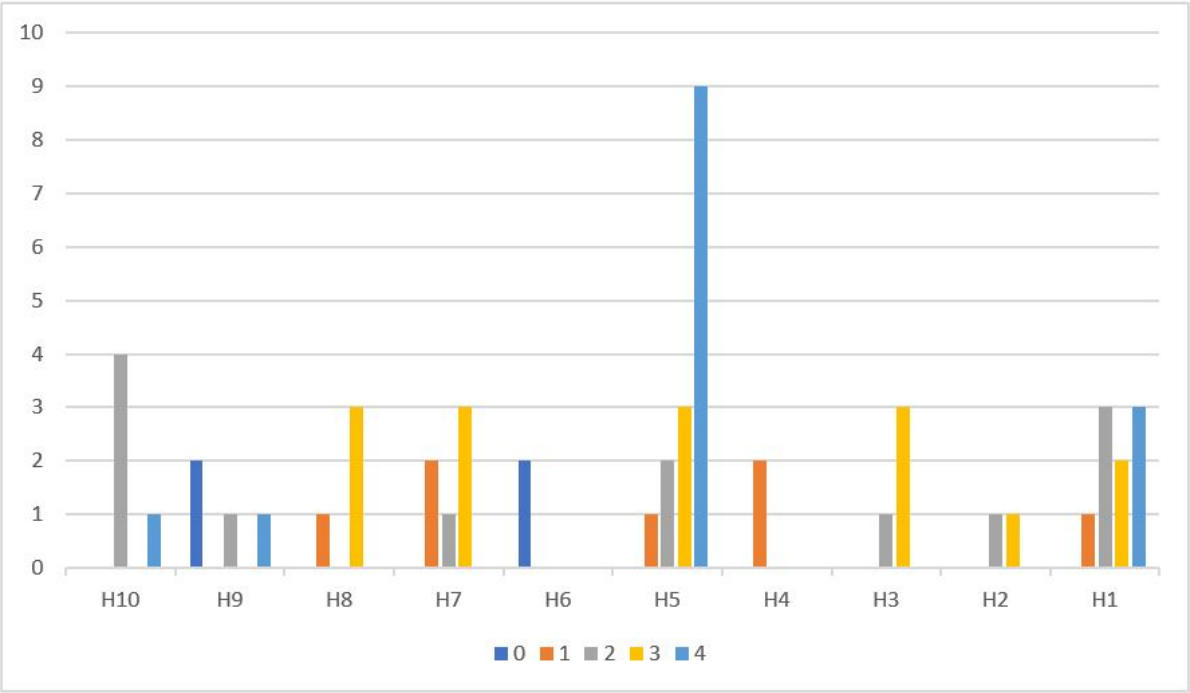
D.6.1 Participant profiles

Six (6) evaluators participated in our inspection. The level of schooling of the participants were: one PhD student; one PhD; one Master; three Master's degree students; and one graduate student. According to their occupations, one was only a student, one was only a teacher, one was only a professional in the IT area, and the others were both students that worked in the IT area. All participants were used to perform heuristic evaluations, and all knew what a systematic review is and how it is used to conduct SLR. Two of them had never used the Thoth tool before.

D.6.2 Heuristic violations reported by the inspectors

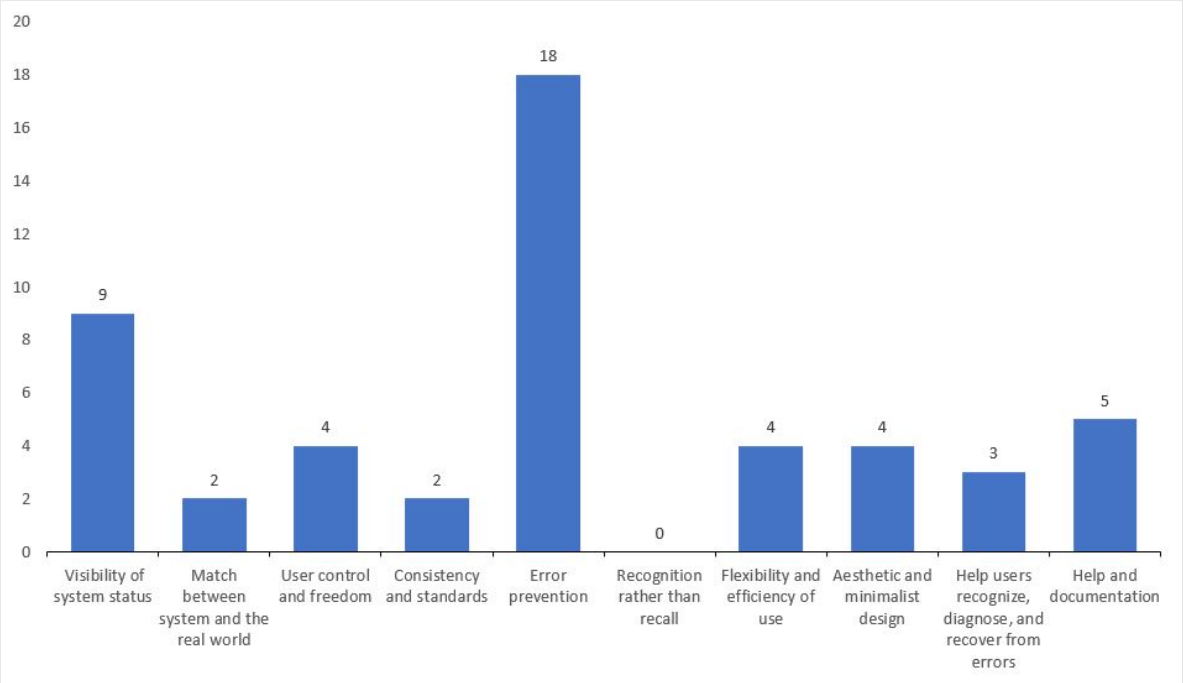
The heuristic evaluation was conducted by six (6) evaluators. Nielsen mentioned that it is impossible for a single person to find all the heuristic violations of a system. Figure 8 presents a graphical view on the degree of severity found for each heuristic inspected from all interfaces combined by their respective value. Figure 9 presents the number of violations found by each heuristic evaluated from different interfaces of the Thoth tool.

Figure 8 – Degree of Severity per Heuristic



Source: Author

Figure 9 – Quantitative Analysis Heuristic



Source: Author

The violated heuristics were classified in order of degree of severity for each evaluator. Among the most pertinent problems that the evaluators considered should be corrected urgently and that severely affected the proper use of the system were the following: participant 2 (P2) felt that it should be possible to make a copy of the planning stage. In the data extraction section at the planning stage, when adding options to a "Multiple Choice List" type question, these options did not appear next to the list of questions. P2 and P5 confirmed that to mitigate that problem the user presses F5 for the browser to update it, an actions that should not be necessary. P2, P4 and P5 classified the system not allowing neither to confirm the password when registering a new user nor to recover it if necessary as a serious fault.

The absence of available support or help to recuperate the data or in other case of need, was classified as degree of urgency, while it was classified as degree two (2) for the participants P1, P2, P3 and P4. The following was classified as grade 4: the impossibility to extract the data in the tool; the fact that when inserting the databases in the conduction stage, an error was generated and the databases appeared with a lot of studies even though they were empty; the system must be refreshed to load the new information entered; the session expired without alerting the user; the errors generated had unidentifiable sources. P6 classified as grade 4 the impossibility encountered to export the Bib from the selected database.

The violations categorized by severity level 3 were the following: P2 related that an error message appeared when the password or e-mail was mistyped during log in, and persisted even after resolving it, while P3 assigned grade 2 for that same problem. He reported that he was unable to import the ACM database. P3 considers that it is a fault of severity 3 to leave the system in English for a public whose first language is not English. P1 qualified it as an error by not being able to visualize the next steps of the planning stage to the conduction. P5, as well as P1, mentioned that the change to be made in the status of a study, for example in quality analysis, did not appear automatically, and the system needed to be reloaded for it to appear. P1 reported that in the planning stage, in the search sequence part, although there was text in the box, an error appeared saying that there was nothing entered.

In the case of P2, the "string improver" function was not defined and the "duplicates" icon looked like the "copy" icon. He assigned a severity grade 3 for both. P5 confirmed that in "Magic Search", searching for some term, e.g., "Heuristic Evaluation", returned all reviews with that name along with an "eye" symbol, and clicking on it did nothing. He also mentioned that when clicking on "Help", the first terms (Sign In, Sign Up, Profile) appeared with no text and nothing to help the user. He placed grade 3 for both of them as well.

The violations that are classified as grade 2 are the following: P1 related the absence of description of the texts that appear on the home page, i.e., it is difficult to

know what each session is about. In the data extraction section, the data does not appear after inserting it, making it mandatory for the user to refresh to make it appear. He also said that in the planning interface, when typing the deadline, it would be ideal to have the division by bars on the date automatically, plus the ability for the user to choose the date via a modal calendar. P4 mentioned that the tool caused some problems when trying to edit the protocol after starting a review. P2 reported in the planning stage that in the search string, when adding an item from the drop-down list, it went out of the list, but when deleting it, it did not go back to the list. He also suggested that the database list be in alphabetical order. P2 and P1 also suggested that the system should offer shortcuts for frequent actions, e.g. create criteria and create questions.

The violations that were reported for grade 1 are the following: P3 suggested that they place effects for commands and actions. He said that it was evident that the colors and buttons had highlights for edit, add and the neutral buttons above. P2 related, regarding standardization, "Data Bases" in the planning interface versus "Database" in the driving interface. P4 mentioned that in some steps of the review, the tool did not check if the user had entered invalid data.

E USABILITY TESTING

E.1 Objective

The developers and designers are not objective users of the products, and many requirements and design solutions are thought out by product designers. Most of the time it is impossible to evaluate with precision the usability of the system, so it is necessary to test with objective users. In this perspective, usability testing through a survey is conducted to collect usability problems in order to improve the tool. The metrics used to evaluate it are Satisfaction, Ease of use, Efficiency, Learn-ability, Effectiveness and Error rate.

E.2 Thoth Evaluation: Survey

For conducting this usability test, we invited some users (researchers and students who are studying the discipline of Academic Research Methodology (MPA)) in order to analyze the following relevant metrics: Ease of Use, Satisfaction, Efficacy and Effectiveness. After performing a task that was submitted, the participants went through the Thoth usability process and provided their ideas about the system.

This test was performed with 25 participants classified by 3 categories, namely a first group of 10 postgraduate researchers who are constant users of the tool, a second group of 4 student users who recently studied the Methodology of Academic Research (MPA), and 9 participants who are currently studying the MPA course. Those nine participants were divided into two groups: 6 of them were software engineering students, more advanced with their revisions, so they did not need a specific task to perform the usability test; and 3 were agricultural engineering students, who were a bit behind with their revisions, so to mitigate this situation, we had to prepare a task for them in order to perform the test. The usability test task is found in this link: <https://zenodo.org/record/7013311#.YwHDxXbMLDc>. Two pilot tests were necessary, one for the participants who did not require homework to conduct their test and one for those who were going to conduct their test after executing their homework.

E.3 Planning

The document we provided to the participants for the test contains three (3) essential parts: Informed Consent Term (ICT), Profile Questionnaire and the test questionnaire. The ICT was presented only to the participants who agreed to answer the profile questionnaire. Instructions on the steps were mailed with details of survey availability, estimated survey duration, information on the number of questions and information that participation would be anonymous for users outside the MPA discipline, and for those

in the discipline, the instructions were made available on the course platform. In order to obtain information about the background of the participants, the following questions were asked: Q1 - What is your education level; Q2 - What is your occupation status; Q3 - If your answer was other, what is your occupation position; and Q4 - Have you already heard about the Thoth tool? After completing the profile questionnaire, the participant continued with the Survey Guidelines.

E.4 Pilot Study

In the perspective of making sure everything was well structured so that the test was successfully executed, two pilot tests were executed. The first one was performed by a graduate student who had already used the tool and was intended for users who were familiar with Thoth. Everything went as planned, without any obstacles. The second pilot test was performed by a user who had no experience with the Thoth tool and was conducting his first systematic review through the assigned task. This test was destined for the 3 three students of the Agricultural Engineering career who were studying the MPA discipline. During the testing of the second pilot test, we were able to discover a potential problem with the databases used. After the test we had the obligation to change the databases to reduce a possible obstacle.

For this second pilot test, a document was prepared for the participant. Within that document was a task on a systematic literature review. The purpose of this task was for the evaluator to explore the tool while conducting their first review. Users who are conducting their first review are also part of Thoth's target public. After the participant finished conducting the review, they went through the usability test stage. The link to the test was inserted at the end of the document containing the task.

E.5 Result Analysis

Twenty-five (25) participants responded to our survey. Regarding educational level, twelve (12) participants were incomplete graduates, nine (9) were incomplete masters, two (2) were complete masters and two (2) were complete graduates. As for their occupations, there were thirteen (13) students, five (5) IT industry professionals, one (1) back-end programmer, one (1) farmer, one (1) freelancer, one (1) student and Information Technology (IT) industry professional, one (1) public official, one (1) contracted public official, and one (1) teacher. Considering the question about whereas they had ever heard of Thoth, the answers were twenty-four (24) "Yes" and one (1) "No". Finally, the answers to the question "have you ever used the Thoth tool to conduct a systematic literature review" were as follows: twenty-two (22) answered "Yes" and three (3) answered "No".

The Perceived Usefulness (PU) was analyzed from fourteen (14) questions and twenty-one (21) responses. There were fifteen (15) questions, but there was one that was

subjective. To measure the internal cohesion of the responses, we applied Cronbach's alpha. In that sense, α=**0.8557176**, which means that the answers of questions one (1) to fourteen (14) had the same line of tendency, as can be seen in Figure 10.

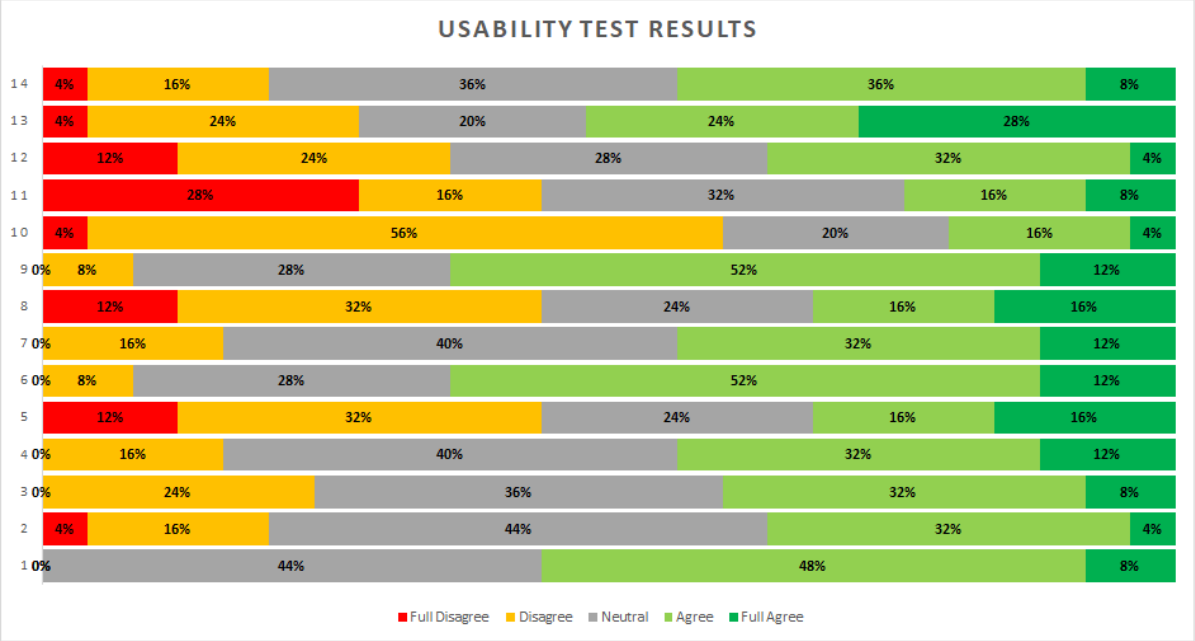
Question fifteen (15) does not appear in the graph because it was an open question, where the participants were free to express and make recommendations according to how they wished to improve performance. The fourteen usability test questions, whose answers are shown in Figure 8, are as follows: Q1- In general, I am satisfied with the use of this system; Q2- The system is simple and easy to use; Q3- I am able to complete my work quickly using this system; Q4- I feel comfortable using this system; Q5- I did not have any difficulties learning how to use this system; Q6- I think I became quickly productive using this system; Q7- The system provides error messages which clearly tell me how to solve the problem; Q8- The information (such as on-line help, on-screen messages, and other documentation) provided with this system is comprehensive; Q9- It is easy to find the information I need in the system; Q10- The information provided by the system is effective in helping me to do my work; Q11- The organization of the information on the system screens is clear; Q12- I like to use the system interface; Q13- The system contemplates all the functionalities and capabilities that it should contemplate; Q14- In general, I am satisfied with this system.

Table 23 – Usability metrics evaluated in the study and their respective survey related questions

Metrics	**Survey question**
Satisfaction	Q1- In general, I am satisfied with this system. Q4- I feel comfortable using this system. Q12- I like to use the system interface. Q13- The system contemplates all the functionalities and capabilities that it should contemplate.
Ease of use	Q2- The system is simple and easy to use. Q9- It is easy to find the information I need in the system.
Efficiency	Q3- I am able to complete my work quickly using this system. Q6- I think I became quickly productive using this system.
Learnability	Q5- I did not have any difficulties learning how to use this system.
Efficacy	Q10- The information provided by the system is effective in helping me to do my work.
Error rate	Q7- The system provides error messages which clearly tell me how to solve the problem. Q8- The information (such as on-line help, on-screen messages, and other documentation) provided with this system is comprehensive.
Memorability	Q11- The organization of the information on the system screens is clear.

Source: Author

Figure 10 – PU Usability Test Results



Source: Author

Analyzing the PU of question 14, we can observe that in terms of satisfaction with the system, four percent (4%) strongly disagreed, sixteen percent (16%) disagreed, thirty-six percent (36%) stayed neutral, thirty-six percent (36%) agreed and eight percent (8%) strongly agreed.

Question thirteen (13) focused on finding out if the system contemplates all the functionalities and capabilities that it should contemplate. The answers were distributed in the following distribution: four percent (4%) of the users completely disagreed, twenty-four percent (24%) disagreed, twenty percent (20%) were neutral, twenty-four percent (24%) agreed and twenty-eight percent (28%) were in complete agreement.

As to whether they liked using the system interface, which was question twelve (12): twelve percent (12%) disagreed completely, twenty-four percent (24%) disagreed, twenty-eight percent (28%) neither agreed nor disagreed, thirty-two percent (32%) expressed agreement and four percent (4%) completely agreed.

Regarding the organization of the information on the system screens, which is the point discussed in the eleventh (11) question, the answers were as follows: twenty-eight percent (28%) of the users totally disagreed that it was adequate, sixteen percent (16%) disagreed, thirty-two percent (32%) neither agreed nor disagreed, sixteen percent (16%) agreed, and only eight percent (8%) totally agreed.

Question ten referred to the effectiveness of the information provided by the system: four percent (4%) totally disagreed, fifty-six percent (56%) disagreed, twenty percent (20%) neither agreed nor disagreed, sixteen percent (16%) agreed and four percent

(4%) totally agreed.

The ninth (9) question deals with the ease of finding needed information in the system. Eight percent (8%) disagreed that it was easy, twenty-eight percent (28%) were neutral, fifty-two percent (52%) agreed and twelve percent (12%) strongly agreed.

Concerning question number eight (8), which was about whether the information (such as on-line help, on-screen messages and other documentation) provided with this system was clear, twelve percent (12%) strongly disagreed, thirty-two percent (32%) disagreed, twenty-four percent (24%) were neutral, sixteen percent (16%) indicated that they agreed, as well as sixteen percent (16%) strongly agreed.

The system provided error messages that clearly indicated how to solve the problem, which was what question seven (7) is about. The answers were as follows: sixteen percent (16%) indicated that they disagreed, forty percent (40%) remained neutral, thirty-two percent (32%) agreed and twelve percent (12%) strongly agreed.

In the sixth (6) question, eight percent (8%) of the participants disagreed that they would be quickly productive by using this system, twenty-eight (28%) neither agreed nor disagreed, fifty-two percent (52%) agreed and twelve percent (12%) were strongly in agreement.

In terms of the ease of learning the system, which was the fifth question, twelve percent (12%) of respondents strongly disagreed, thirty-two percent (32%) disagreed, twenty-four percent (24%) neither agreed nor disagreed, sixteen percent (16%) agreed and sixteen percent (16%) strongly agreed.

For the fourth question, referring to the comfort of using the system, sixteen percent (16%) disagreed, forty percent (40%) remained neutral, thirty-two percent (32%) were in agreement, and twelve percent (12%) were strongly in agreement.

The capacity to complete a task quickly using the system, to which question three (3) concerned, had a twenty-four percent (24%) rate of disagreement, a thirty-six percent (36%) rate of neutrality, thirty-two percent (32%) of agreement, and eight percent (8%) of strong agreement.

The second question dealt with whether the system was simple and easy to use: four percent (4%) of the participants strongly disagreed, sixteen percent (16%) disagreed, forty-four percent (44%) were neutral, thirty-two percent (32%) agreed, and four percent (4%) strongly agreed.

Regarding the first question: forty-four percent (44%) were neutral regarding satisfaction with the use of the system, forty-eight percent (48%) agreed and eight percent (8%) strongly agreed.

Question fifteen (15) was: **Could you give us your opinion so that we can improve the Thoth tool? Feel free to tell us what you think should be changed.** The answers were varied and abundant. In the following paragraphs you will find an analysis of them.

Participant one (P1), participant two (P2), participant three (P3) and participant sixteen (P16) suggested that the system should clearly present the errors made by the users, since the information was not clear and often left the user without knowing how to proceed. They said the system also lacked auxiliary documentation explaining the purpose of the functionalities, leaving the user with no notion of how to get started. They believed that, overall, it was a good tool, as long as you have the help of a colleague or teacher to teach you how to use it.

The recommendations of the participants were the following: improve the search system of the imported studies, since it usually gave an error due to the php timeout (60 seconds), and elaborate a Wiki with the review steps guided by the Thoth tool.

Participant three (P3) said that it was extremely complex to define the weights and quality criteria. He also wanted to have the possibility of exporting all the information per step so that he would able to make a backup copy.

Participant four (P4) had doubts that had not been solved about on-line help, on-screen messages, and other documentation when he used the system for the first time. He needed the help of more experienced users to complete his systematic review, however he thought the system proposal was good, since it reduced the efforts to perform the systematic review, kept a good storage and organization of the articles, plus the visualization of the conclusion in PDF format was good. He suggested the production of a tutorial more oriented to the layman, as Thoth seemed to him as not very intuitive.

Participant five (P5) and participant fifteen (P15) related: "Once you learn how the system works, you can really work productively and justify using the tool instead of spreadsheets. However, there are several improvements that should be made. Some examples are: error reporting is not efficient, database maintenance occurs and RSL data is lost, the quality assessment part of the work should be more intuitive, define a production environment."

According to participant six (P6), the system had bugs to be solved, which made the final development of the system more difficult, but he thought it was a very practical and useful system.

For participant seven (P7) and participant fifteen (P15), the usability issue in the quality review planning and data extraction screens could be very confusing for people without a strong background in systematic reviews mapping.

Participant eight (P8) said that the tool was easy to understand, but data manipulation, some information, and registrations were difficult to interpret, and thus contributed negatively to the learning curve. Some error messages were not clear, as in the bibtext import error in which the correction was based on trial error.

In the opinion of participant nine (P9), the search string generation which was one of the functions of the tool that could be expanded for new databases despite the ones already supported, as well as upgrading the generation of these strings for existing

databases.

Participant ten (P10) stated that he had difficulty starting to use the tool, but managed to understand it very quickly. He also stated that the layout of some functions during planning could be better mapped on the screen.

Participant eleven (P11) affirmed that, in the Planning section, in the Quality Assessment part, at the beginning, it was difficult to understand how the calculation of the General Score worked, as well as the division of the score of the questions. He also complained about the fact that after the pre-filling of the data extraction, the system did not export the data placed there.

For participant twelve (P12), the organization of the forms and the contrast between entries and funds was sometimes a bit confusing. According to him, he was muddled in Planning - Quality Assessment on how to use the scores and its impacts, just like participant one (1).

Participant thirteen (P13) complained about the permanence of the error message that appeared at the top of the system after incorrectly typing the password or email, which remained the same even after correcting it. A point that he found uncomfortable was the issue that the planning submenu, for example, was not at the same height in terms of its buttons, since one was higher than the other, forming a crisscross. Finally, he was forced to re-register simply because the system did not offer the possibility of recovering the password.

Participant number fourteen (P14), he complained about the lack of help from the Help section and the lack of security, because exposing the valid public address of UNIPAMPA to access a web platform makes way for easy attacks.

In the same idea, participant nineteen (P19) stated that there was a lack of information about what should be done in each step by the users of the system. They also thought that the user could be provided with technical information about each stage of planning and execution of systematic studies.

Lastly, participant twenty (P20) said: "When completing some text fields that have a plus button next to them, it could work just by pressing enter, this would save the user from having to click one more time, besides that when completing a information for the first time." He also thought that the information could be better separated, an example of this would be in the planning, in the part of filling out the inclusion and exclusion criteria, there could be different cards.

This participant also reported that, in "Search String", the table has the "synonyms" column, but with each term added, a new table was created inside the table, which made him uncomfortable. These tables within tables were able to pass information, but could be displayed in a better way. In the part of adding the identifier, he thought it would be interesting to fill in an acronym and at each new item added, the system would increment it automatically, an example of this would be in the "Research Question" tab,

it would be interesting to add the id as "RQ", and at each new "RQ" added, the system would add the number automatically. In the tables, the name of each table was at the bottom, and he thought it could be moved to the top of the table. In Conducting, in Study Selection, there was a set of nested buttons, with a gray background, and he said it would be interesting to add a tooltip, to inform what each one of them does.

E.5.1 Threats to Validity

In this section, we analyze the main threats to the validity of our study. Threats are factors that can be internal as well as external that act as obstacles to the reliability of the result (HORNSBY; KURATKO; ZAHRA, 2002). Typical validity problems involve the use of incorrect users or the assignment of incorrect tasks.

Internal validity: To reduce the internal validity of our study, once we decided on the survey method and sampling details, we created a survey form with a limited number of questions and answers for respondents. We put questions that were easy to answer so as not to discourage respondents. We also avoided asking questions that led to a specific answer and that requested respondents to write freely instead of making a choice.

External validity: To mitigate external validity, in order to collect data with a representative sample, we provided within the list of questions an open one where the participants could freely express themselves. Of the twenty-five participants, nineteen (19) of them responded. This was not as representative, but still provided significant data.

Construct validity: To mitigate construct validity, we used the PU for questionnaire organization. We also executed two pilot tests to verify in advance that our questionnaire was adapted and adequate.

Conclusion validity: Although the data obtained are important, we cannot generalize the results by the simple fact that we cannot negate the cases that users have their own point of view.

E.5.2 Discussion

Observing the participants' responses, the following can be deduced: the tool generated a lot of confusion to users at the beginning of use. Due to the lack of documentation to help novice users, it caused fear because they found it complicated. Most of the problems encountered were centered on the planning stage and most of the participants pointed out the bug encountered when loading the databases. One of the most common errors reported by the participants was that the error messages were not clear. They were not able to resolve the errors encountered without the help of an experienced person.

F FINAL CONSIDERATIONS

F.1 Conclusion

The usability has a fundamental role in software development, because it directly impacts the end user. It is understood that there is a large amount of methods and tools to evaluate if the usability aspects are consistent in an application, however, the evaluator does not always know which method is the most suitable for the evaluation of his system and does not even know which tools to use. Aiming to empirically present the advantages of employing methods in usability evaluation, the usability test and heuristic evaluation was applied to Thoth, a support tool for the Systematic Literature Review.

In the usability test, the twenty five evaluators showed that the lack of documentation to guide new users made the system not very efficient. In the same way, the fact that users could not retrieve their passwords was a major obstacle. Error messages did not help users to solve their problems. Likewise, in the heuristic evaluation, five (5) of the six (6) evaluators, demonstrated that the impossibility to retrieve the password was worrying.

It can be concluded that the Thoth tool is a great tool, because it helps in the RSL process, however, the heuristic and usability problems encountered undervalue the great potential that this system may have. A good design analysis paired with screen prototyping and diagrams could have helped in the development process and, of course, in the usability evaluation methods. Usability testing could be used throughout the product development cycle, such as in problem identification; requirements specification; preliminary design; detailed design; and product development. In addition, like the heuristic evaluation, it could also have been applied during product release. As a future work it is desired to use the results collected in the test and heuristic evaluation to fix the problems of the Thoth platform.

F.2 Future Work

This dissertation collects data to provide solutions to usability problems found in the Thoth tool for a long time. During this research, we identified serious problems that obstruct the correct performance of the tool, through a heuristic evaluation and a usability test. In the perspective to provide solutions in an efficient way, the problems found were classified by priority levels and have degrees of urgency. In the same way, through the survey users reported problems that are serious and should be the first to be solved. These are the most urgent:

A large number of the participants, as users, complain about the lack of documentation to help and support users in the real time. In fact, there is no doubt that the

availability of documentation should be one of the first elements to be provided by the tool. To solve this problem, a wiki documentation is in production in order to help users, especially beginners.

One of the most common complaints from users is related to password recovery. The tool does not offer the possibility to recover the password in case the user forgets it. Imagine the case of a user with a revision ready, who is unable to access it because of a forgotten password. This way the tool becomes a waste of time. It is also possible to solve the problem of the password confirmation when creating a new user, while also placing a button so that the user can reveal the password when typing it.

The following problems are in order of priority as reported by users and inspectors. Several participants noted an inconsistency when importing the Bib file from the database. Some report that they were unable to import it because the databases appear with a large number of empty files. There is no doubt that this is very important to be solved as soon as possible, because without the databases it is not possible to progress with the revision.

Users reported that the error messages are not explicit. They did not understand how to solve the problems with the messages provided by the system.

Some inspectors expressed concern that the system did not allow the user to make a copy of the planning phase. This problem can be solved by solving the documentation problem. Because the system, when registering a new project, offers the possibility to copy a planning from another system, if existing. This occurred due to unavailability of information. It is also possible to insert a short description to the buttons or to create a triggering alert to warn the user that a new planning can be imported.

One specialist finds it extremely important that the system is in Portuguese as well, as it is accessible to a public whose first language is not English.

Another problem frequently reported by users is the updating of the data entered. Every time users make a change or enter new information, they have to refresh the browser themselves for those changes to appear.

At the conduction stage, in the data extraction section, the note entry space is not working as it normally should. Every time you write a comment, it appears in all the studies. Also, when you turn off a comment, it turns off in all studies.

In the discussion sections of each evaluation there is a list of violations and usability problems, but here they are so frequent and serious that they all need to be solved as soon as possible.

To keep Thoth up to date, it is recommended to perform regular maintenance as required. For that it is necessary to have a team that gives support to it and to have a complaints and claims section.

BIBLIOGRAPHY

ABELEIN, U.; PAECH, B. Understanding the influence of user participation and involvement on system success a systematic mapping study. **Empirical Software Engineering**, v. 20, n. 1, p. 28–81, 2015. Cited at page 28.

AKAYAMA, N. et al. Salata: A web application for visualizing sensor information in farm fields. In: SCITEPRESS. **9th International Conference on Sensor Networks, SENSORNETS 2020**. [S.l.], 2020. p. 113–120. Cited at page 46.

AL-KILIDAR, H.; COX, K.; KITCHENHAM, B. The use and usefulness of the iso/iec 9126 quality standard. In: IEEE. **2005 International Symposium on Empirical Software Engineering, 2005.** [S.l.], 2005. p. 7–pp. Cited at page 27.

AL-WABIL, A.; AL-KHALIFA, H. A framework for integrating usability evaluations methods: The mawhiba web portal case study. In: IEEE. **2009 International Conference on the Current Trends in Information Technology (CTIT)**. [S.l.], 2009. p. 1–6. Cited at page 48.

ALBERT, B.; TULLIS, T. **Measuring the user experience: collecting, analyzing, and presenting usability metrics**. [S.l.], 2013. Cited at page 30.

ANI, N.; NOPRISSON, H.; ALI, N. M. Measuring usability and purchase intention for online travel booking: A case study. **International Review of Applied Sciences and Engineering**, 2019, v. 10, p. 165–171, 2019. ISSN 20620810. Cited at page 46.

ATTERER, R.; WNUK, M.; SCHMIDT, A. Knowing the user's every move: user activity tracking for website usability evaluation and implicit interaction. In: ASSOCIATION FOR COMPUTING MACHINERY. [S.l.], 2006. p. 203–212. Cited at page 44.

AU, F. et al. Automated usability testing framework. In: CITESEER. [S.l.], 2008. v. 76. Cited at page 30.

BABU, R.; SINGH, R. Evaluation of web accessibility and usability from blind user's perspective: The context of online assessment. In: **AMCIS**. [S.l.: s.n.], 2009. Cited at page 48.

BANGOR, A.; KORTUM, P.; MILLER, J. Determining what individual sus scores mean: Adding an adjective rating scale. **Journal of usability studies**, v. 4, n. 3, p. 114–123, 2009. Cited at page 31.

BARBIERI, L. et al. Mixed prototyping with configurable physical archetype for usability evaluation of product interfaces. **Computers in Industry**, v. 64, n. 3, p. 310–323, 2013. Cited at page 32.

BARGAS-AVILA, J. A.; HORNBÆK, K. Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In: ACM. [S.l.], 2011. p. 2689–2698. Cited at page 32.

BAROLLI, L. et al. A web-based e-learning system for increasing study efficiency by stimulating learner's motivation. **Information Systems Frontiers**, v. 8, n. 4, p. 297–306, 2006. Cited at page 23.

BECCHI, G. et al. A distributed system for multimedia monitoring, publishing and retrieval. **Procedia Computer Science**, 2014, v. 38, p. 100–107, 2014. ISSN 1877-0509. Cited at page 45.

BELE, J. L. et al. ecampus as a platform for ubiquitous learning. In: IEEE. **2014 IEEE Global Engineering Education Conference (EDUCON)**. 2014, 2014. p. 1–7. Cited at page 44.

BEVAN, N. Measuring usability as quality of use. **Software Quality Journal**, v. 4, n. 2, p. 115–130, 1995. Cited at page 23.

BEVAN, N. What is the difference between the purpose of usability and user experience evaluation methods. In: CITESEER. [S.l.], 2009. v. 9, n. 1, p. 1–4. Cited at page 32.

BEVAN, N. et al. New iso standards for usability, usability reports and usability measures. In: SPRINGER. **International conference on human-computer interaction**. [S.l.], 2016. p. 268–278. Cited at page 27.

BEVAN, N.; MACLEOD, M. Usability measurement in context. **Behaviour & information technology**, v. 13, n. 1-2, p. 132–145, 1994. Cited at page 32.

BILJON, J. van; PRETORIUS, M. Usability of a learning management system in interface comparison. **Frontiers in Artificial Intelligence and Applications**, 2019, v. 318, p. 512–521, 2019. ISSN 09226389. Cited at page 46.

BOLCHINI, D.; GARZOTTO, F. Quality of web usability evaluation methods: An empirical study on mile+. **Lecture Notes in Computer Science**, 2007, v. 4832 LNCS, p. 481–492, 2007. ISSN 03029743. Cited at page 46.

BOLCHINI, D.; GARZOTTO, F.; SORCE, F. Does branding need web usability? a value-oriented empirical study. In: SPRINGER. **IFIP Conference on Human-Computer Interaction**. [S.l.], 2009. p. 652–665. Cited at page 45.

BOSENICK, T. et al. Remote usability tests–an extension of the usability toolbox for online-shops. In: SPRINGER. **International Conference on Universal Access in Human-Computer Interaction**. [S.l.], 2007. p. 392–398. Cited at page 45.

BUDGEN, D. et al. Investigating the applicability of the evidence-based paradigm to software engineering. 04 2006. Cited at page 37.

BUITRAGO-CASTRO, L. F. et al. Web application for the teaching of anatomy and physiology of the respiratory system: Usability measurement and metrics. **Communications in Computer and Information Science**, 2020, v. 1274 CCIS, p. 409–419, 2020. ISSN 18650929. Cited at page 45.

CARTA, T.; PATERNÒ, F.; SANTANA, V. F. d. Web usability probe: a tool for supporting remote usability evaluation of web sites. In: SPRINGER. **IFIP Conference on Human-Computer Interaction**. [S.l.], 2011. p. 349–357. Cited at page 45.

CASTILLA, D. et al. Effect of web navigation style in elderly users. **Computers in Human Behavior**, 2016, v. 55, p. 909–920, 2016. ISSN 07475632. Cited at page 47.

CAYOLA, L.; MACíAS, J. A. Systematic guidance on usability methods in user-centered software development. **Information and Software Technology**, 2018, v. 97, p. 163–175, 2018. ISSN 0950-5849. Cited at page 44.

CHYNA, P.; SOBECKI, J. Web-systems remote usability tests and their participant recruitment. **Lecture Notes in Computer Science**, 2015, v. 9169, p. 175–183, 2015. ISSN 03029743. Cited at page 47.

CHYNAL, P. Hybrid approach to web based systems usability evaluation. **Lecture Notes in Computer Science**, 2014, v. 8397 LNAI, p. 384–391, 2014. ISSN 03029743. Cited at page 47.

CHYNAŁ, P.; SOBECKI, J. Application of thermal imaging camera in eye tracking evaluation. In: IEEE. **2016 9th International Conference on Human System Interactions (HSI)**. 2016, 2016. p. 451–457. Cited at page 44.

CLEMMENSEN, T. et al. Do usability professionals think about user experience in the same way as users and developers do? In: SPRINGER. **IFIP Conference on Human-Computer Interaction**. [S.l.], 2013. p. 461–478. Cited at page 45.

CLEMMENSEN, T. et al. Do usability professionals think about user experience in the same way as users and developers do? In: SPRINGER. **IFIP Conference on Human-Computer Interaction**. [S.l.], 2013. p. 461–478. Cited at page 45.

CONTE, T. et al. Improving a web usability inspection technique using qualitative and quantitative data from an observational study. In: IEEE. 2009, 2009. p. 227–235. ISBN 9780769538440. Cited at page 46.

CONTE, T.; SILVA, J. L. M. da. Web usability inspection technique based on design perspectives. In: SBC. 2011, 2011. p. 211–224. Cited at page 48.

DEVI, K.; SHARMA, A. K. Implementation of a framework for website quality evaluation: Himachal pradesh university website. **Indian Journal of Science and Technology**, 2016, v. 9, 2016. ISSN 09746846. Cited at page 46.

DIAS, A. L.; FORTES, R. P. de M.; MASIERO, P. C. Heua: A heuristic evaluation with usability and accessibility requirements to assess web systems. In: ASSOCIATION FOR COMPUTING MACHINERY. 2014, 2014. ISBN 9781450326513. Cited at page 44.

DÍAZ, J. et al. A cultural-oriented usability heuristics proposal. In: ACM. [S.l.], 2013. p. 82–87. Cited at page 48.

DONATI, M.; MORI, G.; PATERNÒ, F. Understanding the transitions between web interfaces designed to stimulate specific emotions. **Universal Access in the Information Society**, v. 19, n. 2, p. 391–407, 2020. Cited at page 45.

DUMAS, J. S.; SALZMAN, M. C. Usability assessment methods. **Reviews of human factors and ergonomics**, v. 2, n. 1, p. 109–140, 2006. Cited at page 30.

DYKSTRA, D. J. **A comparison of heuristic evaluation and usability testing: the efficacy of a domain-specific heuristic checklist**. [S.l.], 1993. Cited at page 29.

ESCALONA, M. J.; KOCH, N. et al. Metamodeling the requirements of web systems. In: SPRINGER. **Proc. International Conference on Web Information System and Technologies (WEBIST 2006), INSTICC**. [S.l.], 2006. p. 310–317. Cited at page 23.

ESTEVEZ SEAN RANKIN, R. S. R.; INDRATMO. A model for web-based course registration systems. **International Journal of Web Information Systems**, 2014, v. 10, p. 51–64, 2014. ISSN 17440084. Cited at page 47.

FALKOWSKA, J.; SOBECKI, J.; PIETRZAK, M. Eye tracking usability testing enhanced with eeg analysis. **Lecture Notes in Computer Science**, 2016, v. 9746, p. 399–411, 2016. ISSN 03029743. Cited at page 47.

FATTO, V. D. et al. Webmgisql 3d - iterating the design process passing through a usability study. In: ACADEMIA. 2020, 2020. p. 69–73. Cited at page 45.

FERNANDES, P.; CONTE, T.; BONIFÁCIO, B. A. We-qt: A web usability inspection technique to support novice inspectors. In: IEEE. [S.l.], 2012. p. 11–20. Cited at page 44.

FERNANDES, P.; CONTE, T.; BONIF'CIO, B. We-qt: A web usability inspection technique to support novice inspectors. In: IEEE. **2012 26th Brazilian Symposium on Software Engineering**. [S.l.], 2012. p. 11–20. Cited at page 48.

FERNANDEZ, A.; ABRAHÃO, S.; INSFRAN, E. Towards to the validation of a usability evaluation method for model-driven web development. In: IEEE. [S.l.], 2010. p. 1–4. Cited at page 46.

FERNANDEZ, A.; ABRAHÃO, S.; INSFRAN, E. A systematic review on the effectiveness of web usability evaluation methods. 2012. Cited at page 32.

FERNANDEZ, A.; ABRAHÃO, S.; INSFRAN, E. Empirical validation of a usability inspection method for model-driven web development. **Journal of Systems and Software**, v. 86, n. 1, p. 161–186, 2013. Cited at page 48.

FERNANDEZ, A. et al. Integrating a usability model into model-driven web development processes. **Lecture Notes in Computer Science**, 2009, v. 5802 LNCS, p. 497–510, 2009. ISSN 03029743. Cited at page 46.

FERNANDEZ, A.; INSFRAN, E.; ABRAHãO, S. Usability evaluation methods for the web: A systematic mapping study. **Information and Software Technology**, v. 53, n. 8, p. 789–817, 2011. ISSN 0950-5849. Advances in functional size measurement and effort estimation - Extended best papers. Cited at page 35.

FERNANDEZ, S. A. A.; INSFRAN, E. A web usability evaluation process for model-driven web development. **Lecture Notes in Computer Science**, 2011, v. 6741 LNCS, p. 108–122, 2011. ISSN 03029743. Cited at page 48.

FERRACIOLI, F.; OLIVEIRA, M. A. de. Using and integrating discount usability engineering in the life cycle of a health care web application. In: CITESEER. 2011, 2011. v. 729. ISSN 16130073. Cited at page 48.

FINSTAD, K. The Usability Metric for User Experience. **Interacting with Computers**, v. 22, n. 5, p. 323–327, 05 2010. Cited 2 times at pages 31 and 66.

FIRMENICH, S. et al. Usability improvement through a/b testing and refactoring. **Software Quality Journal**, 2019, v. 27, p. 203–240, 2019. ISSN 09639314. Cited at page 46.

FIRMENICH, S.; WINCKLER, M.; ROSSI, G. A tool support for web applications adaptation using navigation history. In: SPRINGER. **IFIP Conference on Human-Computer Interaction**. [S.l.], 2011. p. 340–348. Cited at page 45.

FLAVIAN, C.; GUINALíU, M.; GURREA, R. The influence of familiarity and usability on loyalty to online journalistic services: The role of user experience. **Journal of Retailing and Consumer Services**, v. 13, p. 363–375, 09 2006. Cited at page 31.

GARCÍA, F. J. et al. A controlled experiment for measuring the usability of webapps using patterns. In: SPRINGER. **Enterprise Information Systems VII**. [S.l.], 2007. p. 257–264. Cited at page 46.

GARCÍA-PEÑALVO, F. J. et al. Analyzing the usability of the wyred platform with undergraduate students to improve its features. **Universal Access in the information society**, v. 18, n. 3, p. 455–468, 2019. Cited at page 45.

GEORGSSON, M.; STAGGERS, N. Quantifying usability: an evaluation of a diabetes mhealth system on effectiveness, efficiency, and satisfaction metrics with associated user characteristics. **Journal of the American Medical Informatics Association**, v. 23, n. 1, p. 5–11, 2016. Cited at page 23.

GRAY, W. D.; SALZMAN, M. C. Damaged merchandise? a review of experiments that compare usability evaluation methods. **Human–computer interaction**, v. 13, n. 3, p. 203–261, 1998. Cited at page 35.

GRIGERA ALEJANDRA GARRIDO, G. R. J. Kobold: Web usability as a service. In: IEEE. 2017, 2017. p. 990–995. ISBN 9781538626849. Cited at page 47.

GRIGERAA, J. et al. Automatic detection of usability smells in web applications. **International Journal of Human Computer Studies**, 2017, v. 97, p. 129–148, 2017. ISSN 10715819. Cited at page 48.

GRUDIN, J. From tool to partner: The evolution of human-computer interaction. **Synthesis Lectures on Human-Centered Interaction**, v. 10, n. 1, p. i–183, 2017. Cited at page 25.

GUERINO, G. C.; VALENTIM, N. M. C. Usability and user experience evaluation of natural user interfaces: a systematic mapping study. **IET Software**, v. 14, n. 5, p. 451–467, 2020. Cited at page 34.

HAN, S. H. et al. Usability of consumer electronic products. **International Journal of Industrial Ergonomics**, v. 28, n. 3, p. 143–151, 2001. ISSN 0169-8141. 5th Pan-Pacific Conference on Occupational Ergonomics. Cited at page 30.

HARRATI, N. et al. Automating the evaluation of usability remotely for web applications via a model-based approach. In: IEEE. **2015 First International Conference on New Technologies of Information and Communication (NTIC)**. 2015, 2015. p. 1–6. Cited at page 44.

HARUN, Z. M.; ABDULLAH, A.; GUNARATNAM, K. Applying heuristics evaluation to improve the usability of malaysia accounting training management system. In: IOP PUBLISHING. 2020, 2020. v. 769. ISSN 17578981. Cited at page 45.

HASSENZAHL, M. User experience (ux) towards an experiential perspective on product quality. In: ACM. **Proceedings of the 20th Conference on l'Interaction Homme-Machine**. [S.l.], 2008. p. 11–15. Cited at page 24.

HE, X. et al. Aloha: Developing an interactive graph-based visualization for dietary supplement knowledge graph through user-centered design. **BMC Medical Informatics and Decision Making**, 2019, v. 19, 2019. ISSN 14726947. Cited at page 46.

HOLZINGER, A. Usability engineering methods for software developers. **Communications of the ACM**, v. 48, n. 1, p. 71–74, 2005. Cited at page 23.

HORNBAEK, K.; FROKJAER, E. Comparing usability problems and redesign proposals as input to practical systems development. In: CHI '05. **Proceedings of the SIGCHI Conference on Human Factors in Computing Systems**. 2005, 2005. p. 391–400. ISBN 1581139985. Cited at page 46.

HORNSBY, J. S.; KURATKO, D. F.; ZAHRA, S. A. Middle managers' perception of the internal environment for corporate entrepreneurship: assessing a measurement scale. **Journal of business Venturing**, v. 17, n. 3, p. 253–273, 2002. Cited at page 98.

HUSSAIN, A.; KUTAR, M. Usability metric framework for mobile phone application. **PGNet, ISBN**, v. 2099, p. 978–1, 2009. Cited at page 31.

HUSSAIN, A.; MKPOJIOGU, E. Ux evaluation of video streaming application with teenage users. **Journal of Telecommunication, Electronic and Computer Engineering**, 2017, v. 9, p. 129–131, 2017. ISSN 21801843. Cited at page 47.

HUSTAK, T. et al. Principles of usability in human-computer interaction driven by an evaluation framework of user actions. **Lecture Notes in Computer Science**, 2015, v. 9228, p. 51–62, 2015. ISSN 03029743. Cited at page 47.

HVANNBERG, E. T. Identifying and explicating knowledge on method transfer: a sectoral system of innovation approach. 2015, v. 14, 2015. Cited at page 45.

HVANNBERG, E. T.; LAW, E. L.-C.; LÉRUSDÓTTIR, M. K. Heuristic evaluation: Comparing ways of finding and reporting usability problems. **Interacting with computers**, v. 19, n. 2, p. 225–240, 2007. Cited 2 times at pages 30 and 44.

IBRAHIM, N. et al. An evaluation study on dengue-entomological surveillance system using alpha acceptance test. **International Journal on Advanced Science, Engineering and Information Technology**, 2017, v. 7, p. 1574–1850, 2017. ISSN 20885334. Cited at page 47.

ISLAM, M. Beyond users' inaccurate interpretations of web interface signs: a semiotic perception. In: EMERALD. **The IFIP 13th International Conference on Informatics and Semiotics in Organizations (ICISO 2011), Leeuwarden, Netherlands**. [S.l.], 2011. p. 31–40. Cited at page 48.

ISLAM, M. N.; TÉTARD, F. Exploring the impact of interface signs' interpretation accuracy, design, and evaluation on web usability: a semiotics perspective. **Journal of Systems and Information Technology**, 2014. Cited at page 47.

JAZAYERI, M. Some trends in web application development. In: IEEE. **Future of Software Engineering (FOSE '07)**. [S.l.], 2007. p. 199–213. Cited at page 23.

JOHNSON, J.; MARSHALL, C. Convergent usability evaluation: A case study from the eirs project. In: CHI EA '05. **CHI '05 Extended Abstracts on Human Factors in Computing Systems**. 2005, 2005. p. 1501–1504. ISBN 1595930027. Cited at page 44.

JOKELA, T. et al. The standard of user-centered design and the standard definition of usability: analyzing iso 13407 against iso 9241-11. In: ACM. [S.l.], 2003. p. 53–60. Cited at page 32.

JURCAU, D.-A.; STOICU-TIVADAR, V. Evaluating the user experience of a web application for managing electronic health records. **Advances in Intelligent Systems and Computing**, 2018, v. 633, p. 276–289, 2018. ISSN 21945357. Cited at page 47.

KAUR, R.; SHARMA, B. Comparative study for evaluating the usability of web based applications. In: IEEE. **2018 4th International Conference on Computing Sciences (ICCS)**. 2018, 2018. p. 94–97. Cited at page 44.

KAVVADIAS, S.; DROSATOS, G.; KALDOUDI, E. Supporting topic modeling and trends analysis in biomedical literature. **Journal of Biomedical Informatics**, 2020, v. 110, 2020. ISSN 15320464. Cited at page 45.

KERZAZI, N.; LAVALLéE, M. Inquiry on usability of two software process modeling systems using iso/iec 9241. In: IEEE. **2011 24th Canadian Conference on Electrical and Computer Engineering(CCECE)**. [S.l.], 2011. p. 000773–000776. Cited at page 27.

KIENLE, H. M.; DISTANTE, D. Evolution of web systems. In: SPRINGER. **Evolving Software Systems**. [S.l.], 2014. p. 201–228. Cited at page 23.

KRIEKE, L. van der et al. Usability evaluation of a web-based support system for people with a schizophrenia diagnosis. **Journal of medical Internet research**, 2012, v. 14, p. e24, 2012. ISSN 14388871. Cited at page 46.

KUMAR, R.; HASTEER, N. Evaluating usability of a web application: A comparative analysis of open-source tools. In: IEEE. **2017 2nd International Conference on Communication and Electronics Systems (ICCES)**. 2017, 2017. p. 350–354. Cited at page 44.

KUNIAVSKY, M. **Observing the user experience: a practitioner's guide to user research**. [S.l.], 2003. Cited at page 30.

LESTARI, V. A.; AKNURANDA, I.; RAMDANI, F. Usability evaluation of e-government using iso 9241 and fuzzy tsukamoto approach. **Journal of Telecommunication, Electronic and Computer Engineering**, 2017, v. 9, p. 153–157, 2017. ISSN 21801843. Cited at page 47.

LEW, P.; OLSINA, L.; ZHANG, L. Integrating quality, quality in use, actual usability and user experience. In: IEEE. **2010 6th Central and Eastern European Software Engineering Conference (CEE-SECR)**. 2010, 2010. p. 117–123. Cited at page 44.

LIAPIS, A.; KATSANOS, C.; XENOS, M. Don't leave me alone: Retrospective think aloud supported by real-time monitoring of participant's physiology. **Lecture Notes in Computer Science**, 2018, v. 10901 LNCS, p. 148–158, 2018. ISSN 03029743. Cited at page 47.

LILIENTHAL, S. T-prox a user-tracking proxy for usability testing. In: INSTICC. 2009, 2009. p. 225–231. ISBN 9789898111814. Cited at page 46.

LIU, M.; ZHU, Z. A case study of using eye tracking techniques to evaluate the usability of e-learning courses. **International Journal of Learning Technology**, v. 7, n. 2, p. 154–171, 2012. Cited at page 32.

LÓPEZ, J. M.; FAJARDO, I.; ABASCAL, J. Towards remote empirical evaluation of web pages' usability. In: SPRINGER. **International Conference on Human-Computer Interaction**. [S.l.], 2007. p. 594–603. Cited at page 45.

LOWE, D. Web system requirements: an overview. **Requirements Engineering**, v. 8, n. 2, p. 102–113, 2003. Cited at page 23.

MAGYAR, N.; MAHER, M.; XU, X. Creating and evaluating a goal setting prototype for moocs. In: ACM. 2020, 2020. ISBN 9781450368193. Cited at page 45.

MAGYAR, N.; XU, X.; MAHER, M. Creating and evaluating a goal setting prototype for moocs. In: CHI EA '20. **Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems**. 2020, 2020. p. 1–8. ISBN 9781450368193. Cited at page 44.

MALIZIA, A. et al. estorys: A visual storyboard system supporting back-channel communication for emergencies. **Journal of Visual Languages \ Computing**, 2011, v. 22, p. 150–169, 2011. ISSN 1045-926X. Cited at page 45.

MALY, I.; MIKOVEC, Z. Web applications usability testing with task model skeletons. In: SPRINGER. **International Conference on Human-Centred Software Engineering**. [S.l.], 2010. p. 158–165. Cited at page 45.

MANKOFF, J. et al. Heuristic evaluation of ambient displays. In: ACM. **Proceedings of the SIGCHI conference on Human factors in computing systems**. [S.l.], 2003. p. 169–176. Cited at page 29.

MANZARI, L.; TRINIDAD-CHRISTENSEN, J. User-centered design of a web site for library and information science students: Heuristic evaluation and usability testing. **Information technology and libraries**, v. 25, n. 3, p. 163–169, 2006. Cited at page 29.

MARCHEZAN, L. et al. Thoth: A web-based tool to support systematic reviews. In: IEEE. **2019 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)**. [S.l.], 2019. p. 1–6. Cited 3 times at pages 24, 32, and 33.

MARENKOV, J.; ROBAL, T.; KALJA, A. A framework for improving web application user interfaces through immediate evaluation. **Frontiers in Artificial Intelligence and Applications**, 2016, v. 291, p. 283–296, 2016. ISSN 09226389. Cited at page 48.

MARENKOV, J.; ROBAL, T.; KALJA, A. Guideliner a tool to improve web ui development for better usability. In: ACM. **ACM**. 2018, 2018. ISBN 9781450354899. Cited at page 46.

MÄRTIN, C.; RASHID, S.; HERDIN, C. Designing responsive interactive applications by emotion-tracking and pattern-based dynamic user interface adaptation. In: SPRINGER. **International Conference on Human-Computer Interaction**. [S.l.], 2016. p. 28–36. Cited at page 45.

MASIP, L. et al. Towards usability improvement of semantic web applications. In: SCITEPRESS. 2012, 2012. p. 361–366. ISBN 9789898565082. Cited 2 times at pages 30 and 46.

MATERA, M.; RIZZO, F.; CARUGHI, G. T. **Web usability: Principles and evaluation methods**. 2006, 2006. 143-180 p. ISBN 3540281967; 9783540281962. Cited at page 46.

MISTRY, A.; RAJAN, R. A. P. Evaluation of web applications based on ux parameters. **International Journal of Electrical & Computer Engineering (2088-8708)**, v. 9, n. 4, 2019. Cited at page 46.

MORI, G.; PATERNÒ, F.; FURCI, F. Design criteria for web applications adapted to emotions. In: SPRINGER. **International Conference on Web Engineering**. [S.l.], 2014. p. 400–409. Cited at page 45.

MURILLO, B.; SANG, J. P.; PAZ, F. Heuristic evaluation and usability testing as complementary methods: A case study. **Lecture Notes in Computer Science**, 2018, v. 10918 LNCS, p. 470–478, 2018. ISSN 03029743. Cited 2 times at pages 30 and 47.

MURILLO, B. et al. Usability testing as a complement of heuristic evaluation: A case study. **Lecture Notes in Computer Science**, 2017, v. 10288 LNCS, p. 434–444, 2017. ISSN 03029743. Cited at page 47.

NASR, N. et al. The experience of living with stroke and using technology: opportunities to engage and co-design with end users. **Disability and Rehabilitation: Assistive Technology**, v. 11, n. 8, p. 653–660, 2016. Cited at page 24.

NETO ANDRé P. FREIRE, S. S. S. J. M.; ABíLIO, R. S. Usability evaluation of a web system for spatially oriented audio descriptions of images addressed to visually impaired people. **Lecture Notes in Computer Science**, 2014, v. 8514 LNCS, p. 154–165, 2014. ISSN 03029743. Cited at page 47.

NIELSEN, J. Reliability of severity estimates for usability problems found by heuristic evaluation. In: ACM. **Posters and short talks of the 1992 SIGCHI conference on Human factors in computing systems**. [S.l.], 1992. p. 129–130. Cited at page 76.

NIELSEN, J. **Usability engineering**. [S.l.], 1994. Cited 3 times at pages 24, 28, and 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. Cited at page 28.

OGNJANOVIC, S.; RALLS, J. Don't talk to strangers!' peer tutoring versus active intervention methodologies in interviewing children. In: CHI EA '13. **CHI '13 Extended Abstracts on Human Factors in Computing Systems**. 2013, 2013. p. 2337–2340. ISBN 9781450319522. Cited at page 44.

OLSINA, L. et al. Updating quality models for evaluating new generation web applications. **Journal of Web Engineering**, 2012, v. 11, p. 209–246, 2012. ISSN 15409589. Cited at page 46.

OLSINA, L. et al. Incremental quality improvement in web applications using web model refactoring. **Lecture Notes in Computer Science**, 2007, v. 4832 LNCS, p. 411–422, 2007. ISSN 03029743. Cited at page 46.

OREHOVACKI, T. Proposal for a set of quality attributes relevant for web 2.0 application success. In: IEEE. 2010, 2010. p. 319–326. Cited at page 44.

OREHOVACKI, T.; HRUSTEK, N. Towards a framework for usability evaluation of educational artifacts created with web 2.0 applications: A pilot study. In: IEEE. 2013, 2013. p. 565–570. ISBN 9789532330762. Cited at page 46.

OTAIZA, R.; RUSU, C.; RONCAGLIOLO, S. Evaluating the usability of transactional web sites. In: IEEE. 2010, 2010. p. 32–37. ISBN 9780769539577. Cited at page 46.

PANACH, J. I. et al. Towards an early usability evaluation for web applications. **Lecture Notes in Computer Science**, 2008, v. 4895 LNCS, p. 32–45, 2008. ISSN 03029743. Cited at page 46.

PAOLINI, P. Hypermedia, the web and usability issues. In: IEEE. 1999, 1999. v. 1, p. 111–115 vol.1. Cited at page 44.

PATERNÒ, F.; PIRUZZA, A.; SANTORO, C. Remote web usability evaluation exploiting multimodal information on user behavior. In: SPRINGER. **Computer-Aided Design of User Interfaces V**. [S.l.], 2007. p. 287–298. Cited at page 45.

PAZ, F. et al. Heuristic evaluation as a complement to usability testing: A case study in web domain. In: IEEE. 2015, 2015. p. 546–551. ISBN 9781479988273. Cited at page 47.

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. Cited at page 34.

PAZ, F. A. P. F.; POW-SANG, J. A. Comparing the effectiveness and accuracy of new usability heuristics. **Advances in Intelligent Systems and Computing**, 2017, v. 497, p. 163–175, 2017. ISSN 21945357. Cited at page 47.

PUUSKA, S. et al. Nationwide critical infrastructure monitoring using a common operating picture framework. **International Journal of Critical Infrastructure Protection**, 2018, v. 20, p. 28–47, 2018. ISSN 18745482. Cited at page 48.

QADOUMI, B.; AL-SHURUFAT, B. Towards evaluation method of usability engineering for web application sites using 3d approach. In: ACM. 2015, 2015. v. 23-25-November-2015. ISBN 9781450334587. Cited at page 46.

RAMLI, M. H. et al. The adaptive model driven approach for enhancing usability of user interface design: A review process. In: CHIUXID'19. 2019, 2019. p. 65–69. ISBN 9781450361873. Cited at page 44.

RESKI, N. et al. "oh, that's where you are!'" – towards a hybrid asymmetric collaborative immersive analytics system. In: ASSOCIATION FOR COMPUTING MACHINERY. 2020, 2020. ISBN 9781450375795. Cited at page 44.

RIBEIRO, R. F. et al. Usability problems discovery based on the automatic detection of usability smells. In: SAC '19. 2019, 2019. p. 2328–2335. ISBN 9781450359337. Cited at page 44.

RIIHIAHO, S. et al. Experiences with usability evaluation methods. **Licentiate thesis. Helsinki University of Technology. Laboratory of Information Processing Science**, 2000. Cited at page 28.

RIVERO, L.; CONTE, T. Using an empirical study to evaluate the feasibility of a new usability inspection technique for paper based prototypes of web applications. In: SPRINGEROPEN. **2012 26th Brazilian Symposium on Software Engineering**. 2012, 2012. p. 81–90. Cited at page 44.

RIVERO, L.; CONTE, T. Improving usability inspection technologies for web mockups through empirical studies. In: RESEARCHGATE. **SEKE**. [S.l.], 2013. p. 172–177. Cited at page 48.

RIVERO, L.; KALINOWSKI, M.; CONTE, T. Practical findings from applying innovative design usability evaluation technologies for mockups of web applications. In: IEEE. 2014, 2014. p. 3054–3063. ISBN 9781479925049. ISSN 15301605. Cited at page 47.

RIVERO, L. et al. Evaluating software engineers' acceptance of a technique and tool for web usability inspection. In: RESEARCHGATE. 2015, 2015. v. 2015-January, p. 140–145. ISBN 1891706373. ISSN 23259000. Cited at page 47.

RODRÍGUEZ, F. D.; ACUÑA, S. T.; JURISTO, N. Design and programming patterns for implementing usability functionalities in web applications. **Journal of Systems and Software**, v. 105, p. 107–124, 2015. Cited at page 46.

RUSSELL, P. Infrastructure-make or break your e-business. **TOOLS-Pacific 2000: Technology of Object-Oriented Languages and Systems**, 2000. Cited at page 23.

RYBARCZYK, Y. et al. On the use of natural user interfaces in physical rehabilitation: A web-based application for patients with hip prosthesis. **Journal of Science and Technology of the Arts**, 2018, v. 10, 2018. ISSN 16469798. Cited at page 47.

SALAU, S. A. et al. Usability effectiveness of a federated search system for electronic theses and dissertations in nigerian institutional repositories. **Performance Measurement and Metrics**, 2020, v. 22, p. 1–14, 2020. ISSN 14678047. Cited at page 45.

SARRAJ, W. A.; TROYER, O. D. Web mashup makers for casual users: A user experiment. In: IIWAS '10. 2010, 2010. p. 239–246. ISBN 9781450304214. Cited at page 44.

SCHMIDT-KRAEPELIN, T. D. M.; SUNYAEV, A. Usability of patient-centered health it: Mixed-methods usability study of epill. In: BOOKS.GOOGLE. 2014, 2014. v. 198, p. 32–39. ISBN 9781614993964. ISSN 09269630. Cited at page 46.

SCHREPP, M. Goms analysis as a tool to investigate the usability of web units for disabled users. 2010, v. 9, 2010. Cited at page 45.

SHACKEL, B.; RICHARDSON, S. J. **Human factors for informatics usability**. [S.l.], 1991. Cited at page 30.

SHAMSUDDIN, N. A.; SYED-MOHAMAD, S. M.; SULAIMAN, S. Capturing users' actions in a web application to support learnability. In: IEEE. **2014 8th. Malaysian Software Engineering Conference (MySEC)**. 2014, 2014. p. 142–147. Cited at page 44.

SHIGA, Y.; TAKAMI, K. Development and evaluation of a contact center application system to integrate multiple communication channels using webrtc. **International Journal of Computer Networks and Communications**, 2017, v. 9, p. 1–20, 2017. ISSN 09752293. Cited at page 47.

SNYDER, C. **Paper prototyping: The fast and easy way to design and refine user interfaces**. [S.l.], 2003. Cited at page 30.

SOUTH, H. et al. Digitising a medical clerking system with multimodal interaction support. In: ICMI '17. 2017, 2017. p. 238–242. ISBN 9781450355438. Cited at page 44.

SOUTH, H. et al. Digitising a medical clerking system with multimodal interaction support. In: ACM. 2017, 2017. v. 2017-January, p. 238–242. ISBN 9781450355438. Cited at page 47.

SPEICHER, M.; BOTH, A.; GAEDKE, M. Ensuring web interface quality through usability-based split testing. In: SPRINGER. **International conference on web engineering**. [S.l.], 2014. p. 93–110. Cited at page 45.

SPIELER, B. et al. Development and evaluation of a web-based application for digital findings and documentation in physiotherapy education. **Studies in Health Technology and Informatics**, 2015, v. 212, p. 182–189, 2015. ISSN 09269630. Cited at page 47.

STANDARD, I. Ergonomic requirements for office work with visual display terminals (vdts)–part 11: Guidance on usability. iso standard 9241-11: 1998. **International Organization for Standardization**, v. 55, 1998. Cited at page 27.

SUDUC, A.-M.; BIZOI, M.; FILIP, F. G. User awareness about information systems usability. **Studies in Informatics and Control**, v. 19, n. 2, p. 145–152, 2010. Cited at page 23.

SWEDBERG, B.; PEUQUET, D. An evaluation of a visual analytics prototype for calendar-related spatiotemporal periodicity detection and analysis. **Cartographica**, 2017, v. 52, p. 63–79, 2017. ISSN 03177173. Cited at page 47.

TARKKANEN, K. et al. Back to user-centered usability testing. In: SPRINGER. **International Conference on Human Factors in Computing and Informatics**. [S.l.], 2013. p. 91–106. Cited at page 30.

TORRENTE, M. C. S. et al. Sirius: A heuristic-based framework for measuring web usability adapted to the type of website. **Journal of Systems and Software**, 2013, v. 86, p. 649–663, 2013. ISSN 0164-1212. Cited at page 45.

TUNÇ, S. K.; KÜLCÜ, Ö. A web application path analysis through server logs. In: ELSEVIER. **KDIR**. [S.l.], 2018. p. 425–428. Cited at page 47.

VALENCIA, X. et al. Assisted interaction data analysis of web-based user studies. In: SPRINGER. **IFIP Conference on Human-Computer Interaction**. [S.l.], 2015. p. 1–19. Cited at page 45.

VARGAS, A.; WEFFERS, H.; ROCHA, H. V. da. Analyzing user interaction logs to evaluate the usability of web applications. In: IEEE. **2011 3rd Symposium on Web Society**. 2011, 2011. p. 61–67. Cited at page 44.

VASCONCELOS, L.; JR, L. A. B. Usatasker: a task definition tool for supporting the usability evaluation of web applications. In: IADIS. **Proceedings of the IADIS International Conference on WWW/Internet 2012**. [S.l.], 2012. p. 307–314. Cited at page 46.

VASCONCELOS, L. G.; BALDOCHI, L. A.; SANTOS, R. D. C. An approach to support the construction of adaptive web applications. **International Journal of Web Information Systems**, 2020, v. 16, p. 171–199, 2020. ISSN 17440084. Cited at page 45.

VASCONCELOS, L. G. de; JR, L. A. B. Towards an automatic evaluation of web applications. In: ACM. **Proceedings of the 27th Annual ACM Symposium on Applied Computing**. [S.l.], 2012. p. 709–716. Cited at page 46.

VENKATESH, H. H. V.; ALJAFARI, R. A usability evaluation of the obamacare website. **Government Information Quarterly**, 2014, v. 31, p. 669–680, 2014. ISSN 0740624X. Cited at page 47.

WAKIL, K.; JAWAWI, D. N. A. A new framework for usability evaluation web engineering methods. **Journal of Theoretical and Applied Information Technology**, 2018, v. 96, p. 354–364, 2018. ISSN 19928645. Cited at page 47.

WICHIENNIT, N. et al. Design and development of application for crime scene notification system. In: IEEE. **2017 10th International Conference on Ubi-media Computing and Workshops (Ubi-Media)**. 2017, 2017. p. 1–6. Cited at page 44.

XEXAKIS, G.; TRUTNEVYTE, E. Are interactive web-tools for environmental scenario visualization worth the effort?' an experimental study on the swiss electricity supply scenarios 2035. **Environmental Modelling \ Software**, 2019, v. 119, p. 124–134, 2019. ISSN 1364-8152. Cited at page 44.

YAN, P.; GUO, J. The research of web usability design. In: IEEE. **2010 The 2nd International Conference on Computer and Automation Engineering (ICCAE)**. [S.l.], 2010. v. 4, p. 480–483. Cited at page 23.

YEOW, R. J. N. H. P. An empirical study of factors affecting the perceived usability of websites for student internet users. 2009, v. 8, 2009. Cited at page 45.

YOGASARA, T. et al. General characteristics of anticipated user experience (aux) with interactive products. In: DELFT UNIVERSITY OF TECHNOLOGY. [S.l.], 2011. p. 1–11. Cited at page 32.

ANNEX A – USABILITY TEST FORM

Usability Test of the Thoth Tool

Dear, this questionnaire is destined to collect data for a research in Software Engineering, conducted by the Master in Software Engineering at UNIPAMPA, with the objective of evaluating usability issues of the Thoth tool. Your participation is very important to improve the tool..

Master student: Fabienne Charles, fabiennecharles.aluno@unipampa.edu.br, UNIPAMPA.
Advisor: Prof.Dr.Elder de Macedo Rodrigues.
Co-advisor: Dra.Ildevana Poltronieri

*Obrigatório

1. E-mail *

FREE INFORMED CONSENT FORM

2. I agree to participate in this study and declare that I have read the details described in this document. I understand that I am free to accept or decline, and that I may discontinue my participation at any time without giving a reason. I agree that the data collected will be used for the purposes described above. I understand the information presented in this TERM OF CONSENT. I state that I have had the opportunity to ask questions and all of my questions have been answered. *

Marcar apenas uma oval.

- ◯ I agree

What is your education level?

3. Qual é o seu grau de escolaridade? *

Marcar apenas uma oval.

- ◯ Incomplete graduate
- ◯ Graduate degree complete
- ◯ Master uncomplete
- ◯ Master's degree complete
- ◯ PhD incomplete
- ◯ Complete PhD

4. What is your occupation? *

Marcar apenas uma oval.

- ◯ Teacher
- ◯ Student
- ◯ IT Industry
- ◯ Outro: _____

5. If your answer is other, tell what your working position is:

_____

6. Have you already heard of the Thoth tool? *

Marcar apenas uma oval.

- ◯ Yes
- ◯ No

7. Have you ever used the Thoth tool to conduct a Systematic Literature Review? *

Marcar apenas uma oval.

- Yes
- No

Usability Test

In this section we ask you to give us your opinion about the use of Thoth.
You will answer 14 multiple choice questions, these answers have a score on a scale from 1 to 5, being 1 very bad and 5 very good, and 1 open question, in which we want to know what improvements should be applied to Thoth.

8. In general, I am satisfied with the use of this system. *

Marcar apenas uma oval.

	1	2	3	4	5	
Totally Disagree						Strongly Agree

9. The system is simple and easy to use. *

Marcar apenas uma oval.

	1	2	3	4	5	
Totally Disagree						Strongly Agree

10. I am able to complete my work quickly using this system. *

Marcar apenas uma oval.

	1	2	3	4	5	
Totally Disagree						Strongly Agree

11. I feel comfortable using this system. *

Marcar apenas uma oval.

	1	2	3	4	5	
Totally Disagree	◯	◯	◯	◯	◯	Strongly Agree

12. I did not have any difficulties learning how to use this system. *

Marcar apenas uma oval.

	1	2	3	4	5	
Totally Disagree	◯	◯	◯	◯	◯	Strongly Agree

13. I think I became quickly productive using this system. *

Marcar apenas uma oval.

	1	2	3	4	5	
Totally Disagree	◯	◯	◯	◯	◯	Strongly Agree

14. The system provides error messages which clearly tell me how to solve the problem. *

Marcar apenas uma oval.

	1	2	3	4	5	
Totally Disagree	◯	◯	◯	◯	◯	Strongly Agree

15. The information (such as on-line help, on-screen messages, and other documentation) provided with this system is comprehensive. *

Marcar apenas uma oval.

	1	2	3	4	5	
Totally Disagree	◯	◯	◯	◯	◯	Strongly Agree

16. It is easy to find the information I need in the system. *

Marcar apenas uma oval.

	1	2	3	4	5	
Totally Disagree	◯	◯	◯	◯	◯	Concordo plenamente

17. The information provided by the system is effective in helping me to do my work. *

Marcar apenas uma oval.

	1	2	3	4	5	
Totally Disagree	◯	◯	◯	◯	◯	Strongly Agree

18. The organization of the information on the system screens is clear. *

Marcar apenas uma oval.

	1	2	3	4	5	
Totally Disagree	◯	◯	◯	◯	◯	Strongly Agree

19. I like to use the system interface. *

Marcar apenas uma oval.

	1	2	3	4	5	
Totally Disagree	◯	◯	◯	◯	◯	Strongly Agree

20. The system contemplates all the functionalities and capabilities that it should contemplate. *

Marcar apenas uma oval.

	1	2	3	4	5	
Totally Disagree	◯	◯	◯	◯	◯	Strongly Agree

21. In general, I am satisfied with this system. *

Marcar apenas uma oval.

	1	2	3	4	5	
Totally Disagree	◯	◯	◯	◯	◯	Strongly Agree

22. Could you give us your feedback so we can improve the Thoth tool. Be comfortable to explain to us what you think should be changed.

Thank you for your participation!

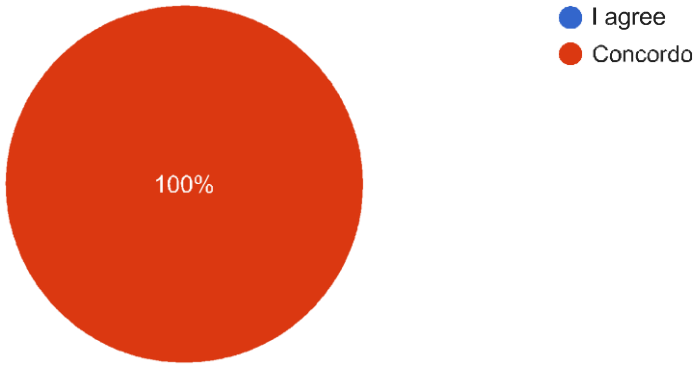
ANNEX B – THOTH TOOL USABILITY TEST ANSWER

FREE INFORMED CONSENT FORM

Copiar

I agree to participate in this study and declare that I have read the details described in this document. I understand that I am free to accept or decline, and that I may discontinue my participation at any time without giving a reason. I agree that the data collected will be used for the purposes described above. I understand the information presented in this TERM OF CONSENT. I state that I have had the opportunity to ask questions and all of my questions have been answered.

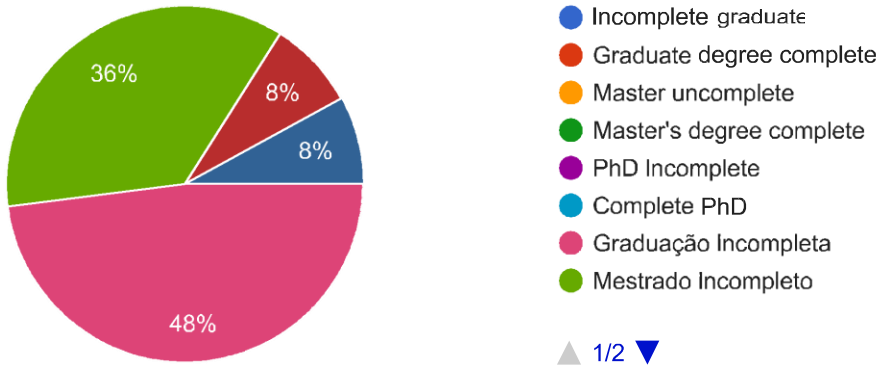
25 respostas



Copiar

01) What is your education level?

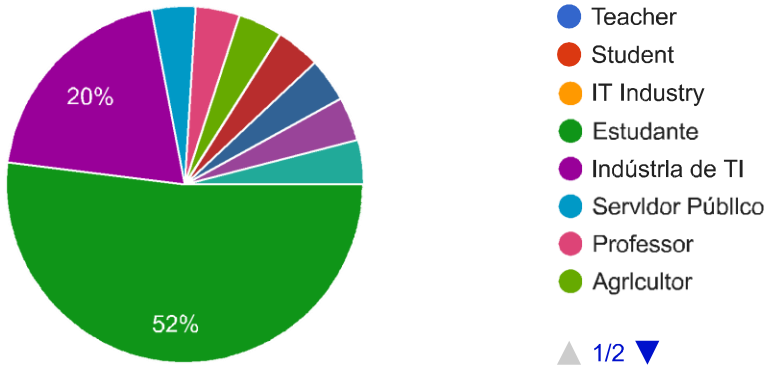
25 respostas



02) What is your occupation?

Copiar

25 respostas



03) If your answer is other, tell what your working position is:

7 respostas

Técnico em TI

Co-proprietário

Trabalhador Freelancer

Engenheiro de Software

Secretário Escolar

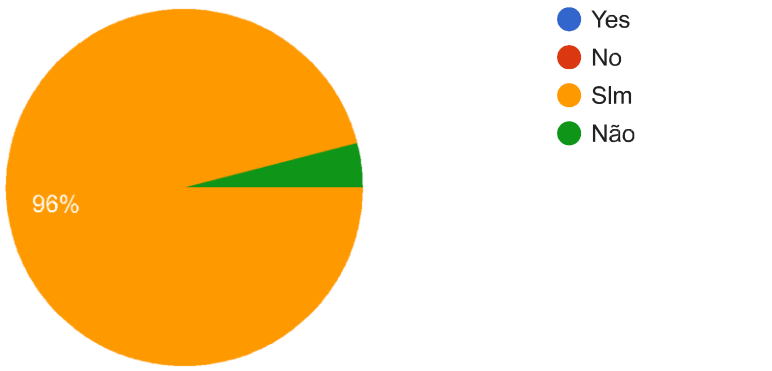
Programador Backend

Estudante de mestrado e da indústria

04) Have you already heard of the Thoth tool?

Copiar

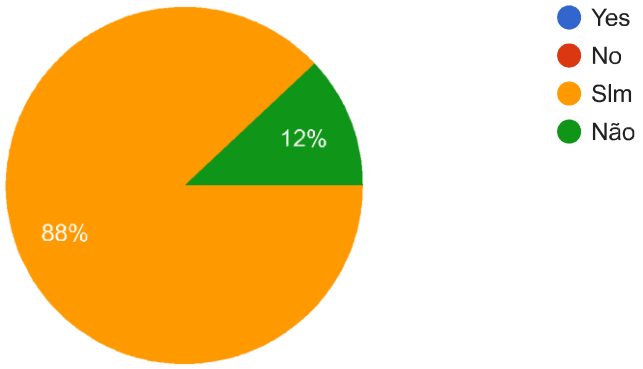
25 respostas



05) Have you ever used the Thoth tool to conduct a Systematic Literature Review?

Copiar

25 respostas

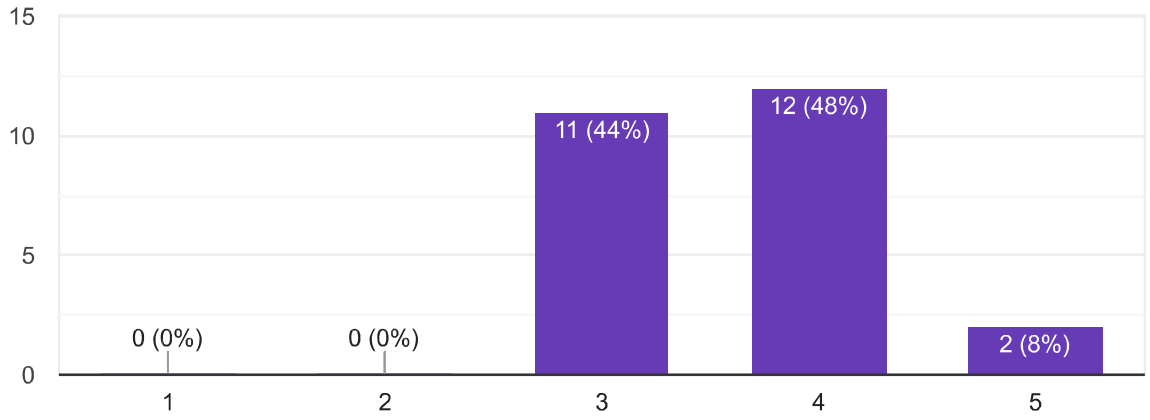


Usability Test

1- In general, I am satisfied with the use of this system.

Copiar

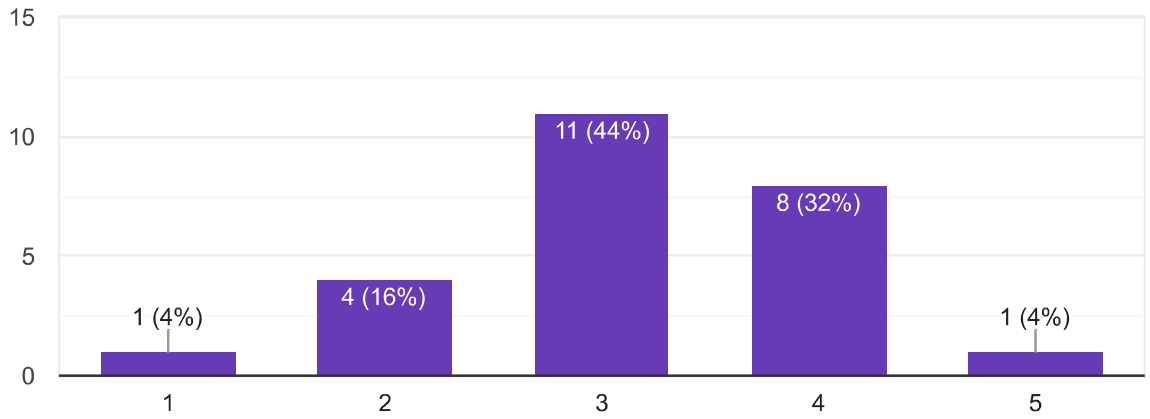
25 respostas



2- The system is simple and easy to use.

Copiar

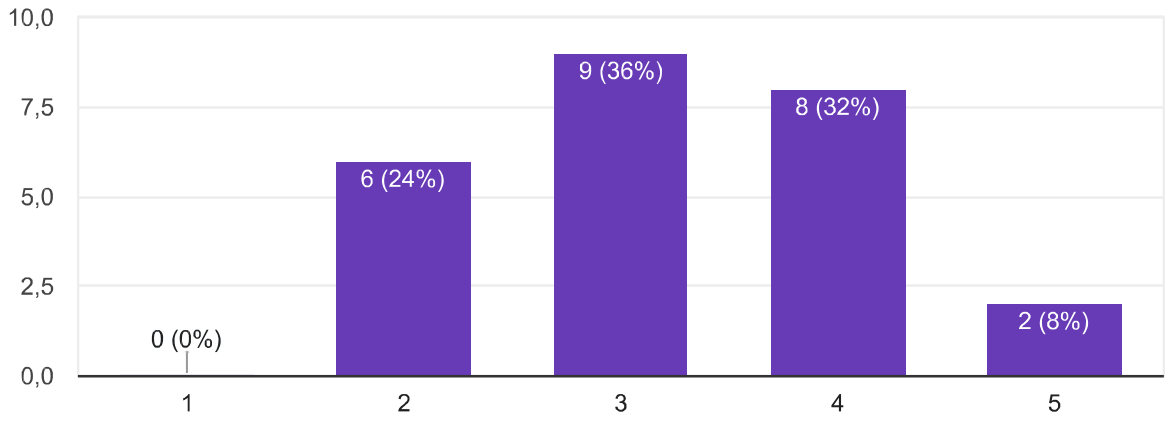
25 respostas



3- I am able to complete my work quickly using this system.

Copiar

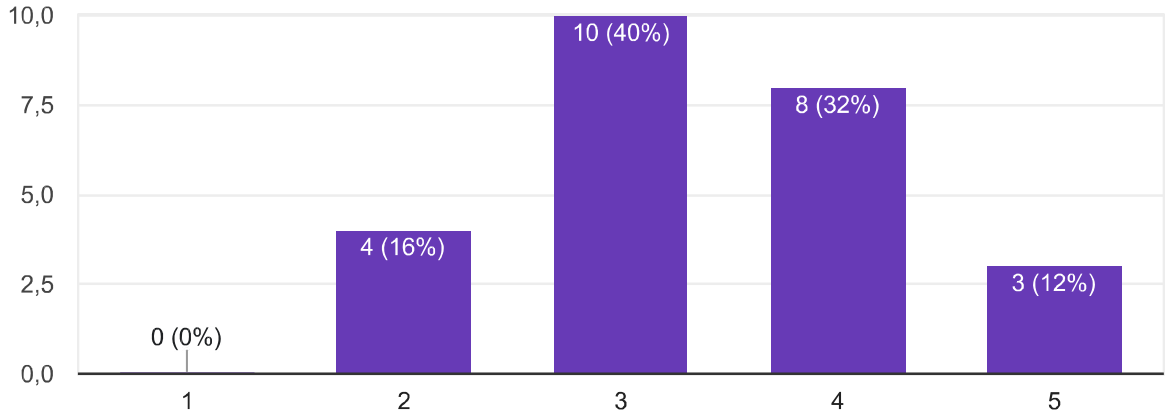
25 respostas



4- I feel comfortable using this system.

Copiar

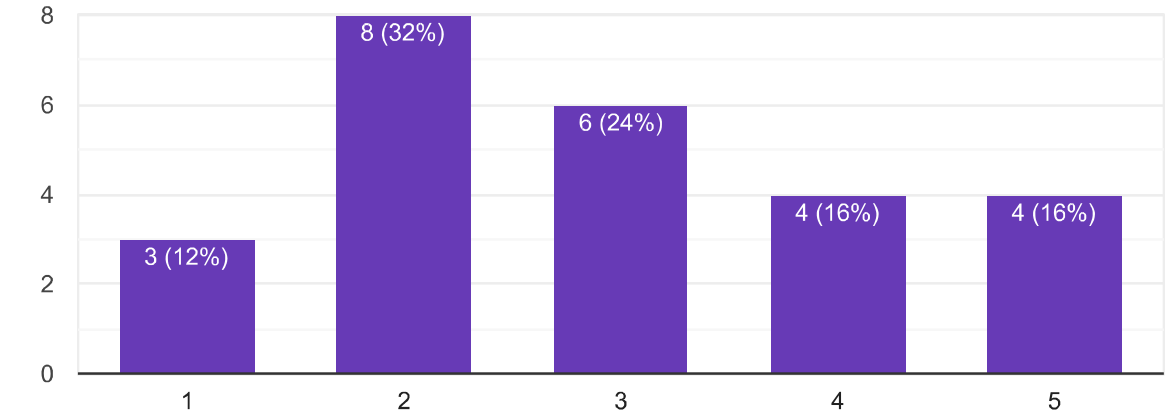
25 respostas



5- I did not have any difficulties learning how to use this system.

Copiar

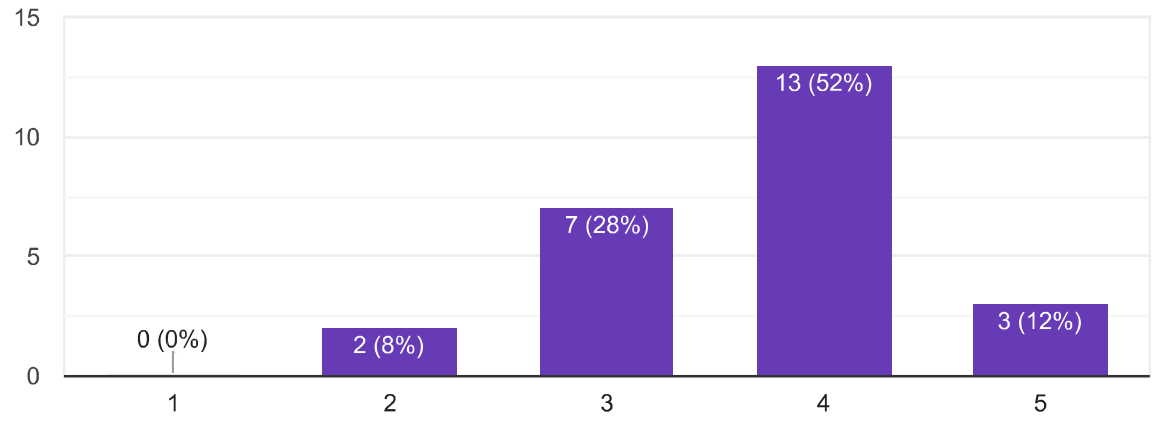
25 respostas



6- I think I became quickly productive using this system.

Copiar

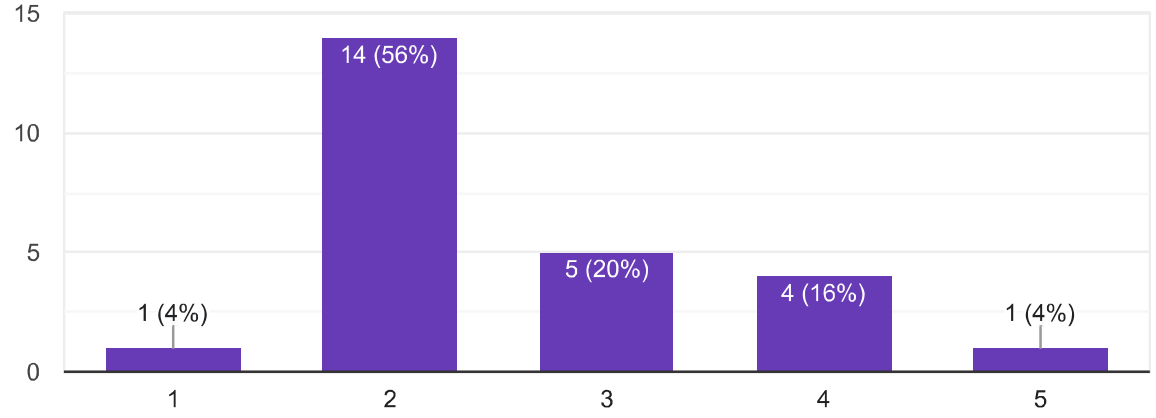
25 respostas



7- The system provides error messages which clearly tell me how to solve the problem.

Copiar

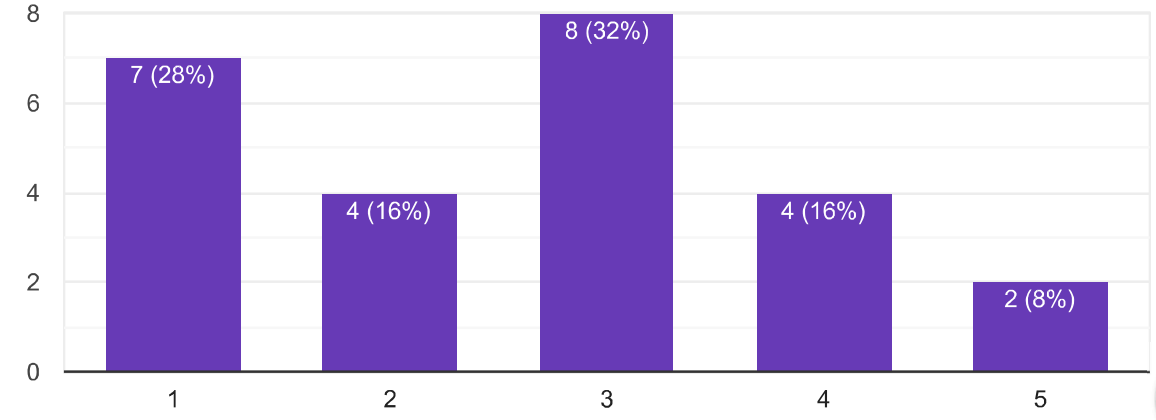
25 respostas



8- The information (such as on-line help, on-screen messages, and other documentation) provided with this system is comprehensive.

Copiar

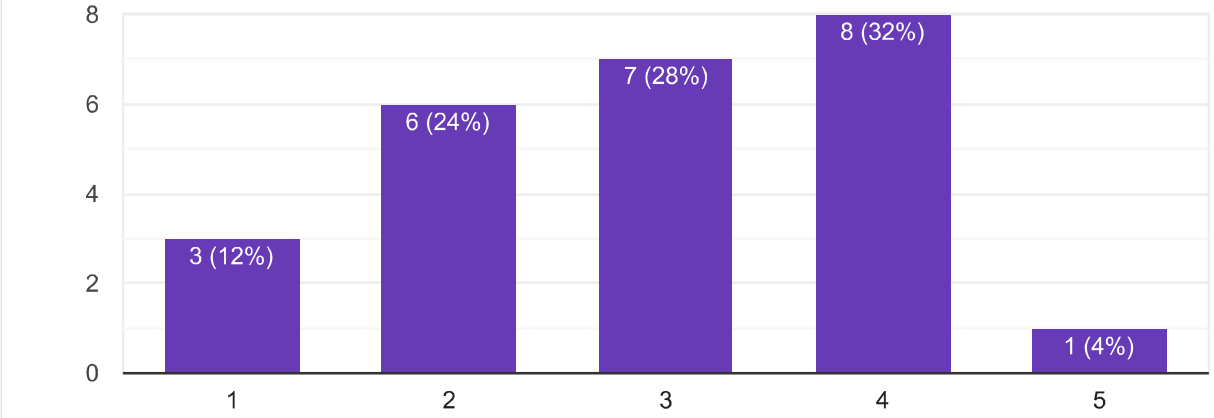
25 respostas



9- It is easy to find the information I need in the system.

Copiar

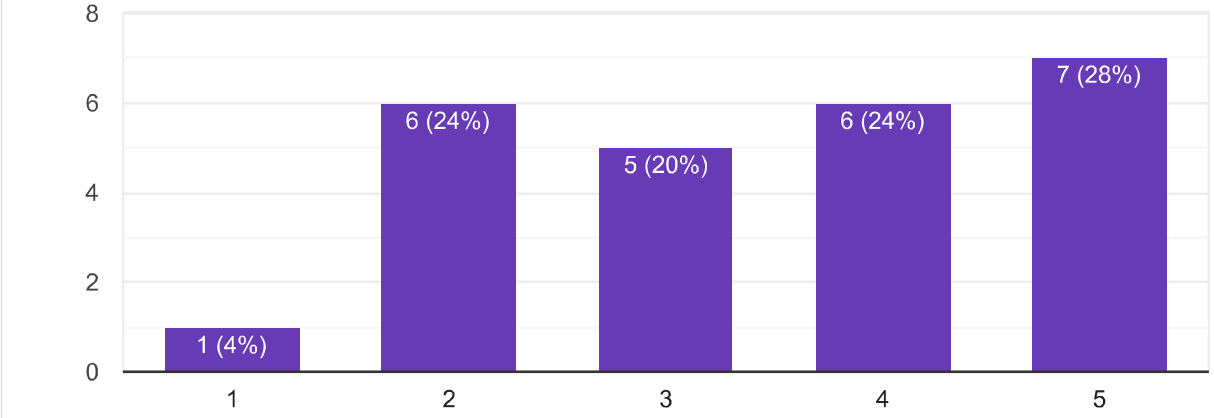
25 respostas



10- The information provided by the system is effective in helping me to do my work.

Copiar

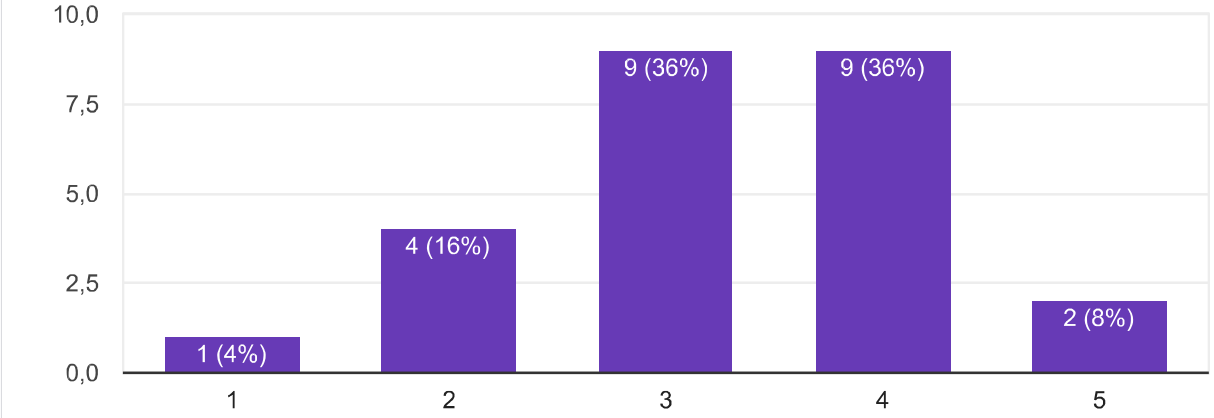
25 respostas



11- The organization of the information on the system screens is clear.

Copiar

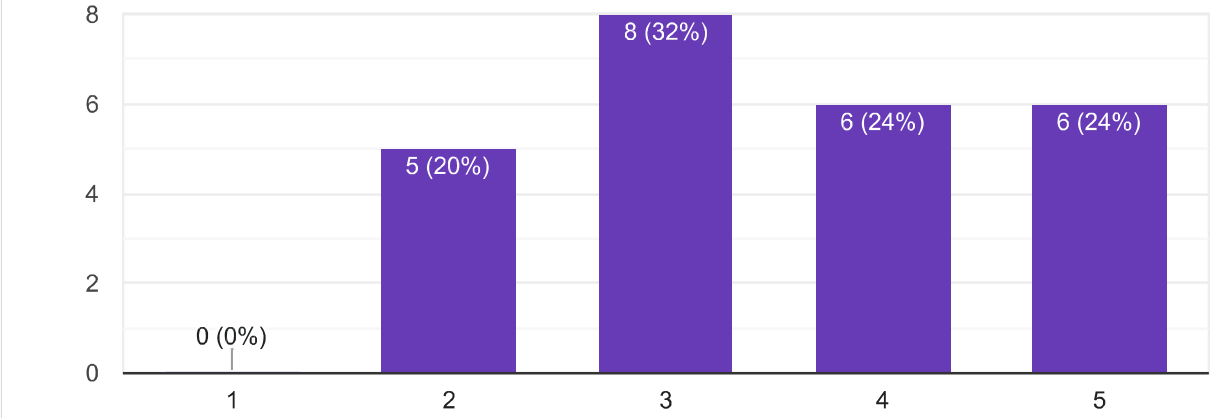
25 respostas



12- I like to use the system interface.

Copiar

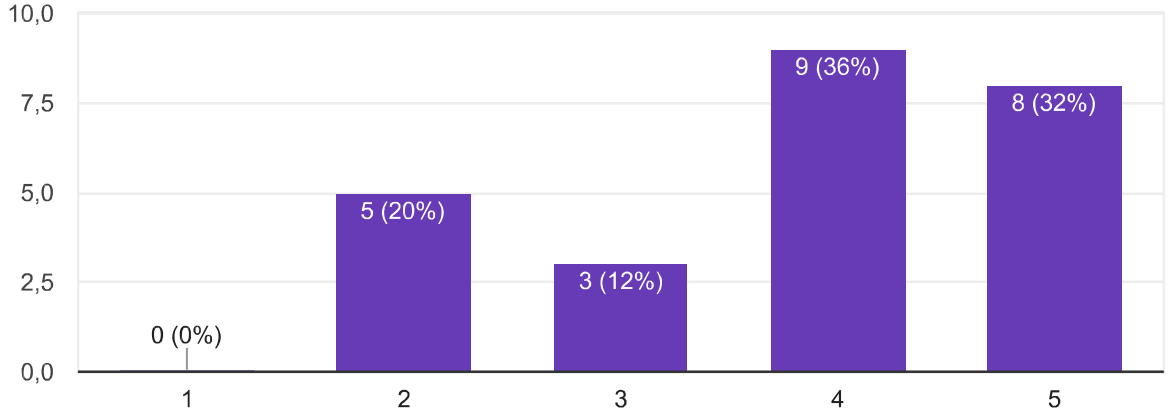
25 respostas



13- The system contemplates all the functionalities and capabilities that it should contemplate.

Copiar

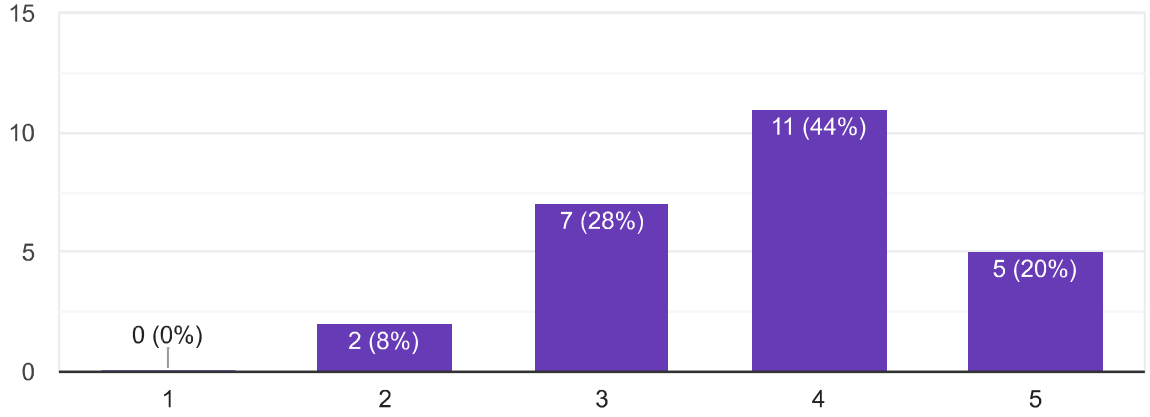
25 respostas



14- In general, I am satisfied with this system.

Copiar

25 respostas



ANNEX C – HEURISTIC EVALUATION, CONSENT FORM AND PROFILE EVALUATION

Heuristic Evaluation

Dear, this questionnaire is destined to collect data for a research in Software Engineering, conducted by UNIPAMPA, with the objective to conduct a heuristic evaluation of the Thoth tool. Your participation is of extreme importance so that the tool can be improved.
Master student: Fabienne Charles, fabiennecharles.aluno@unipampa.edu.br, UNIPAMPA.
Advisor: Dr.Elder de Macedo Rodrigues.
Co-advisor: Dra.Ildevana Poltronieri

*Obrigatório

1. E-mail *

Term of Consent

2. STATEMENT OF CONSENT TO THE PURPOSE OF THE RESEARCH I agree to participate in this study and declare that I have read the details described in this document. I understand that I am free to accept or decline, and that I may discontinue my participation at any time without giving a reason. I agree that the data collected will be used for the purposes described above. I understand the information presented in this AGREEMENT. I have had the opportunity to ask questions and all my questions have been answered. I will receive a signed and dated copy of this CONSENT FREE AND CLARIFIED document. *

Marcar apenas uma oval.

- ◯ I agree

Profile Evaluation

3. 01) What is your education level? *

Marcar apenas uma oval.

- ◯ Incomplete graduate
- ◯ Graduate degree complete
- ◯ Master Incomplete
- ◯ Master's degree complete
- ◯ PhD incomplete
- ◯ PhD complete

4. 02) What is your occupation? *

Marque todas que se aplicam.

- ☐ Teacher
- ☐ Student
- ☐ IT Industry
- ☐ Other

5. If your answer is other tell what your working position is: 0 pontos

_____

6. Have you ever conducted a Heuristic Evaluation? *

Marcar apenas uma oval.

- ◯ Yes
- ◯ No

7. Have you already conducted a Systematic Literature Review? *

Marcar apenas uma oval.

- Yes
- No

8. Have you already used or heard of the Thoth tool? *

Marcar apenas uma oval.

- Yes
- No

9. If you know Thoth, tell us how you heard about it: *

10. If you have already used Thoth how was your experience *

Thank you so much for your participation!!!

INDEX