


User perception as a factor for improving Trustworthiness in e-commerce systems

Andréia Rodrigues Casare  [Universidade Estadual de Campinas | casareandrea@gmail.com]

Celmar Guimarães da Silva  [Universidade Estadual de Campinas | celmar@unicamp.br]

Regina Moraes  [Universidade Estadual de Campinas | regina@ft.unicamp.br]

✉ School of Technology, University of Campinas. R. Paschoal Marmo, 1888, Jardim Nova Itália. Limeira, SP, 13484-332, Brazil

Received: 11 October 2023 • Accepted: 28 February 2024 • Published: 08 March 2024

Abstract

The User Interface (UI) is the first artefact that the user interacts with while developing a sense of trust that motivates him to use software applications more effectively. A badly designed UI can deceive users and bring the system into disrepute. Trustworthiness in UI is mandatory, as a poorly implemented UI can lead to the user misusing the system and jeopardizing the expected result. Trust in computational systems involves not only technical aspects, such as computational infrastructure, storage space, and service composition, but also aspects of Human-Computer Interaction (HCI). While technological aspects have received considerable attention, there are few research on human-computer interaction in terms of trust. This paper describes a way for assessing a system's trustworthiness based on the user's perception. The approach relies on a quality model to aggregate interface quality criteria in order to get a trustworthiness score. Three sets of experiments involving more than 300 individuals were carried out to validate the suggested methodology. A comparison was made between the trustworthiness score obtained through the methodology and the answers to open questions obtained through the users' questionnaires. The results were consistent, and statistical analysis corroborated the positive assessment. Based on these results, examples of improvements were developed to highlight the usefulness of the approach for developing more trustworthiness interfaces.

Keywords: Trust, Trustworthiness, User Experience, Quality Model, Usability, Accessibility

1 Introduction

In a globalized market, people are using online systems and applications more frequently, which has created new challenges for software development, including problems with human-computer interaction. However, difficulties with non-functional criteria, such as security, privacy, and trust, can impact the company's value or even prevent its relationship with the customer from being strengthened. Customer expectations must be met if organisations are to promote consumer trust and confidence and maintain their competitiveness in a global market.

Trust permeates every aspect of human life. Individuals in societies connect with one another with the hopes of building trusting relationships. There can be no business transactions or the introduction of new technology without trust. This occurs in the digital world as well, where a computational environment or product is selected based on the manufacturer's ability to create trust with the user who will utilise the target product or environment. In computer science, trust can be defined as the likelihood that an entity would display reliable behaviour for certain operations in a risky environment [Cho *et al.*, 2015], whereas trustworthiness is a system feature that can influence this person's trust in the system [Gao *et al.*, 2021].

To create economic and/or societal benefits, trust needs to be taken into account during all the software life cycle [Mouratidis and Cofta, 2010]. Furthermore, each user's perspective and the context of use influence trust. Thus, trust

involves not only technological qualities, but also human interaction characteristics such as user experience, accessibility, and usability. Particularly, the impact of user interface usability on trust in online stores was explored by Chen and Dibb (2010), and a substantial correlation between interface quality and trust was established. If a user interface is easy to use and understand, offering a positive user experience in addition to being functionally correct, it will be used again and trusted by those who use it [Sharma and Lijuan, 2015].

Trust is a ubiquitous and fundamental notion in many industries, but it is especially critical in e-commerce. According to our understanding, mainly in this context, trust derives not only from the right operation of e-commerce systems, but also from how effortlessly the user completes his work and how much confidence and pleasure he gets when using the system.

Online shopping has grown exponentially over the last decade and is regarded as a way for businesses to reach new customers ([Habib *et al.*, 2022]; [ITA Publishing, 2024]). Particularly, during the period of the COVID 19 pandemic, significant growth was reported by several studies ([ITA Publishing, 2024], [Al-Azzawi *et al.*, 2021], [Parlakkiliã *et al.*, 2020], [McKibbin and Fernando, 2021]). E-commerce websites play an essential part in online shopping because of their ability to reach and attract customers online, increasing user satisfaction and, as a result, attracting the attention of marketing practitioners, society, and academics.

Any e-commerce application uses the internet as a backdrop and a web page to encourage buyer-vendor interaction.

Its user interface is so critical that in a study by the Bentley University Design and Usability Centre, it was suggested that one only has about 6 seconds to make a positive impression on clients when they first browse to a website [Albert, 2012]. The result of the experiments suggests that the user's feeling of trust is mostly determined by the quality of the user interface.

Some studies that have been developed focusing on the pandemic period confirm the same and report that purchasing decisions were positively affected by trust, which in turn presents a positive relationship with satisfaction when shopping online ([Parlakkiliã *et al.*, 2020], [Attar *et al.*, 2020]). However, some authors go so far as to say that the shopper had no alternatives in a lockdown context and trust could be put aside but the lack of confidence could motivate a return to traditional commerce when shoppers can do so (Bonisoli and Castillo Leyva [2022]). Also, no studies were found in the literature that published a clear impact of the COVID pandemic in the trust feeling related to e-commerce systems.

So, defining a set of attributes and metrics based on the user interface is critical for determining how trustworthy an e-commerce service is. However, because each layer of the environment relies on a different set of attributes, different attributes are required to compose a computational environment that gives the user a sense of trust (e.g., scalability, availability, QoS, robustness, security, privacy assurance, dependability, and so on). These metrics can serve as an objective measure of a system's trustworthiness at a certain degree of confidence, and they can inspire adjustments to boost trust.

In this paper, we discuss the main challenges of measuring trustworthiness from the standpoint of the user experience, with a focus on e-commerce because they are widely used and typically collect and manipulate sensitive data (such as e-mail addresses, phone numbers, credit card numbers, and addresses, just to mention a few). In the context of e-commerce and for simplicity, the word "interface" is employed throughout this paper with the meaning of "user interface". Also, it is important to emphasize that this work extends and complements other works of the same group, which are cited throughout the text.

With an emphasis on user experience, we give an overview of various methodologies, techniques, and tools for measuring some trustworthiness characteristics. In addition, we developed and formalised a collection of user interface-based attributes and sub-attributes that characterize users' reported feelings of trust. Using the Quality Model given by the ISO/IEC 25000 (SQuaRE) standard [ISO, 2014], we propose the full directions to compose all these attributes towards an interface trustworthiness score, highlighting the normalization, weighting, and aggregation processes.

The composition and validation processes of the proposed approach encompassed several phases. Firstly, a systematic literature review was carried out to identify methodologies, techniques, and tools for measuring trustworthiness through the user experience in the context of e-commerce. Also, we added attributes that can be visually observed, such as a padlock, company information, privacy policies, and customer evaluations. This step allowed us to build the first version of such as performance of page up (i.e. website load time),

broken links, and AccessTx rate.

Following, an evaluation was carried out with 105 users through a questionnaire after using three e-commerce websites, with the aim of assessing the importance of the identified attributes as well as identifying new attributes that users cited as relevant. The results obtained were positive and reflected the users' perception of trust on those websites. Based on these results, the model was updated to accommodate new attributes suggested by the users and considered relevant. With the complete set of attributes and sub-attributes, we formalized all of them to make it clear how they are understood.

In the next phase, a new validation with more than 150 users was carried out, and the scores of other three e-commerce websites were calculated and compared with the users' answer about those websites' ranking. Again, the results were positive and fit the perception of the participants.

A new validation was performed in the last phase. This time, 50 users evaluated three websites from different context (three bank websites) that require even greater dependability and security solutions. Again, the results obtained through the answers to the questionnaires reflected the users' perception of trust. We realized that some variations in the weights of the component attributes occurred, and the model easily accommodated those variations. Complementing these last two phases, statistical analyzes were carried out to understand the significance of the results. Finally, based on these results, examples of improvements were developed to highlight the usefulness of the approach, always using Nielsen's heuristics as a guide. All details of the research, including questionnaires and raw data obtained, are available on our group's website ¹.

The rest of the work is organized as follows. Relevant concepts are presented in Section 2 followed by related work that guided our study in Section 3. The methodology used to get the final trustworthiness score, the metrics that compose the user interface quality model, and the model itself can be found in Section 4. Section 5 presents the results of the experiments, including two categories of websites. Finally, the conclusions and future work are presented in Section 6.

2 Background

This section briefly outlines the concepts that underpin this work. Essentially, it involves trust, user experience, measurements, and the quality model.

2.1 Trust

The concepts of trust and trustworthiness have been published in a variety of contexts, including people's social and business contexts. For example, the OECD (Organisation for Economic Cooperation and Development) defines trust in the social context as *a person's belief that another person or institution will act consistently with their expectations of positive behaviour* [OECD, 2017].

Trust is founded on a bilateral relationship between a subject (e.g., a user, the truster) and an object of trust - a tar-

¹<https://wordpress.ft.unicamp.br/seis/teste-piloto-trustworthiness/>

get entity, i.e., the entity that is trusted and is known as the trustee (e.g., a store, a bank, or a service) [Hussain and Chang, 2007]. An entity's decision to interact with others is an act of trust. The truster, in this circumstance, relies on and trusts the trustee to do the work as promised [Aljazzaf *et al.*, 2010]. In this sense, trustworthiness refers to the likelihood that a trustee would act in the manner expected by the truster [Bauer, 2019]. Mohammadi *et al.* (2014) argue that, while trust is an individual's concern based on their personal observations, trustworthiness is a system quality that might influence this person's trust in the system in either a positive or negative way.

Although these ideas are defined differently in diverse sectors, one of the universal key goals is to appropriately assess the amount of trust as a solid basis for decision making. One major issue is that trust levels are unclear and can fluctuate dynamically, making the development of trustworthy services difficult. It is mostly dependent on a user's feelings when interacting with the system, i.e., the quality of the human-system interaction. As a result, the user experience should be considered when calculating a system's trustworthiness score.

2.2 User Experience and Human-Computer Interface (HCI)

User experience is described as *the user's perceptions and reactions as a result of using a software product, system, or service*, according to ISO 9241-210 [ISO, 2019]. All of the user's emotions, perceptions, preferences, bodily and psychological responses, behaviours, and successes that occur before, during, and after use are included in the user experience. Also, according to Nielsen Norman Group ([Norman and Nielsen, 2024]) user experience "encompasses all aspects of the end-user's interaction with the company, its services, and its products".

To provide a good user experience, the software product needs to achieve excellence in a number of software quality attributes [Guerino and Valentim, 2020]. Usability is a quality HCI attribute that assesses how easy user interfaces are to use [Nielsen, 2024]. It is related to the simplicity of use and learning, as well as the ability of users to engage with the system to achieve their goals and happiness in utilising the computer system [Filho *et al.*, 2022]. Another HCI attribute that has a similar impact is accessibility. Web accessibility can be defined as the characteristic that allows citizens with any disability (visual, auditory, physical, cognitive, and neurological) to use, understand, contribute, interact, and navigate the internet without any type of barrier [Henry *et al.*, 2023].

The intention of a consumer to keep a relationship with a company is defined by his or her assessment of the benefits and high-quality service that provide a constant stream of value [Patterson *et al.*, 2006], and is highly tied to the system interface. Service quality refers to how well an information service provider's internal organisation, external supplier, and third parties are served. User perceptions are strongly tied to the process of creating consumer trust in e-commerce. Companies and organisations that offer e-commerce services must understand their customers' percep-

tions and how the companies connect with them. People typically avoid transacting over the Internet due to worries regarding consumer information exploitation, reliability, fraud, and payment [Jiménez *et al.*, 2021]. So, user experience is a result of the features, performance, system interactivity, or products that the user has experienced as a result of previous experiences, abilities, and context of use.

2.3 Metrics and Measurements

A software metric can be defined as any type of measurement technique that refers to a software product or system, process, or documentation. There are some examples of software metrics, such as the number of lines of code, the number of defects, and the number of error messages [Somerville, 2011]. Software metrics, according to Malhotra (2016), are the continual application of measurement-based methodologies to the software development process and its products, with the purpose of delivering relevant and management information to help the development and product improvement process. Software metrics can give engineers with the information they need to make technical decisions as well as project management information.

Software metrics can be classified as control or prediction metrics. Control metrics are used in software processes, whereas prediction metrics are used with software products, usually to measure software quality. But measuring the quality of a product in software development is not an easy task, as quality attributes such as ease of maintenance, ease of understanding, and ease of use are external attributes that are related to the perception of developers and end users. To measure the quality of these attributes, you must measure some internal attribute of the software, such as the code size or message count. Therefore, there must be a clear relationship between internal and external software attributes [Somerville, 2011], which should composed a model to provide a more complete measurement.

2.4 Quality Model

Trustworthiness is a multidimensional term that combines certain features, properties, and characteristics (such as security, privacy, fairness, transparency, and reliability, only to mention a few). These properties or characteristics have additional sub-attributes that enhance the number of possible solutions.

A Multi-Criteria Decision-Making (MCDM)-based technique can be effective in establishing how to compute a service's global score due to the analysis's multiple contradicting qualities. Logic Score of Preferences (LSP) [Dujmovic, 2007] was adopted in a similar way to the work of Olsina *et al.* [2008]. It is composed of numerous aggregation blocks that define how the various elements should be aggregated to produce a final score.

Service measures typically have distinct scales and dimensions. Before aggregation, the measures must be brought to the same scale. To do this, we used the normalising routines provided in the work of Frigal *et al.* (2016).

To use the LSP approach, first create a Quality Model [ISO, 2014], which is essentially a conceptual representation

of the system’s characteristics, weights, thresholds, and operators (for example, the tree structure in Figure 1). The blocks indicate attributes (leaf or composite) that are collected (by the operators). Bottom-level data (leaf attributes) are aggregated to form upper-level values (composite attributes), which are then used to determine the system’s overall score on a single 0-to-100 scale. Thresholds are normalisation function elements that determine the range of acceptable leaf-level attribute input values. Weight is an adjustable factor that specifies a preference for one or more system parameters (for example, memory utilisation may be more important than performance in particular situations).

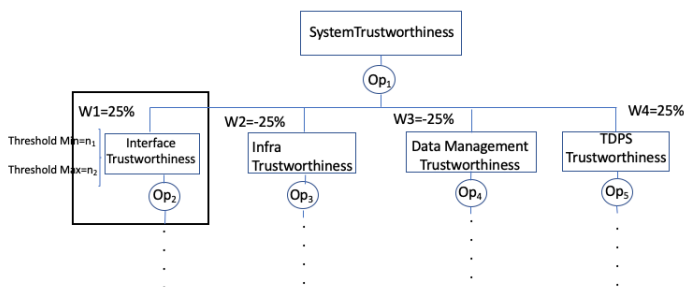


Figure 1. Quality Model

3 Related Work

To the best of our knowledge, trustworthiness measurement from the standpoint of user experience (i.e., user perception based on the user interface) has not been substantially researched. We were unable to find any work that provided a comprehensive solution for assessing the trustworthiness of an e-commerce system from the user’s perspective, ranging from the selection of appropriate attributes and sub-attributes to the definition of how to extract appropriate measures and the assess of a score that would allow comparison of the best available solutions.

This section resumes the result of a systematic literature review using the PRISMA framework [Moher *et al.*, 2009]. The research was carried out in studies published during the years 2009-2024. In addition to those, some classical references were added. Only English language works were considered, and peer-reviewed scientific papers such as journal articles, conference papers and books were included. The search was carried out using four multidisciplinary electronic databases specialized in the field of Computer Science: ACM Digital Library, IEEE Xplore, Web of Science, and Springer Digital Library. It involves the works focused on trust, trustworthiness in Subsection 3.1; the Human-Computer Interface and its impact on the user experience and system trust in Subsection 3.2; Subsection 3.3 report the works about trustworthiness metrics and measurement and finally works related to the Quality Model in Subsection 3.4.

3.1 Trust and Trustworthiness

As mentioned before, the notions of trust and trustworthiness have been approached in several areas. Mayer *et al.*

(1995) addresses the concept in terms of organizational relationships and suggests a model that defines trust. Shankar *et al.* (2002) developed an online trust conceptual framework based on the diverse views and demands of numerous stakeholders (such as customers, suppliers, employees, partners, and so on). McKinght *et al.* (2002) established a multidimensional model of online trust that includes four high-level components (disposition to trust, institution-based trust, trusting beliefs, and trusting intentions) and sixteen quantitative sub-constructs.

Focusing on identifying trustworthiness attributes, Hussin, Macaulay, and Keeling (2007) carried out a survey, whose objective was to identify attributes that impact the trust of e-commerce websites. Five e-commerce trust models were used to survey those attributes, which were extracted using online questionnaires, answered by 1230 participants. In addition to the attribute identification, the authors provide a ranking of the importance of each attribute on the user’s perception of trust based on the questionnaires’ top ranking on the Likert scale (It means that in the case of using a 7-point Likert scale, the attributes scored as 7). We borrowed this idea from them to weight the attributes selected in our study. A survey on systematic analysis of metrics, measurements, metrics properties, and associated ontologies was presented in the study of Cho *et al.* (2019). The authors improves the TRAM framework, previously published by the same group, by including security issues, addressing the measurement of four major system quality aspects: security, reliability, resilience, and agility. The STRAM (Security, Trust, Resilience, and Agility Metrics) structure is a hierarchical ontology structure for system trustworthiness, with each sub-metric defined as a subontology. In addition to the metrics, the paper outlines crucial evaluation procedures such as vulnerability assessment, risk assessment, and team building. However, various drawbacks are revealed in the framework, such as a lack of a clear description of used attributes, a lack of criteria to measure structures, and a lack of data to confirm the proposed metrics. Tao *et al.* (2015) use axiomatic methodologies to quantify software trustworthiness based on trustworthy attribute decomposition as well as the trust criterion (i.e. monotonicity, acceleration, sensitivity, substitutability, and expectability). To demonstrate the usefulness of the suggested metric, the reliability of network software was tested using a set of attributes and sub-attributes. The software’s critical attributes were reliability and maintainability, while portability and testability were non-critical. The results demonstrated that this model outperforms other measures provided in the literature, and the reliable qualities demonstrate that it meets the trust criterion. Part of the attributes that made up our approach were extracted from these works, which were essential in the composition of the approach’s quality model.

Tsuda *et al.* (2019) developed a methodology for assessing the quality of software products in development and use. The WSQF framework (Waseda Software Quality Framework) was evaluated using 21 software products and covers evaluations of the planning, development, maintenance, and operations phases. ISO 25023 [ISO, 2016b] and ISO 25022 [ISO, 2016a] were used as a base to develop measurement methodologies. The framework was a significant starting point for

our approach as it includes metrics for usability and accessibility. According to Basso et al. (2001), items such as photographs of company employees are a simple and efficient technique to build the customer connection. Lists of phone numbers or internet addresses are also typical elements in social-cue design. Providing the physical address of the store as well as its operating hours might help to create the impression that Internet clients can rely on the organisation.

3.2 User Experience and Human-Computer Interface (HCI)

The process of creating consumer trust in e-commerce is inextricably tied to the user's experience with the website. Companies and organisations who provide e-commerce services must understand their customers' perceptions and how they engage with the online interface, which is why a score that reflects user perception of the web interface is crucial for the industry.

Concerns regarding the exploitation of consumer sensitive data, which allows for fraud and monetary value theft, restrict people from conducting internet transactions [Jiménez et al., 2021]. According to Ang and Lee (2000), no purchase choice will be made until the website persuades the buyer that the merchant is reliable. The design of the interface for e-commerce transactions is one type of impact that may affect an online shopper's faith in the company. Websites with a more visually appealing interface typically result in greater service quality and can influence a user's experience and, ultimately, his or her long-term relationship with a service provider, thereby increasing potential consumers' perceived trust [Wang and Emurian, 2005].

According to several research in the Human-Computer Interaction (HCI) literature, user experience, usability, accessibility, and product quality as attributes (or characteristics) are factors that influence consumers' perceptions of trust when using a website or application [Mohammadi et al., 2013] [Hussin et al., 2007] [Henry et al., 2023].

The work most closely related to ours is Ramadhan and Iqbal (2018). The authors evaluate the user experience in terms of elements that influence user trust via the website interface design. They evaluated the three most popular bitcoin websites in Indonesia using approaches such as Experiential Overview, Post-Task Rating, Performance Metrics, Post-Session Rating, and an Eye-tracking device. However, they did not use a model to capture the multidimensional data, nor did they compute trustworthiness ratings to help identify the most trusted website. This work proposes a way to calculate a score that reflects this perspective, allowing comparison across different software products, in addition to defining and formalising interface design aspects (characteristics and subcharacteristics) that influence the user's feeling of trust.

3.3 Metrics and Measurements

Casare et al. (2021) recently found a collection of user interface-based traits that characterize users' perceived feelings of trust and codified a set of related trustworthiness metrics based on usability, accessibility, and user experience.

Concerning user experience, we have searched works that had previously been published to assess the user's trust, but those works are mainly focused on usability, and it was not possible to find any relevant work that goes deep into user experience in e-commerce.

ISO/IEC standards established a set of usability and accessibility attributes. ISO/IEC 25022 [ISO, 2016a] specifies measurements for the quality of user-system interaction. Measures include, among other things, satisfaction, efficiency, and effectiveness. Also, ISO/IEC 25023 [ISO, 2016b] specifies a set of quality metrics such as learnability, operability, and user interface aesthetics. Each characteristic and sub-characteristic has its own set of quality measures, as well as instructions on how to utilise them.

Regarding accessibility, Parmanto and Zeng (2005) developed WAB (Web Accessibility Barrier) metric to quantify web content accessibility. The score is based on the Web Content Accessibility Guidelines (WCAG) 1.0 milestones. In addition, the Unified Web Assessment Methodology (UWEM) [Nietzio et al., 2008] provides test descriptions for WCAG 1.0 conformance. Song et al. (2017) proposed the Web Accessibility Experience Metric (WAEM), which combines accessibility evaluation findings with user experience via paired websites comparisons. One year later the same group [Song et al., 2018] presented Reliability Aware Web Accessibility Experience Metric (RA-WAEM), an extension of WAEM, which takes into account user experience and dependability when determining the severity of accessibility barriers. The Web Accessibility Quantitative Metric (WAQM) was proposed by Vigo et al. (2007), which generates an accessibility score using evaluation reports provided by tools (for example, EvalAccess and LIFT tools).

In terms of usability, John Brooke (1996) proposed SUS (System Usability Scale), a set of usability measures that assesses the efficiency, effectiveness, user happiness, and ease of learning features. SUS scores are derived using a ten-item questionnaire given to respondents after they use the system under evaluation. Complementing this effort, Bangor et al. [2009] propose the inclusion of another question in the SUS questionnaire. This new question contains a list of adjectives that uses the 7-point Likert scale, which aims to help answer how the SUS numerical score is translated into usability judgement. Sauro, Jeff [2024] proposed a package that contains a calculator and a practical guide to the SUS. The calculator's role is to avoid inappropriate coding, perform statistical comparisons between two SUS scores, deal with missing values, calculate sample size, convert SUS scores to rating percentages and letter grades, and verify the reliability of responses.

Similarly, Kraig Finstad (2010) introduced a four-item questionnaire to quantify the same features, called UMUX (Usability Metric for User Experience). Also, the UMUX-LITE [Lewis et al., 2013] is based on UMUX and employs a two-item questionnaire. In the same vein, Seffah et al. (2006) introduced the Quality in Use Integrated Measurement (QUIM), a ten-usability factors model including ease of learning, satisfaction in use, efficiency, and effectiveness. These ten elements are further subdivided into a total of 26 subfactors or quantifiable criteria based on 127 distinct measures. Sauro and Kindlund's (2005) and Veral and Macias'

(2019) studies define usability as the mix of efficacy, efficiency, learning ease, and satisfaction. The SUM (Single Usability Metric) score was introduced in their previous study, which integrates the majority of information in four basic usability metrics: task completion rates, average number of errors, average length on task, and post-task satisfaction. The previous work presented a reaction card-based assessment of usability perception - a popular method for getting subjective user satisfaction in user experience ratings.

We looked into works involving questionnaires that had previously been used. The Computer System Usability Questionnaire (CSUQ) was created by Lewis (1995) and consists of 19 statements. They are rated by the user on a seven-point scale (plus N/A). It assesses factors such as ease of learning, ease of use, pleasure, utility, efficiency, and satisfaction. Similarly, Lund (2001) introduced the USE, which is a 30-question questionnaire with a 7-point Likert scale that rates ease of learning, usefulness, easy of use, and satisfaction. Lin et al. (1997) introduced the Purdue Usability Testing Questionnaire (PUTQ), which has 100 questions regarding computer interfaces and asks users to rate agreement on a 7-point scale (plus N/A). Hendradjaya and Praptini (2015) also propose a questionnaire with 9 questions that analyzes usability aspects such as ease of learning, simplicity of use, navigability, and consistency using the same scale.

Perlman (1997) created the Practical Heuristics for Usability Evaluation questionnaire, which is based on Nielsen's Heuristics and Norman's concepts. It comprises of 13 statements with which the user evaluates agreement on a 7-point scale (plus N/A). The assertions are separated into four categories: errors, feedback, learning, and user adaptation. Grannollers (2018) offered a set of principles for assessing user interfaces, as well as a series of 56 questions to be answered when studying each concept.

We take advantage of the works presented in this section to identify and complement attributes and sub-attributes, resulting in a set of 26 characteristics that fit better with e-commerce systems. Also, we formalize each related trustworthiness metric. To deal with the subjectivity of some characteristics, we use the questionnaire-related works to adapt our own questionnaire questions. Besides the professional profile questionnaire, other two questionnaires were used throughout the work reported in this article, one more general and one specific to identify the flexibility-related attributes.

3.4 Quality Model

The most known quality model is the one presented by the ISO/IEC 25000 (SQuaRE) standard [ISO, 2014]. Its tree structure formalizes metric interpretation and interrelationships. We opted to adopt this quality model in this task. In addition to being ISO/IEC standardised, it allows for the representation of several attributes as well as the determination of how measurements should be aggregated and what methodologies must be used to homogenise their values. Each attribute can have its own quality model, which can subsequently be aggregated using a hierarchical structure. This structure fits perfectly with our goal. Although the quality model is a standard, its content, weights, and thresholds must be defined for each scenario. This composition was created

for the context of e-commerce systems with the assistance of users, and preliminary versions have been published by the same group [Casare *et al.*, 2020] [Casare *et al.*, 2021]. It also allows for the setup of thresholds, weights, and operators.

Lew, Olsina, and Zhang (2010), also based on ISO/IEC SquaRE standard [ISO, 2014], established a framework for modeling quality, usability, and user experience requirements. Their purpose is to assess the quality of software of the Web applications. Accuracy, suitability, accessibility, and legal compliance are some of the characteristics used by the authors. Joseph and Mariappan (2018) created a platform called Trust Score that is a dynamic trustworthiness scoring approach based on five parameters: competency, perseverance, credibility, reputation, and integrity. The Trust Score is calculated by combining the 5 components and a K-factor, which is a trust normalisation constant (parameter weight). To the Trust Score is assigned a number between 0 and 1 (each parameter is normalized). Although the parameter set is different, some attributes and normalization are also present in our model.

Other authors used quality models for different purposes. To evaluate e-government websites, Hendradjaya and Praptini (2015) established a quality model including attributes, such as productivity, functionality, usability, dependability, efficiency, and portability. In their experiments, the data was collected using specialized Web tools and surveys. Olsina et al. (2008) provided an evaluation architecture that permits preserving values for particular real-world measurement and evaluation tasks. Their approach is very similar to ours in that it makes use of software quality attributes, metrics, weights, aggregation, operators, and the Logic Score of Preferences (LSP) technique. However, our model considers attributes that influence user trust and computes a final score that may be used to select the most trustworthy website (e.g., the one with the highest trustworthiness score). Other previous works ([Lew *et al.*, 2010], [Hendradjaya and Praptini, 2015]) presented quality models linked to usability and user experience, as well as some quality attributes related to trustworthiness (e.g., reliability, accessibility), but they did not focus on these characteristics.

4 Metrics Selection and Formalization – User Interface Quality Model

Identifying reliable measures of trustworthiness in the information from a system proved difficult. Given the complexities of trustworthiness, evaluating it only based on a quality attribute is extremely unlikely. Instead, the trustworthiness measure will be composed of several characteristics (or attributes, which is used interchangeably in this work) at different scales. To score based on a criterion, attribute values will need to be aggregated using a given procedure, which will almost certainly require that attribute values be expressed in the same units in order to be able to operate with them. Then, a multidimensional approach is needed to combine software attributes, which must be organized into a measurement model, to allow the scores to be calculated using the model and appropriate attributes [Tao and Zhao, 2018].

Normally, in an e-commerce context, the essential functionalities are displayed to users, and interactions are carried out via service interfaces. Then, a user can access e-commerce via various presentation channels (e.g., mobile application, website, social network, among others). Typically, e-commerce makes use of third-party services (such as payment, producer, distribution, and inventory), which must be transparent to the user. E-commerce can rely on numerous backup features to improve the company’s business and performance (for example, data analytics and storage performance), as well as infrastructure elements that help improve both the company’s business and the user experience (e.g., security and privacy protection, exceptions treatment). Each of those layers relies on a set of different attributes. Most work in this context covers small groups of associated attributes. We combined the results of several studies to compose our own set of attributes.

As previously stated, we chose the ISO/IEC Quality Model [ISO, 2014] for this work since it is flexible enough to represent multiple qualities, setup the thresholds, weights, and operators, and stipulate how the measures should be aggregated, as well as which techniques must be utilized to homogenize their values. It is feasible to design one quality model for each attribute, and then aggregate these diverse perspectives using a hierarchical structure.

The rest of this section presents the elicitation process of trustworthiness attributes in Subsection 4.1, the formalization of the metrics to measure these attributes in Subsection 4.2, and in the Subsection 4.3 the Interface Quality Model used to evaluate the trustworthiness of a system based on its interface.

4.1 The Elicitation Process of Trustworthiness Attributes

Like any part of a software product, measuring interface quality helps to understand deficiencies and guide improvements. Based on the findings of several studies in the literature, which were previously discussed in Section 3, it was possible to identify an initial set of quality attributes that influence user trust during the e-commerce process. The attributes were classified into three main groups: usability, accessibility, and user experience. The usefulness attribute a component of usability, as suggested by ISO 25022 [ISO, 2016a]. After this first phase, the research group analyzed the results obtained and added some attributes based on their own experience in using e-commerce systems. Then, a pilot test was applied, in which a group of specialists analysed three e-commerce systems and suggested a few more attributes that were incorporated to compose the final set of 26 attributes presented in Table 1. The first column brings the three groups of attributes, which are listed in the second column. The third column shows the way that the metrics were extracted in our validation process, and the last column shows the domain of the metric values. It is worth noting that one of the metrics (Performance) was found to be relevant to both usability and accessibility, thus falling into both categories. However, the weights assigned to this attribute in each group are different, since the importance of the attribute varies in the context of Usability and Accessibility.

Table 1. Measurable attributes that can impact trust

Type	Attributes	Measurement way	Metric
Usability	Coherent Buttons	Questionnaires	rate between 0 - 1
	Coherent Menus	Questionnaires	rate between 0 - 1
	Navigation	Questionnaires	rate between 0 - 1
	Easy of learning	Questionnaires	rate between 0 - 1
	Usefulness	Questionnaires	rate between 0 - 1
	Failure Handling	Questionnaires	rate between 0 - 1
	Number of failures	Questionnaires	rate between 0 - 1
	Users facing failures	Questionnaires	rate between 0 - 1
	Satisfaction	Questionnaires	rate between 0 - 1
	General Flexibility	Questionnaires	rate between 0 - 1
	Environment Flexibility	Questionnaires	rate between 0 - 1
	Responsive	Automatic tool	yes/no
	Performance	Automatic tool	rate between 0 - 1
Accessibility	Performance	Automatic tool	rate between 0 - 1
	Simple Screen	Questionnaires	rate between 0 - 1
	Colors and Fonts	Questionnaires	rate between 0 - 1
	Visibility of system status	Questionnaires	rate between 0 - 1
	AccessTlx	Automatic tool	rate between 0 - 1
	Back Button	Questionnaires	rate between 0 - 1
	Broken Links	Automatic tool	rate between 0 - 1
	Visible Focus	Questionnaires	rate between 0 - 1
User Experience	Company Information	Questionnaires	rate between 0 - 1
	Company Reputation	Questionnaires	rate between 0 - 1
	Customer Opinions	Questionnaires	rate between 0 - 1
	Padlock	Automatic visible	yes/no
	Pleasure	Questionnaires	rate between 0 - 1
	Privacy Policies	Questionnaires	rate between 0 - 1

Besides the metrics identification, the literature review was also used to identify examples of questionnaires to be applied, since the majority of the attributes (22 up to 26) depends on the users’ evaluation. Three questionnaires were generated. One of them collects the participant’s profile with personal data and professional training. A second questionnaire focused only on the general and environmental flexibility attributes since, to assess them, it was necessary to use different devices (for example, cell phones, tablets and computers). The third, more general, was focused on the remaining attributes.

Table 2 presents some statements composing the more general questionnaire, in this case related to the failure handling attribute. In this questionnaire, we want to ask users how much they agree or disagree with these statements. This agreement level was based on a 7-points Likert scale. It is important to notice that the first statement was suggested by the specialists during the pilot test, three statements were adapted from Granollers (2018), and one of them was adapted from Perlman (1997). Table 3, on the other hand, presents some statements related to the flexibility questionnaire, particularly the flexibility of the environment and also used a 7-points Likert scale. Those statements were suggested by the members of the research group. The full format questionnaires are available in our research group website.

Table 2. Failure Handling statements

Statements
The website provides error messages that clearly say how to fix the problems.
The site makes sure that the user can easily get out of an undesirable state [Perlman 1997].
Errors are shown in real time [Granollers 2018].
“Automatic saving” is implemented [Granollers 2018].
The website responds well to external failures (Power cut, internet does not work, among others) [Granollers 2018].

In addition to the attributes that were evaluated with the

Table 3. Environment Flexibility statements

Statements
The website is flexible to be used in different browsers.
The website is flexible to be used in different devices (smartphones, tablets).

help of the users, there is one attribute that can be visually evaluated in the interface (the Padlock) and 4 other attributes that can be evaluated with automatic tools (Responsive, Performance, AccessTx and Broken Links) once they are objectives attributes. The Mobile Friendly Test tool ² is the only stable tool identified to obtain responsiveness metric, which verifies if the website is ready to run on mobile devices. The tool's result is 0 (non-responsive) or 1 (responsive). Performance (the time/rate to load a page in the website) and AccessTx scores (how many accessibility recommendations are met) are calculated using the average of the measurements. The former is provided by Page Speed ³, PingDom ⁴ and GT-Metrix ⁵ and the latter uses the Ases ⁶, Nibbler ⁷ and Access Monitor ⁸ tools. The Broken Link score is calculated based on the maximum rate obtained by any of the Dead Link Checker ⁹, Screaming Frog ¹⁰ or Xenu's tools ¹¹. The tools, based on an initial URL, scan the website and analyze links and objects on the web page, calculating, according to their own criteria, the metric on which each one of them is focused and outputting a report. The reported metric value is then inserted into the Quality Model (until now entered manually). The formalization of these metrics is presented in the next section.

4.2 The Metrics Formalization

There are three groups of attributes, related to: User Experience, Accessibility, and Usability. Each attribute and sub-attribute that makes up our model (a tree-model) was formalized in line with ISO/IEC 25022 / 23 [ISO, 2016a] [ISO, 2016b] structure. When data is obtained from the appropriate source (e.g., an automatic tool or a questionnaire), the document specifies how each one should be calculated. The ID (Identification code), name, description, measurement function (formula detailing how the quality measure elements are integrated to produce the quality measure), and method (the source or type of method that can be employed to obtain the measure) comprise this structure. A quality measure ID is a code that represents the quality attributes and sub-attributes. It is a sequential number within a quality sub-attribute and G – for generally applicable – or S – Specialised for particular needs. For example, Op-3-G means the third sub-attribute generally applicable to compose Operable attribute. Some annotations are supplied for each attribute (or sub-attribute) to supplement comprehension of some aspect of the measure-

ment function or sources used to derive the measure.

The document containing all formalized metrics are available on our research group's website, and complements the previously publication of the same group [Casare *et al.*, 2021]. Figures 2,3, and 4 depict the formalization of three of those attributes. The first, Failure Handling, reveals whether or not the website handles and treats mistakes and errors that occur during user interaction. It is a leaf attribute (sub-attribute) of Safety in Use (composite attribute). The one shown in Figure 3, Environment Flexibility, assesses the website's ability to be used in many browsers and devices and is a sub-attribute of Efficiency in Use, which in turn composes the Usability attribute. Broken Links is a sub-attribute of Operable, which is a component of the Accessibility attribute. The method in the last column shows that it is obtained automatically by tools in this case. The use of at least three tools to determine the total number of links inspected and the number of broken links is advised in the notes. The measurement function is the instrument with the highest rate of broken links (assuming the worst situation).

Figures 2 and 3 both formalize sub-attributes extracted from questionnaires. The fourth column shows the expression used to obtain the score of a leaf sub-attribute as the average of the scores indicated on the answers of respondents for the questions related to this sub-attribute. In our case, as we are using a 7-point Likert scale questionnaires, S_{ij} is the value of the Likert scale (1 up to 7) that has been chosen by the respondents for each question related to this sub-attribute.

Expression 1 indicates the same calculation formula already expressed (and more simplified) based on the Likert scale. Once the average is calculated, we also report the standard deviation in Expression 2, which can be used for adjustments in the measurement of a particular case (but we did not use it in our experiments). The expressions consider the set of questions Q (e.g., $j(1), j(2), \dots, j(m)$) related to each attribute k . The variable i is the Likert Scale value, and n_{ij} is the number of times the value i of the Likert Scale was pointed out (by all the participants) for each question j of the attribute k . The average score considering all questions j belonging to the set of questions $Q(k)$ is AVG_{attr_k} , and the standard deviation of the scores considering the same set of questions is SD_{attr_k} .

$$AVG_{attr_k} = \frac{\sum_{j \in Q(k)} \sum_{i=1}^7 i * n_{ij}}{\sum_{j \in Q(k)} \sum_{i=1}^7 n_{ij}} \quad (1)$$

$$SD_{attr_k} = \sqrt{\frac{\sum_{j \in Q(k)} \sum_{i=1}^7 (i - AVG_{attr_k})^2 * n_{ij}}{\sum_{j \in Q(k)} \sum_{i=1}^7 n_{ij}}} \quad (2)$$

The last sub-attribute, Broken Links, is being used here to illustrate how the score for a sub-attribute extracted by automatic tools is calculated. In this case, as presented in the Sub-section 4.1, the metric is extracted relying on at least three automatic tools (in our case, Dead Link Checker, Screaming Frog, and Xenu's tools). These tools return the overall number of website links as well as the number of broken links. In this scenario, the measurement is relatively objective and is unaffected by the computational environment or network,

²<https://search.google.com/test/mobile-friendly>

³<https://developers.google.com/speed/pagespeed/insights/>

⁴<https://tools.pingdom.com/>

⁵<https://gtmetrix.com/>

⁶<https://asesweb.governoeletronico.gov.br/>

⁷<https://nibbler.insites.com>

⁸<https://wordpress.org/plugins/access-monitor/>

⁹<https://www.deadlinkchecker.com/>

¹⁰<https://www.screamingfrog.co.uk/>

¹¹<https://xenus-link-sleuth.en.softonic.com>

ID	Name	Description	Measurement function	Method
Sf-1-S	Failure handling	The site handles and treats the errors.	$X = \frac{\sum_{i=1}^U \sum_{j=1}^Q S_{ij}}{U * Q}$ i = user identifier j = question identifier U = the total number of users Q = the total number of questions S _{ij} = the score of the question j given by the user i	Questionnaire
Note 1 The questionnaire uses as Likert scale point (1 to 7) and the related questions can be found in Appendix G. Note 2 To answer the questionnaire, users must use the site before. Note 3 Minimum 31 users (to meet statistical significance in accordance to Triola(1999)). Note 4 The site must be tested in different browsers (e.g.,Firefox, Chrome, Safari) and with a minimum of 10 users in each browser.				

Figure 2. Formalization of General Interface Metric - Failure Handling

ID	Name	Description	Measurement function	Method
Ef-2-S	Environment flexibility	The site is flexible to be used in different browsers and devices.	$X = \frac{\sum_{i=1}^U \sum_{j=1}^Q S_{ij}}{U * Q}$ i = user identifier j = question identifier U = the total number of users Q = the total number of questions S _{ij} = the score of the question j given by the user i	Questionnaire
Note 1 The questionnaire uses as Likert scale point (1 to 7) and the related questions can be found in Appendix F. Note 2 To answer the questionnaire, users must use the site before. Note 3 Minimum 31 users (to meet statistical significance in accordance to Triola(1999)). Note 4 The site must be tested in different browsers (e.g.,Firefox, Chrome, Safari) and with a minimum of 10 users in each browser.				

Figure 3. Formalization of General Flexibility Metric - Environment Flexibility

ID	Name	Description	Measurement function	Method
Op-3-G	Broken links	The amount of website links pointing to non-existent web pages	$X = \text{Max}(\sum_{i=1}^t (B_i/L_i))$ i = tool identifier t = the total number of tools B = the amount of broken links L = the total links of site	Automatic tool
Note 1 Minimum of 3 automatic tools. For example, Dead Link Checker (https://www.deadlinkchecker.com/), Xenu's Link Sleuth (https://xenu-link-sleuth.softonic.com.br/), Screaming Frog Seo Spider (https://www.screamingfrog.co.uk/) can be used. Note 2 Each tool must return the total number of links inspected and the number of defective links.				

Figure 4. Formalization of Automatic Tool Metric - Broken Link

relying solely on the tool’s accuracy when scanning the page. The broken link rate is then determined based on the highest rate obtained by any of the tools, i.e., it is calculated by the Expression 3 (see Figure 4):

$$BrokenLinkRate = \max_{t \in T} \frac{B_t}{L_t}, \tag{3}$$

where B_t is the number of broken links found by the tool t ; L_t is the total number of links found by the tool t ; and $T = \{DeadLink_Checker, Screaming_Frog, Xenu's_Link_Sleuth\}$ is the set of tools that were used to detect broken links.

The next section presents how we accommodate the attributes and sub-attributes in the Quality Model.

4.3 The User Interface Quality Model

As mentioned before, the Quality Model (QM) is a reference model proposed in the ISO/IEC 25000 (SQuARE) standard [ISO, 2014] that formalizes the interpretation of measures and their relationships. It enables the description of several attributes as well as the specification of how measurements should be aggregated and what approaches must be used to homogenize their values. A QM is organized in a hierarchical framework. Its leaf properties represent metric definitions (see Subsection 4.2), and their associated scores are dependent on some measurement processes. The collected

leaf attributes match against thresholds (*Threshold Min* and *Threshold Max* features) to guarantee that only relevant and authentic data is included. A min-max normalization (using *Normal Min* and *Normal Max* features) ensures that operators aggregate values at the same scales.

In the context of Figure 5, the three child attributes of Interface Trustworthiness have distinct weights (W_1 , W_2 and W_3). Usability ($W_1 = 35\%$) is equally significant as Accessibility ($W_2 = 35\%$) and both are more essential than User Experience ($W_3 = 30\%$). The final score is calculated by aggregating the attribute values from the leaf-level to the root attributes, using the Operators (OP_n) that explain the relationships between them. To define the conditions under which composite attributes are aggregated, several operators such as simultaneity (all requirements must be satisfied), replaceability (used when one of the requirements has a higher priority than the remaining requirements), and neutrality (combination of simultaneity and replaceability) can be used.

In addition to the Quality Model (QM) for Interface Trustworthiness, the ATMOSPHERE project¹² created the following QMs (see Figure 1): Infra Trustworthiness (in charge of assessing the trustworthiness of the system infrastructure), Data Management Trustworthiness (in charge of assessing data storage resources), and Trustworthy Data Processing Services – TDPS Trustworthiness (in charge of defining the attributes of the services that are running in order to provide

¹²<https://www.atmosphere-eubrazil.eu>

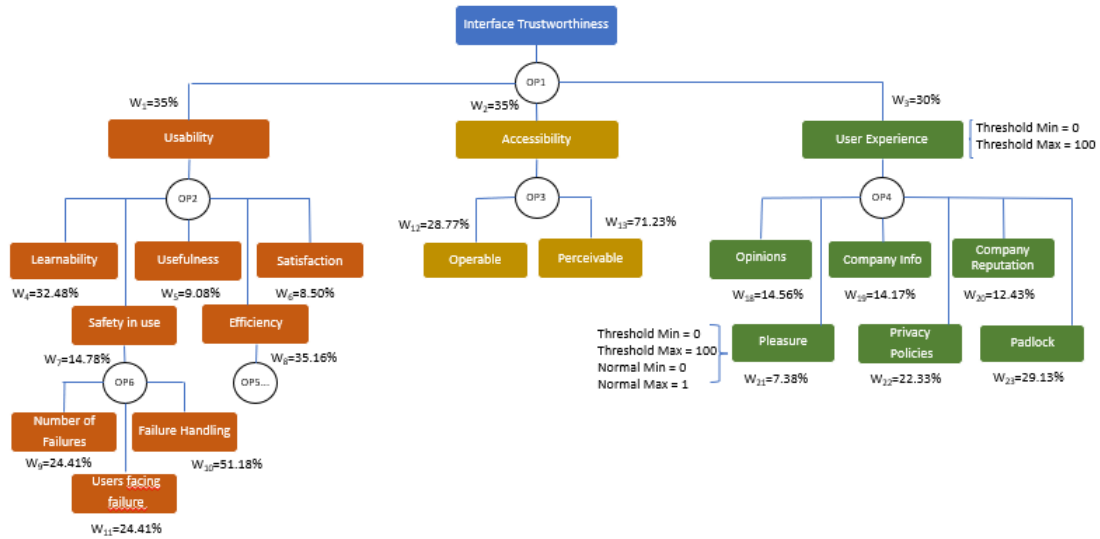


Figure 5. Interface Trustworthiness Quality Model

the expected results to the users). More information on these other QMs can be found on the website of our research group.

To get the score for each QM attribute, a min-max normalization must be done to transform the value AVG_{attr_k} from the Likert Scale [1,7] to the interval score [0,1], as defined in the Expression 4, where $S_{min} = 1$ and $S_{max} = 7$ in our case.

$$Score_{attr_k} = \frac{AVG_{attr_k} - S_{min}}{S_{max} - S_{min}} \quad (4)$$

One important element of the QM is the weight of the attributes. It represents the importance of the specific attribute (or sub-attribute) in the score composition of their parents. We examined the questionnaire responses and utilized them to calculate the weights for each composite attribute in the Interface QM. These weights were determined based on the perceptions of the participants, i.e., the attributes (or sub-attributes) with the greatest score (7) are considered more significant for gauging trust. The equation for computing the weight for each composite attribute is presented in the Expression 5. $Weight_j$ is the relative relevance of the attribute j to its parent attribute. If two attributes (or sub-attributes) have the same number of respondents who gave them the highest score (7), the second highest score (6) will be used to determine which one is the most essential; then the third highest score (5) will be used, and so on and so forth.

$$Weight_j = \frac{n7_j}{\sum_{x=1}^n n7_x} \quad (5)$$

As the Interface Trustworthiness QM has many attributes and sub-attributes and the presentation of the complete model makes it difficult to read, the Figures 6, 7 and 8 present the sub-QMs of Usability, Accessibility and User Experience, respectively (the second level sub-attributes of our QM). As already mentioned, it is possible to define a QM for each characteristic and then group them. The grouping of the three second-level characteristics can be seen in Figure 5.

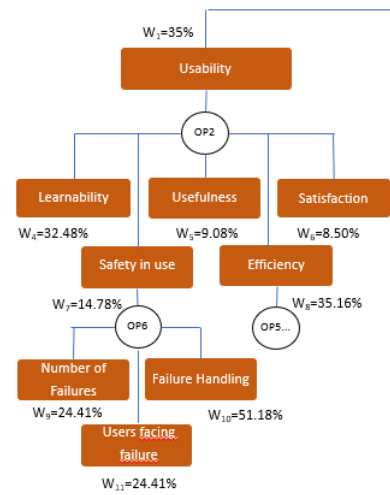


Figure 6. Usability Quality Model

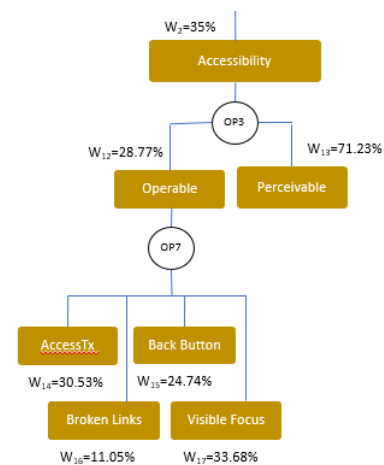


Figure 7. Accessibility Quality Model

5 Trustworthiness Measurement Experiments

This section describes the experiments carried out to evaluate and validate the suggested methodology for using the Interface Trustworthiness Quality Model to produce interface-

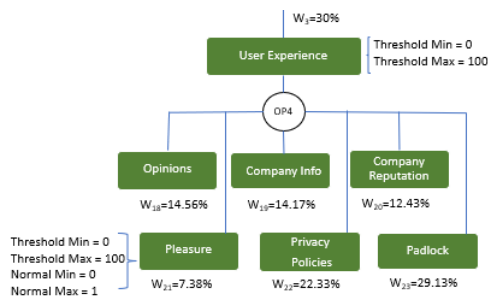


Figure 8. User Experience Quality Model

based trustworthiness scores for websites, as well as to help development professionals to improve the interface design. Firstly, we report in Subsection 5.1 the results of a pilot test, which was important to evaluate the QM's applicability and to complement the set of metrics used to compose the model. To validate the methodology, two sets of experiments were conducted, both of which used e-commerce websites and are detailed in Subsections 5.2 and 5.3. Finally, the Subsection 5.4 presents a validation based on a bank website, a system context that requires even more reliable solutions.

To meet the standards of the research ethics board, all volunteers in all human experiments were required to complete a Informed Consent Form (*Termo de Consentimento Livre e Esclarecido – TCLE*, in Portuguese). This document is part of the process approved by Ethics Committee - CAAE n. 30471320.8.0000.5404.

The task script followed by participants in any of the experiments is as follows: (i) read and agree with the TCLE; (ii) if agreed, answer the profile questionnaire; (iii) read the questionnaires to be aware of the questions to be answered; (iv) perform the experiments as requested (depending on the group they belong) until the point before the payment task; (v) answer the test questionnaire.

5.1 The Pilot Test

This section describes the preliminary experiments (pilot test) that were conducted in order to determine interface-based trustworthiness scores for e-commerce websites using the Interface Trustworthiness Quality Model. The goal of this pilot test was to refine the user testing task elements, which includes the following actions: improving the instructions on how to use the websites that will be the focus of assessments; improving the questions in the surveys; complementing the set of metrics; and verifying planned calculations on acquired data using the proposed metrics, aiming to understand the weak points of the whole process.

Twenty-one persons (12 men and 9 women) between the ages of 21 and 52 took part in the tests to evaluate three e-commerce websites. The users were divided into two groups: 9 participants performed the test using one website and different devices (laptops and smartphones or tablets) and browsers (Safari, Firefox or Chrome), and answered the questionnaire composed of general and environment flexibility questions (see Table 3); the other 12 participants, using the device of their choice, performed the test on the three websites and answered one questionnaire for each evaluated

website, with questions about failure handling, satisfaction, usefulness, and so on. In all experiments reported in this work, the participants were free to choose their own devices and to visit the websites more than once as well.

Based on the questionnaire responses, the average, standard deviation, and score for each attribute represented in the Interface QM were determined.

Moreover, the scores for Broken Link Rate, Performance Page Up, Responsive Rate, and AccessTx Rate attributes were also calculated based on the results of the automatic tools.

After calculating the scores for the leaf attributes (i.e., the attributes whose calculus were obtained through questionnaires or automatic tools), these values were used to calculate the scores of their respective composite attributes in the Interface QM. To do this, firstly the weight for each composite attribute was calculated. Finally, the whole process was applied to get the Interface Trustworthiness score.

We did not publish here the results collected because the purpose of this experiment was not to provide test results or conclusions (given that it was too early to deduce anything based on measurements obtained with a small number of users at the time). In any case the results are available elsewhere [Casare *et al.*, 2022b]. The pilot test met the desired goals once we were able to discover and fix weaknesses in some parts of the process (task script and elements) prior to testing with a larger number of users.

In a first comment, the participants noted that understanding the post-test questionnaire prior to beginning the test itself helped them pay more attention to some features of the interface and the task that had to be completed during the test on the website. As a result, we revised our task script to recommend that participants read the post-test questionnaire before interacting with the website.

Some participants complained that the option “not applicable” (N/A) was missing from some questions. For example, they argued that they had doubts about how to grade questions related to website problems if there is no failure. The questionnaire was reviewed, and this option was added to the questions concerning Failure Handling, as well as an observation in the instruction to select the option “4” (neutral score) if in doubt about which answer to choose.

During the results computation stage, there was a lack of information about each evaluated website's “start and end time”. In addition to measuring the test effort, it is important to quantify the failure rate, which is one of the model's properties and was completely overlooked. The information is now required in the questionnaire, and we added an alert to the guideline to emphasize its importance.

We were first skeptical about the value of conducting the entire evaluation procedure because the results with only a few people would not be reliable enough to draw any firm conclusions. Fortunately, we persisted in finishing the Quality Model with all of the collected metrics and calculating the outcomes (all of the scores). As a result, we discovered that the weights of the metrics acquired by the automated tools were overestimated. This was happening because the value assigned to the weights of those attributes was the complement to 100%, considering the set of other sub-attributes in the same group, which was not correct. To address the is-

sue, a question on the relevance seen by the participants in relation to the automatic attribute was included to each tool.

During the pilot test for the more comprehensive test, the following issues were resolved: (i) a suggestion to read the post-test questionnaire before accessing the website was added to the task script, as well as a highlight on the importance of filling out the time of the test start and end time on each website; (ii) the option “N/A” was added to some questions, along with a remark linking option “4” when no answer is adequate; (iii) the start and end time were added to the post-test questionnaire; and (iv) a question for each measure acquired by the automatic tools was included to capture the participant’s judgement of its weight.

5.2 Methodology Validation - First set of E-commerce websites

This experiment aimed to validate the methodology, all metrics created, the quality model, and generate the trustworthiness score of each website.

5.2.1 First Experiments Description and Results

Three tasks were carried out: a set of participants answered the more general questionnaire (with questions about satisfaction, usefulness, failure handling, among others); other set of participants answered the flexibility context questionnaire; and the automatic tools were used to get the pertinent metrics. At this point, three e-commerce websites (referred as A, B, and C to preserve their identity) were used: one of a world-renowned e-commerce, one of a famous Brazilian e-commerce and one of a famous Brazilian product company. Before starting the experiments with the participants, the team of researchers evaluated the websites proposed in the experiment, and each one cast a vote, placing the order in which they classified the three websites in terms of trust. These votes were saved to be used at the end of the first experiment to compare the results obtained.

From November 8th to December 17th, 2021, 105 people (82 males and 23 females), aged from 18 to 60 years old, took part in tests and answered questionnaires after using the three websites. In terms of professional background, 55% of the participants do not work in information technology and have never worked with interfaces or computer systems; 32% have worked with computer systems; and 13% are now working with computer systems. The participants were divided into two groups: 51 took the test on each of the three websites and answered the more general questionnaire for each of them; the other 54 took the test on one website using different devices (laptops and smartphones or tablets) and browsers (Chrome, Firefox, Safari), and answered a flexibility context questionnaire. The research group’s website has further information about the surveys and their responses.

The average, standard deviation, and score for each attribute comprising the QM were calculated using the questionnaire responses. Table 4 presents these calculations of the leaf sub-attributes for the three websites.

The leaf attributes that are evaluated with the help of automatic tools are presented in the next tables. Table 5 shows the performance of the page up score, which is computed

by using the average of three automatic tools’ measurements (*Page Speed*, *PingDom*, and *GTMetric*). The same calculation (i.e. the average) is used to obtain the AccessTx rate (Table 6), which uses other three tools (*ASES*, *Nibbler*, and *Access Monitor*). The broken links score is calculated based on the maximum rate obtained by any of the tools (*Dead Link Checker*, *Xenu’s Link*, and *Screaming Frog*), as can be seen in Table 7.

The next step was to use the QM to guide the composition of higher level qualities attributes (i.e., composite attributes). To that aim, the weights of the attributes were calculated before proceeding with the composition of parent attributes. Using the Expression 5 on the answer of the 7-point Likert scale, the respective weights were calculated and shown in Table 8. According to the experiments, the most important Usability sub-attribute was Efficiency, with 35.16% of the weight; for Accessibility it was Perceivable, with 68.12%; and for User Experience it was the Padlock, with 29.13%. All the weights were considered to complete the Interface QM. Following, some examples of the calculation are provided.

The Learnability attribute is composed of the sub-attributes Coherent Buttons, Coherent Menus, Navigation, and Easy of Learning. All of them were evaluated through questionnaires. To calculate the weights of the attributes, we must use Expression 5. Exemplifying the calculation with the Coherent Buttons attribute, 64 respondents scored it as 7 on the Likert scale. The total number of respondents who scored 7 for all the attributes in this group (i.e., the Learnability group) is 279. So, the coherent buttons weight will be:

$$\text{Weight} = 64 / 279 = 0.22939 = 22.94\%$$

The weights of the other attributes in this group are 32.97%, 27.60%, and 16.49%, respectively (always summing up 100% in each group). Then, considering the e-commerce A, the score of Learnability is obtained by the Expression 6:

$$\begin{aligned} \text{Score}_{\text{Learnability}} &= \text{Score}_{\text{CoherentButton}} * 0.2294 \\ &+ \text{Score}_{\text{CoherentMenu}} * 0.3297 \\ &+ \text{Score}_{\text{Navigation}} * 0.276 \\ &+ \text{Score}_{\text{EasyOfLearning}} * 0.1649 \\ &= 0.783 \end{aligned} \quad (6)$$

In other example, the attribute Efficiency in Use is composed of Environment Flexibility, General Flexibility, Responsiveness rate, and Performance page up. The General Flexibility and Environment Flexibility attributes are evaluated through questionnaires. Their weights were calculated as 36.75% and 38.41%, respectively. In the case of the attributes Responsiveness rate and Performance page up, they are evaluated using automatic tools and their weights are obtained through the specific questions in the questionnaire (average of importance expressed by the participants and not based on the Expression 5) as 12.91% and 11.92%, respectively. So, the score of Efficiency still considering e-commerce A is calculated as in the Expression 7:

Table 4. E-commerce websites - Average, Standard Deviation and Score - Experiment 1

Attributes Trustworthiness	e-commerce A			e-commerce B			e-commerce C		
	Average	Standard Deviation	Score	Average	Standard Deviation	Score	Average	Standard Deviation	Score
Coherent Buttons	6.255	1.355	0.876	5.980	1.276	0.830	4.176	2.112	0.529
Coherent Menus	5.314	1.703	0.719	5.539	1.696	0.757	3.775	2.053	0.462
Navigation	5.608	1.529	0.768	5.402	1.437	0.734	3.471	2.033	0.412
Easy of Learning	5.843	1.447	0.807	5.412	1.510	0.735	3.569	2.163	0.428
Usefulness	5.294	1.459	0.716	4.935	1.575	0.656	3.458	1.868	0.410
Failure Handling	4.392	1.344	0.565	4.314	1.132	0.552	3.949	1.275	0.492
Number of failures	0.206	—	0.794	0.208	—	0.792	0.472	—	0.528
User facing failures	—	—	0.863	—	—	0.843	—	—	0.765
Satisfaction	5.824	1.389	0.804	4.971	1.735	0.662	3.206	1.992	0.368
General Flexibility	4.850	1.908	0.642	4.674	2.034	0.612	4.195	2.288	0.533
Environment Flexibility	5.375	1.911	0.729	4.954	1.883	0.659	5.052	1.960	0.675
Responsive	—	—	1.000	—	—	1.000	—	—	1.000
Performance	—	—	0.917	—	—	0.817	—	—	0.486
Simple Screen	5.634	1.537	0.772	5.346	1.689	0.724	4.020	2.128	0.503
Color and Font	5.765	1.490	0.794	5.657	1.543	0.776	3.735	2.266	0.456
Visibility of System Status	5.582	1.549	0.764	5.529	1.473	0.755	4.052	1.953	0.509
AccessTx	—	—	0.738	—	—	0.689	—	—	0.798
Back Button	5.275	1.805	0.712	5.471	1.719	0.745	4.255	2.150	0.542
Broken Link	0.011	—	0.980	0.00	—	1.000	0.029	—	0.928
Visible Focus	5.922	1.426	0.820	6.020	1.321	0.837	4.667	1.917	0.611
Company Information	5.843	1.786	0.807	5.706	1.648	0.784	5.137	2.179	0.690
Company Reputation	4.402	2.315	0.567	6.069	1.308	0.845	3.402	1.567	0.400
Customers Opinion	6.020	1.590	0.837	5.902	1.485	0.817	5.196	1.951	0.699
Padlock	—	—	1.000	—	—	1.000	—	—	1.000
Pleasure	5.627	1.584	0.771	5.039	1.825	0.673	3.078	2.113	0.346
Privacy Policies	5.059	2.028	0.676	5.333	1.992	0.722	5.127	2.047	0.688

Table 5. Performance Measurements and Scores for e-commerce websites (Automatic Tools)

Website	Automatic Tools			Score
	PageSpeed Average	PingDom Average	GTMetrix Average	
e-comm A	88.55	95.55	91.00	0.917
e-comm B	76.77	91.77	76.44	0.817
e-comm C	29.22	83.55	33.11	0.486

Table 6. AccessTx Measurements and Scores for e-commerce websites (Automatic Tools)

Website	Automatic Tools			Score
	ASES %	Nibbler %	Access Monitor %	
e-commerce A	90.45	85.00	46.00	0.738
e-commerce B	89.72	not work	48.00	0.689
e-commerce C	93.39	82.00	64.00	0.798

$$\begin{aligned}
 Score_{EfficiencyInUse} &= Score_{GeneralFlexibility} * 0.3675 \\
 &+ Score_{Env.Flexibility} * 0.3841 \\
 &+ Score_{Responsive} * 0.1291 \\
 &+ Score_{Performance} * 0.1192) \\
 &= 0.753
 \end{aligned}$$

(7)

Table 7. Broken Links Measurements and Scores for e-commerce websites (Automatic Tools)

Website	Automatic Tools						Score
	Dead link Checker		Xenu's Link		Screaming Frog		
	Broken Links	Total	Broken Links	Total	Broken Links	Total	
e-com A	8	607	0	1	1406	69444	0.020
e-com B	0	1	1	1	33	165743	0.000
e-com C	145	2000	31	4272	38	6395	0.073

Considering that we were able to calculate all the parents of the leaf sub-attributes, and following the proposed QM, the next calculation, for example, computes the score of Usability. Usability sub-attribute is composed of Learnability, Usefulness, Safety in Use, Satisfaction and Efficiency. Similarly, the calculation of Usability Scores for the same website is as presented by Expression 8:

$$\begin{aligned}
 Score_{Usability} &= Score_{Learnability} * 0.3248 \\
 &+ Score_{Usefulness} * 0.0908 \\
 &+ Score_{SafetyInUse} * 0.1478 \\
 &+ Score_{Satisfaction} * 0.0850 \\
 &+ Score_{EfficiencyInUse} * 0.3516 \\
 &= 0.755
 \end{aligned}$$

(8)

Table 8. QM Sub-attributes Weights - Experiment 1

Attribute	Sub-attribute	Sub-attribute	Weight %
Usability 35%	Learnability 32.48%	Coherent Buttons	22.94%
		Coherent Menus	32.97%
		Navigation	27.60%
		Easy of Learning	16.49%
	Usefulness		9.08%
	Safety in Use 14.78%	Failure Handling	51.18%
		Number of Failures	24.41%
		Users Facing Failures	24.41%
	Satisfaction		8.50%
	Efficiency in Use 35.16%	General Flexibility	36.75%
		Environment Flexibility	38.41%
		Responsive	12.91%
Performance		11.92%	
Accessibility 35%	Perceivable 68.12%	Performance	8.87%
		Simple Screen	33.74%
		Color and Fonts	25.37%
		Visibility of System Status	32.02%
	Operable 31.88%	Access IX	30.53%
		Back Button	24.74%
		Broken Links	11.05%
		Visible Focus	33.68%
User Experience 30%	Company Information	14.17%	
	Company Reputation	12.43%	
	Customers Opinions	14.56%	
	Pleasure	7.38%	
	Padlock	29.13%	
	Privacy Policies	22.33%	

Table 9. Trustworthiness Scores for e-commerce websites – Experiment 1

Scores	e-comm A	e-comm B	e-comm C
Learnability	0.783	0.764	0.458
Usefulness	0.716	0.656	0.410
Safety in Use	0.694	0.682	0.567
Satisfaction	0.804	0.662	0.368
Efficiency in Use	0.753	0.705	0.642
Usability	0.755	0.712	0.527
Perceivable	0.788	0.755	0.491
Operable	0.786	0.787	0.686
Accessibility	0.787	0.765	0.554
User Experience	0.806	0.837	0.720
Interface Trustworthiness	0.782	0.768	0.594

Table 10. Questionnaires Results - AVG and SD

	AVG	SD
e-commerce A	5.467	1.640
e-commerce B	5.382	1.594
e-commerce C	4.080	2.001

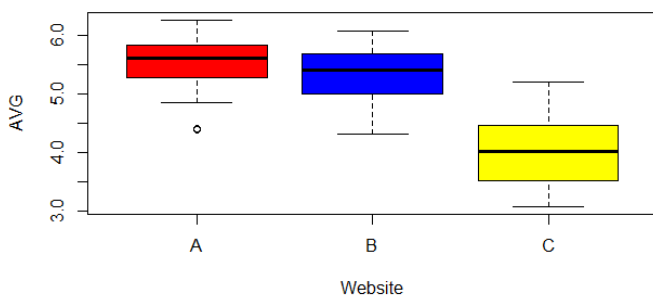


Figure 9. Statistical Analysis Experiment 1

The same logic applies to the qualities Accessibility and User Experience. Finally, the Interface Trustworthiness score is determined, which is made up of Usability, Acces-

Table 11. Shapiro-Wilk Test – Experiment 1

	e-commerce A	e-commerce B	e-commerce C
W	0.92453	0.95942	0.92391
p-value	0.1373	0.561	0.1336

sibility, and User Experience. Table 9 presents the highest level attributes of the Interface Quality Model.

Analyzing the results obtained in the first experiment, it can be seen that e-commerce A obtained the best interface confidence score (0.782), closely followed by e-commerce B (0.768), with the worst score (a difference of more than seventeen percentage points) being for e-commerce C (0.594). It is also noted that e-commerce A obtained better scores in the Usability and Accessibility attributes, while e-commerce B obtained better scores in the User Experience attribute. On the other hand, e-commerce C, obtained the lowest scores in all attributes when compared to e-commerces A and B. The only attribute that reached a score similar to the others (but even lower) was User Experience. Recovering the votes casted by the team’s researchers, it was noted that all researchers voted in the same order, and this order was coincident with what was observed as a result of the Experiment 1.

To verify the significance of the results obtained in this first experiment, a statistical analysis was performed. Firstly, the Shapiro-Wilk test was applied, and the result is that normality was met ($p > 0.05$ in Table 11) in all cases. Results were standardized on a 1-7 pt scale based on post-test questionnaire questions. For this analysis, the tool RSTUDIO¹³ was used. Table 10 presents the general results of the evaluation of the three websites of e-commerce. As observed in Table 10, the websites A and B obtained very close mean and standard deviation values, while website C obtained a worse average in relation to websites A and B. These differences can also be seen in Figure 9.

The homogeneity of variances was also evaluated using the Levene test ($df = 54$, F value = 1.4192, and $p = 0.2508$) accepting the null hypothesis, and therefore assuming that the variances are homogeneous. Thus, proving the normality and homogeneity of the variances, a normal distribution graph was made to observe the behavior of the groups. Therefore, the T test of the sites was elaborated. To perform this test, the following hypotheses were assumed: H_0 – The averages of the websites are equal; H_1 – The average of one of the websites is bigger than the others. With the results obtained, it can be concluded that the difference between the values of website A and website B are not statistically significant, and we can say that website A is equivalent to website B. Applying the T test for websites A - C and B - C, it can be concluded that in both cases the difference is statistically significant, with the average of site A being 34% higher than that of website C and the average of website B 31.9% higher than website C.

To complement this statistical analysis, a correlation analysis of some attributes of the Quality Model was also carried out. The correlation in statistics indicates the existence (or not) and the degree of dependence between two variables.

¹³<https://www.rstudio.com/categories/rstudio-ide/>

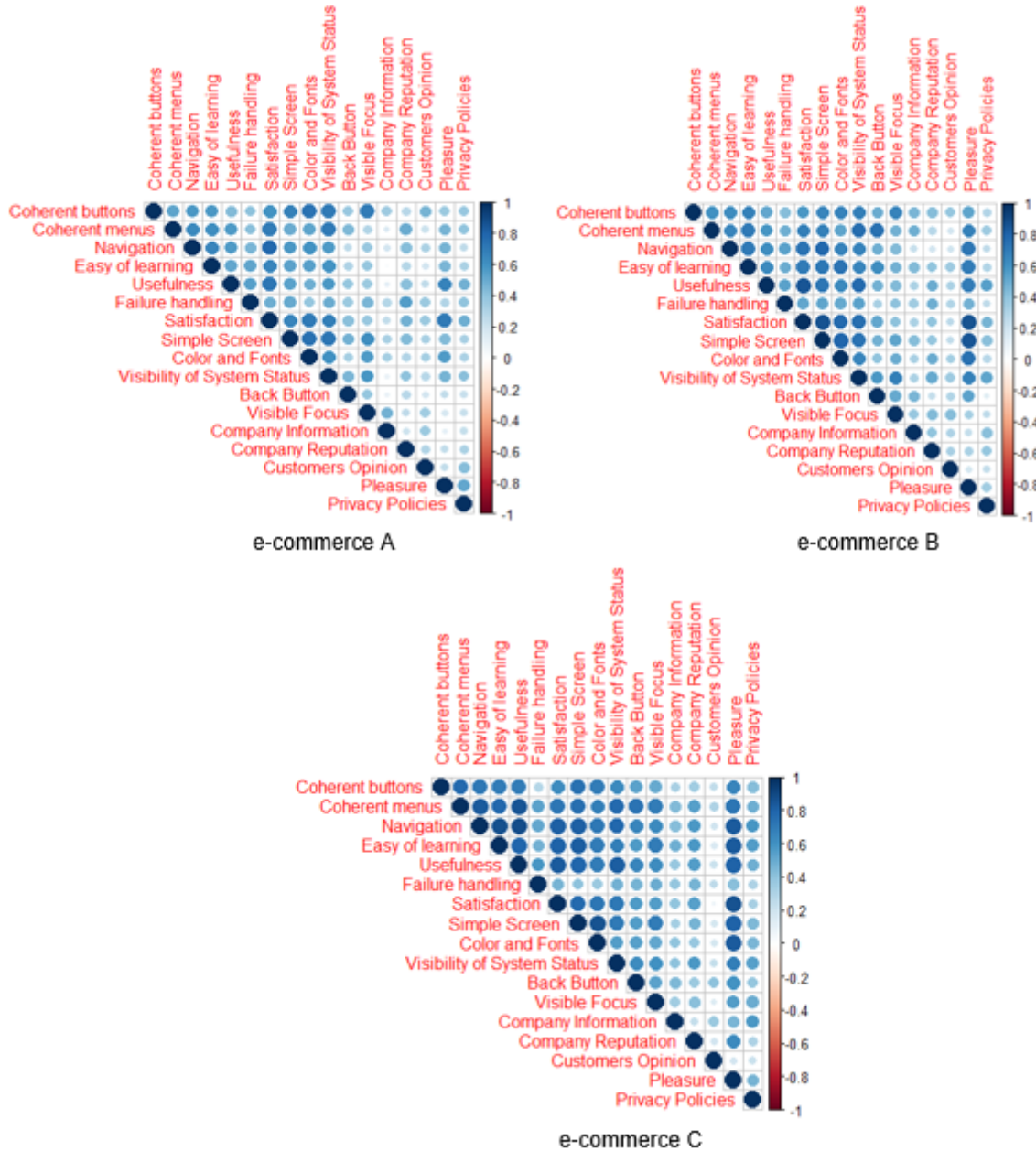


Figure 10. Correlation Matrix - Trustworthiness Attributes Experiment 1

Figure 10 presents the correlation matrix of the trustworthiness attributes of e-commerce websites A, B, and C. Each matrix cell shows the correlation between two attributes, and the darker and larger the circle, the more correlated are the two attributes (i.e. the correlation value is closer to 1).

Analyzing the matrix, it is noticed that the correlations in e-commerce C are more dense, presenting a highly correlated set composed of simple screen, navigation, usefulness, satisfaction, and pleasure. The same did not occur with the other two websites. Observing the reasons why it probably happens, we face the lack of quality in most of these attributes as the majority of the attributes were classified around 2 up to 4 on the Likert scale. The other two websites have more attributes that reach the higher Likert scale levels, and then they are more scattered and put down the correlation of some specific attributes. Even so, the correlation between navigation, simple screen and pleasure is still observed.

5.2.2 Improvements based on the First Experiments

This subsection presents an analysis carried out on the attributes that scored worst on Experiment 1 and also some suggestions for improvements to increase the scores of these attributes, thus increasing the feeling of trust that the user has. In this first experiment, website C obtained the worst scores based on the proposed methodology. The following attributes were analyzed: Navigation, Coherent Buttons, Coherent Menus, Visible Focus, Back Button, Company Information, and Privacy Policy. In addition to having received a low score, these attributes are visible in the interface of this website. Some attributes (such as Usefulness and User Satisfaction) were not selected because they are more subjective and depend on a user’s personal perception. Other attributes related to Safety in Use, in addition to not being visible in the interface, depend on technical decisions that would require a member of the development company’s team.

Considering Usability, we analyzed the sub-attributes Coherent Buttons, Coherent Menus and Navigation, which make up the sub-attribute Learnability.

Sub-attribute: Coherent Buttons (The image or text of the buttons matches the functionality they perform). Coherent Buttons is a sub-attribute that contributes to usability by making it easier to learn how to use the website. If the illustration is clearly related to the functionality, it is easier for the user to understand that it is the point of access to that functionality (in other words, you learn to use the website in a more intuitive way).

As an example, each of the websites presented different icons in the button that gives access to the list of items selected for purchase: Website A used a shopping cart, Website B used a shopping basket, and Website C used a shopping bag (see Figures 11, 12, and 13). Given that Website A received the highest score in this sub-attribute, it suggests that the shopping cart icon was easily connected to the idea of its functionality. The same cannot be said about the bag and basket icons.



Figure 11. Shopping cart button - website A

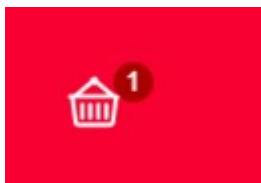


Figure 12. Shopping basket button - website B

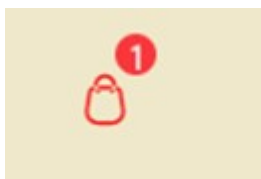


Figure 13. Shopping bag button - website C

We suggest that both website B and C change the button image (Figure 12 and 13) so that it is clearer to the user what its functionality is about (for example, considering a laptop user, avoiding the need to hover the mouse cursor over the figure to understand its associated functionality). There are several shopping cart vector graphics available royalty-free. Figure 14 presents a few to illustrate a possible redesign.



Figure 14. Redesign Suggestions - website B and C

Sub-attribute: Coherent Menus (Returning to the top level menu requires only one simple action and few steps;

Current position in website structure is provided). Coherent menus is a sub-attribute that facilitates usability by making it easier to navigate through the website.

In websites A and B, with just one click, it is possible to return to the previous menu, as can be seen in Figures 15 (“Computadores e Informática” menu) and 16 (“celulares e smartphones” menu). Conversely, website C neither provides the path to return to the previous menu nor does it provide the current position in the website structure. Figure 17 shows an example of when the “Explore” menu is selected. As one can see, a way back to the previous menu is not provided.

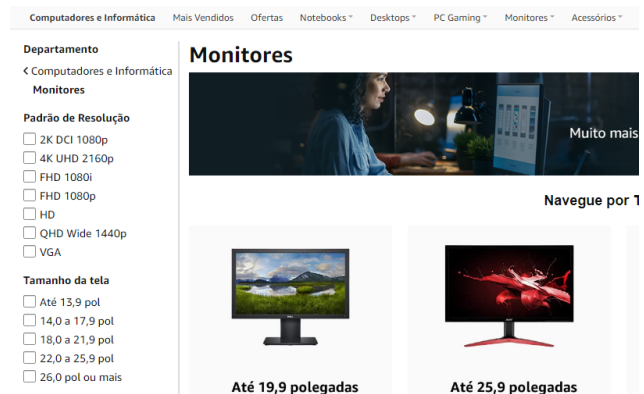


Figure 15. Website A - Menus



Figure 16. Website B - Menus

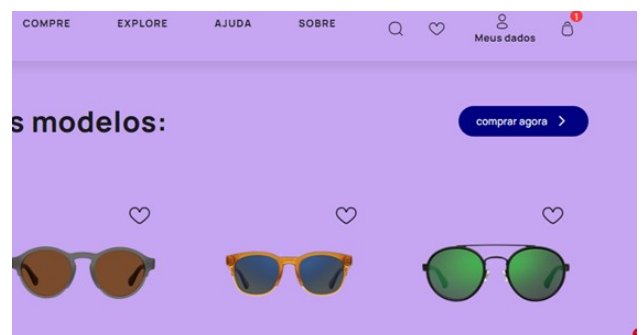


Figure 17. Website C - Menus

Our suggestion for website C is that every page has a link to return to the previously visited menu. In this way, the user will be able to know in which position of the website he is browsing and can return easily for previously visited pages.

Sub-attribute: Navigation (Provide maps of the website, which allow the user to visualize paths to follow). Navigation through the website is greatly facilitated when menus are provided, making available paths visible. This facilitates usability, since the user can have a view of the navigation possibilities.

Website A provides vertical and horizontal menus that facilitate the visualization of the paths to follow for the product purchase operation (see Figures 18 and 19). Besides, it shows at the top of the page the path to return to the initial page (Figure 19, right – link “Menu Principal” (“Main Menu” in English)). Website B also provides maps through the menus shown in Figures 20 and 21, which facilitate the tasks to be performed on the website. For website C, the home page menu does not provide a map of the website, which makes it difficult to navigate and find the product one want to buy. Information on each menu item is only shown at the bottom of the page. At several points, when activating the menu, no action takes place or an unwanted menu is presented. For example, when pressing the “Compre” (means “Buy” in English) button, but when using a laptop, the categories appeared (Figure 22), a wrong menu that refers to the Mother’s Day (“Dia das Mães” in Portuguese) promotion is presented (Figure 23), leaving the user confused.

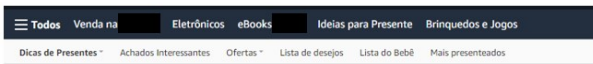


Figure 18. Horizontal Menu - website A

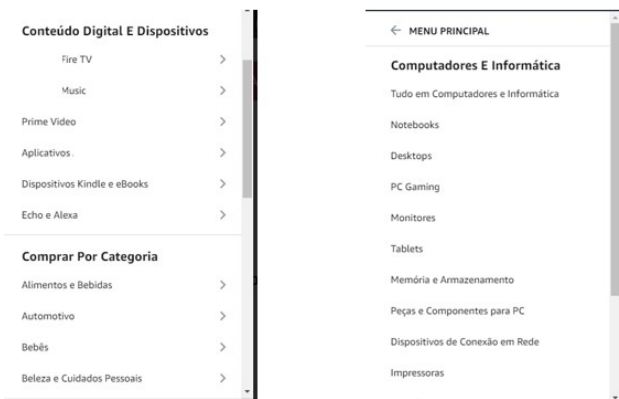


Figure 19. Vertical Menu - website A

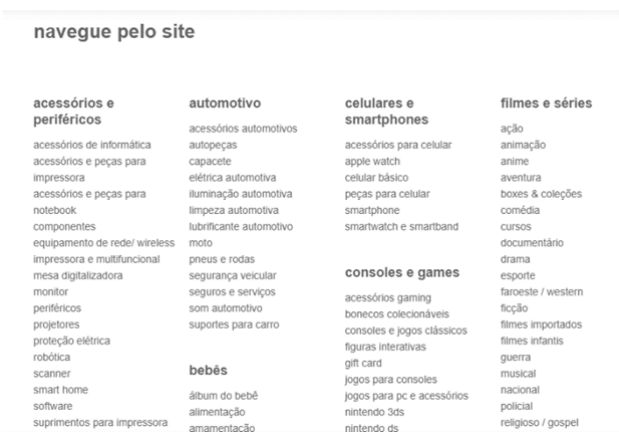


Figure 20. Horizontal Menu - website B

The suggestions for improving website C are: (i) construction of vertical menus that show a map of the website with categories and subcategories, making easier to locate the product one want to buy; (ii) handle page events so that the click of the mouse is obeyed and perceived by the user.

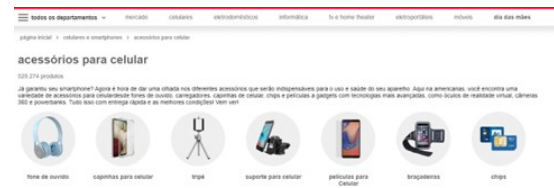


Figure 21. Vertical Menu - website B

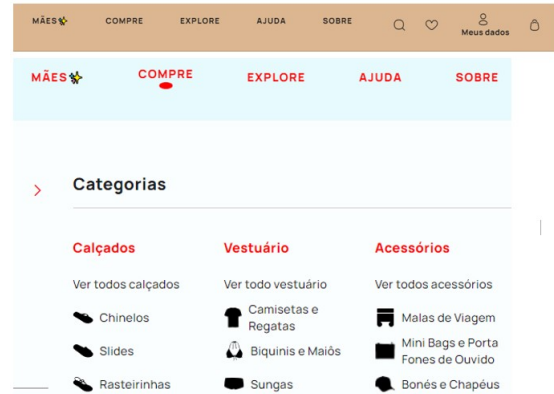


Figure 22. Menus - website C



Figure 23. Menus - website C

In the accessibility perspective, we analyzed the sub-attributes **Back Button** and **Visible Focus**, which make up the sub-attribute **Operable**.

Sub-attribute: Back Button (Most pages on the website allow users to return to the previously visited page.) Back Button facilitates the operation being carried out on the website, making it possible to quickly return to the previous page. This facilitates accessibility when using the website.

In the analysis carried out on website A, we did not find a Back Button on all the pages visited. To return to the previous page, it is necessary to use the navigation link; an example can be seen in Figure 24. The same happens with website B; in this website, one need to use the navigation buttons of the browser to go back to the previous page (Figure 25). Website C also does not provide a back button. Several pages were analyzed, but on none of them was possible to find a button to return to the previous page, being necessary to use the links that are at the top of the page (Figure 26).



Figure 24. Website A example

We suggest that the three websites analyzed add the back button on most pages. This button helps the user while browsing, making the website more operable, improving accessibility and consequently contributing to a better user experience,



Figure 25. Website B example

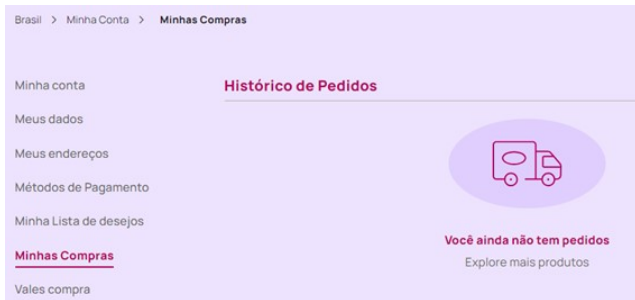


Figure 26. Website C example

which makes the website more reliable.

With regard to User Experience, we analyzed the sub-attributes **Company Information** and **Privacy Policy**.

Sub-attribute: Company Information (Company name, address, e-mail, and telephone are present on the website). The Company Information sub-attribute is a set of information that, once present, conveys confidence to the user, making the user feel more confident that the company is well established. This decreases uncertainty, improves the user experience, and automatically contributes to a better interface trust score. On the websites of the first experiment, it was verified that website A does not have the company's contact telephone information, and websites B and C do not have the contact email. We suggest that in these 3 websites, the following information be added on the first page: company name, address, e-mail, and telephone number. This information contributes to greater confidence in the website.

Sub-attribute: Privacy Policy (The website informs users about the use of cookies; Privacy policies are visible on the website). The Privacy Policy sub-attribute refers to information about the collection, storage, processing, and sharing of user data by websites, apps, and systems. It also highlights the need to inform the user about the destination of their data in different usage scenarios. These days, this information is important and needs to be present on any website, complying with the Brazilian General Data Protection Law (LGPD) and automatically contributing to a better user experience by increasing the level of trust in the interface.

On websites A and C, cookies and privacy policies are linked on the main page. On website B, this information does not appear on the first page in a visible way. The user needs to click on "information" for the privacy policy link to be displayed. We suggest that website B put a link on the first page so that the user can access information on privacy and cookie policies. It contributes to a better user experience and feeling of trust and automatically contributes to an increase in the trust score.

5.3 Methodology Validation - Second set of E-commerce websites

A second experiment was carried out to validate the approach, which uses a new set of three different websites. The websites chosen for this experiment were those identified by the previous experiment participants as the best and worst e-commerce websites (based on the questions inserted in the Experiment 1 questionnaire, i.e., "Which website have you ever used that you rank as the best?" and "Which website have you ever used and which do you classify as the worst?"). The responses were analyzed, and two websites were chosen from those most cited as the best and the most cited as the worst. All steps of the approach were redone, allowing a comparison of results.

5.3.1 Second Experiment Description and Results

One hundred and fifty-nine people between 18 and 60 years old performed tests to evaluate the e-commerce websites. The participants, who had no overlap with those who participated in the previous experiment (120 males and 39 females), performed the test from May 2, 2022, to June 17, 2022, and answered questionnaires about their experience using these websites. In terms of professional profile, 71% are not in the area of information technology and have never worked with user interfaces or computer systems; 21% have worked with computer systems, and 8% actively work with computer systems. The participants were split into two groups: 86 participants did the test on the three websites and complete the questionnaire for each evaluated website, which included questions about privacy policies, learnability, usefulness, satisfaction, among others; the remaining 73 participants completed the test using a website and several devices (smartphones or tablets and laptops) and browsers (Chrome, Firefox, Safari), and answered a questionnaire containing questions regarding general flexibility and environmental flexibility.

Based on the answers obtained through the questionnaires, the average, the standard deviation and the score were calculated for each attribute composing the Quality Model. Table 12 presents these calculations for the three evaluated e-commerce websites.

Again, to ensure the companies' privacy, the presentation of the results does not name them and the e-commerce websites are referred to in the text as *e-commerce A*, *B* and *C*, in no particular order. As can be seen, for *e-commerce A*, the Environment Flexibility attribute showed the best values for mean (6.360), and score (0.893); standard deviation (SD) = 1.145. For *e-commerce B*, the Visible Focus attribute had the best values for mean (6.151), and score (0.859); standard deviation (SD) = 1.244. For *e-commerce C*, the best values were presented for the Customer Opinions attribute (mean 6.221, standard deviation 0.969 and score 0.870). Excluding the attributes whose responses are binary (responsiveness and padlock) the highest score for all three websites was for the broken link attribute, meaning that the websites have practically no broken links.

The less subjective attributes that were evaluated using automatic tools (performance page up, AccessTx rate, broken links, and responsive rate) are presented in Tables 13,

Table 12. E-commerce websites - Average, Standard Deviation and Score - Experiment 2

Attributes Trustworthiness	e-commerce A			e-commerce B			e-commerce C		
	Average	Standard Deviation	Score	Average	Standard Deviation	Score	Average	Standard Deviation	Score
Coherent buttons	6.326	1.252	0.888	6.081	1.383	0.847	5.965	1.401	0.828
Coherent Menus	5.488	1.854	0.748	5.895	1.552	0.816	5.680	1.606	0.780
Navigation	5.587	1.470	0.765	5.581	1.663	0.764	5.384	1.686	0.731
Easy of Learning	6.070	1.179	0.845	5.767	1.590	0.795	5.302	1.685	0.717
Usefulness	5.407	1.605	0.734	5.174	1.677	0.696	5.174	1.658	0.696
Failure handling	4.293	1.486	0.549	4.416	1.439	0.569	4.370	1.321	0.562
Number of failures	0.154	—	0.846	0.153	—	0.847	0.330	—	0.670
User facing failures	—	—	0.628	—	—	0.674	—	—	0.686
Satisfaction	5.494	1.764	0.749	4.965	1.858	0.661	4.977	1.791	0.663
General Flexibility	5.085	1.892	0.681	4.736	1.949	0.623	4.886	1.840	0.648
Environment Flexibility	6.360	1.145	0.893	5.808	1.301	0.801	5.955	1.278	0.826
Responsive	—	—	1.000	—	—	1.000	—	—	1.000
Performance	—	—	0.936	—	—	0.579	—	—	0.491
Simple Screen	5.961	1.488	0.827	5.647	1.735	0.775	5.267	1.870	0.711
Color and Font	6.244	1.289	0.874	5.814	1.533	0.802	5.605	1.662	0.767
Visibility of System Status	5.640	1.672	0.773	5.736	1.535	0.789	5.504	1.588	0.751
AccessTx	—	—	0.605	—	—	0.734	—	—	0.530
Back Button	5.198	2.011	0.700	5.616	1.658	0.769	5.756	1.509	0.793
Broken Link	0.001	—	0.998	0.001	—	0.998	0.002	—	0.994
Visible Focus	6.198	1.362	0.866	6,151	1.244	0.859	5.953	1.539	0.826
Company Information	5.628	1.692	0.771	6.093	1.291	0.849	4.616	2.141	0.603
Company Reputation	3.192	1.740	0.365	6.087	1.333	0.848	3.198	2.076	0.366
Customers Opinion	5.919	1.314	0.820	5.977	1.414	0.829	6.221	0.969	0.870
Padlock	—	—	1.000	—	—	1.000	—	—	1.000
Pleasure	5.779	1.324	0.797	5.267	1.788	0.711	4.733	1.845	0.622
Privacy Policies	5.692	1.850	0.782	5.895	1.801	0.816	4.372	2.313	0.562

14 and 15 respectively, for page loading performance, AccessTx rate, and broken links. Regarding the responsive rate attribute, all *e-commerce websites* used in this second experiment are considered responsive (scored as 1).

Table 13. Measurement and Scores of Performance Page Up for e-commerce, based on automatics tools - Experiment 2

Website	Automatic Tools			Score
	PageSpeed AVG	PingDom AVG	GTMetrix AVG	
e-comm A	91.77	95.88	93.11	0.936
e-comm B	64.88	61.00	47.88	0.579
e-comm C	44.33	73.33	29.66	0.491

Also, the weights of each of the leaf attributes were calculated and they are coincident to the ones used in Experiment 1, which are presented in Table 8. Also in a similar way, the composite attributes' scores were calculated (presented in Table 16) following the Quality Model hierarchy. The means were presented in Table 12, and the weights were presented in Table 8.

Analyzing the results obtained in the second experiment

Table 14. Measurement and Scores of AccessTx, based on automatic tools - Experiment 2

Website	Automatic Tools			Score
	ASES %	Nibbler %	Access Monitor %	
e-commerce A	75.03	46.00	não avaliado	0.605
e-commerce B	95.29	86.00	39.00	0.734
e-commerce C	66.07	57.00	36.00	0.530

Table 15. Measurement and Scores of broken links for e-commerce, based on automatic tools - Experiment 2

Website	Automatic Tools						Score
	Dead link Checker		Xenu's Link		Screaming Frog		
	Broken Links	Total	Broken Links	Total	Broken Links	Total	
e-com A	4	2000	34	24512	0	295605	0.002
e-com B	2	2000	0	1	2	1138	0.002
e-com C	1	181	1	1	0	15	0.006

and presented in Table 16, considering the interface trustworthiness, there is practically a tie between e-commerces A (score 0.795) and B (score 0.794), and the worst score

Table 16. Trustworthiness Scores for websites of the e-commerce - Experiment 2

Scores	e-comm A	e-comm B	e-comm C
Learnability	0.801	0.805	0.767
Usefulness	0.734	0.696	0.696
Safety in Use	0.641	0.663	0.618
Satisfaction	0.749	0.661	0.663
Efficiency in Use	0.834	0.735	0.743
Usability	0.778	0.737	0.721
Perceivable	0.831	0.769	0.719
Operable	0.760	0.814	0.746
Accessibility	0.809	0.783	0.727
User Experience	0.799	0.872	0.720
Interface Trustworthiness	0.795	0.794	0.723

was assigned to e-commerce C (score 0.723). E-commerce B scored better on the User Experience attribute, while e-commerce A scored better on the Usability and Accessibility attributes. E-commerce A (slightly higher) and B divided the best scores almost equally. Contrarily, E-commerce C obtained worse scores in all attributes, which justifies the worst interface trustworthiness score.

The statistical analysis of the questionnaire results is presented in Table 17, which shows the general results for the three e-commerce websites. Results were standardized on a 0-7 pt scale based on post-test questionnaire questions. The website of e-commerce A obtained the best average, followed by e-commerce B, with a very small difference. E-commerce C, on the other hand, obtained a worse average in relation to websites A and B, which can be seen in Figure 27.

Table 17. Questionnaires Results - average (AVG) and standard deviation (SD)

	AVG	SD
e-commerce A	5.556	1.547
e-commerce B	5.616	1.565
e-commerce C	5.206	1.673

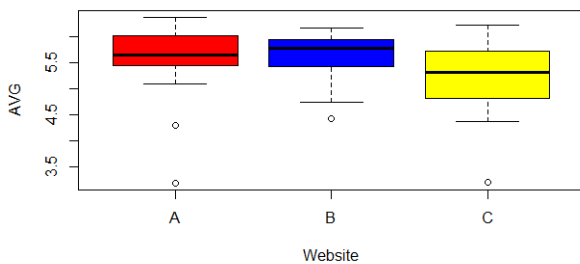


Figure 27. Experiment 2 Boxplot

The Shapiro-Wilk test was applied to check the normality of the data groups. As seen in Table 18, the p-value of website C is greater than 0.05 ($p > 0.05$); therefore, it is assumed that the data have a normal distribution. Opposite to that, for websites A and B, the p-values are not greater than 0.05 ($p < 0.05$), therefore the data for websites A and B does not have a normal distribution. As the normal distribution of data only occurred with e-commerce C, the Mann-Whitney Test was performed, as presented in Table 19.

Table 18. Shapiro-Wilk Test - Experiment 2

	e-commerce A	e-commerce B	e-commerce C
W	0.81708	0.87485	0.92809
p-value	0.002046	0.01749	0.1597

Table 19. Mann-Whitney Test - Experiment 2

	e-commerces A / B	e-commerces A / C	e-commerces B / C
W	179	241	247
p-value	0.9767	0.07983	0.05396

The result of the Mann-Whitney test for e-commerces A, B, and C showed that the p-values obtained were greater than 0.05. Therefore, it can be concluded that the difference between them is not significant. Another important detail is that the Mann-Whitney test also verifies the median of the data group. The test showed that the median of e-commerce A (median = 5.64) is considered equal to e-commerce B (median = 5.77), and the same happened with the test between e-commerces A and C, the median of e-commerce C (median = 5.30). Therefore, the difference between the medians of e-commerces A, B, and C is not statistically significant.

To complement the statistical analysis, a correlation matrix was created between the attributes that were measured through questionnaires. Figure 28 shows the correlation of the trust attributes of the interfaces of e-commerces A, B, and C. When the correlation matrix is examined, it is possible to observe that there is a stronger correlation between the attributes utility and satisfaction, ease of learning and satisfaction, simple screens and pleasure in e-commerces A and B. In e-commerce C, there is a strong correlation between the attributes navigation, usefulness, and satisfaction, and between the attributes ease of learning and pleasure. However, a highly correlated set is not observed here as observed in Experiment 1.

5.3.2 Improvements based on the Second Experiment

In the second experiment, the three websites obtained better scores compared to the websites of Experiment 1, and the interface trust score obtained using the proposed methodology was very similar for the three websites. We focused on the analysis of the following attributes: Navigation, Company Information, and Privacy Policy, as they were the ones that presented the lowest scores among the attributes that do not require technical knowledge of the website.

Sub-attribute: Navigation (Provides maps of the website, allowing you to visualize paths to follow). As mentioned before, when menus are provided with visible paths, navigation is facilitated and, consequently, usability is improved.

Website A has vertical (categories) and horizontal (sub-categories) menus that help the user navigate the website to find the product they want to buy. Browsing the searched product is simple and easy. Even so, due to the fact that there are many products to be sold, an improvement would be to transform the horizontal menu into a vertical one to facilitate the location of the various sub-categories (Figure 29).

On website B, the department menu helps the user find the product he wants to buy. It has a more complete vertical

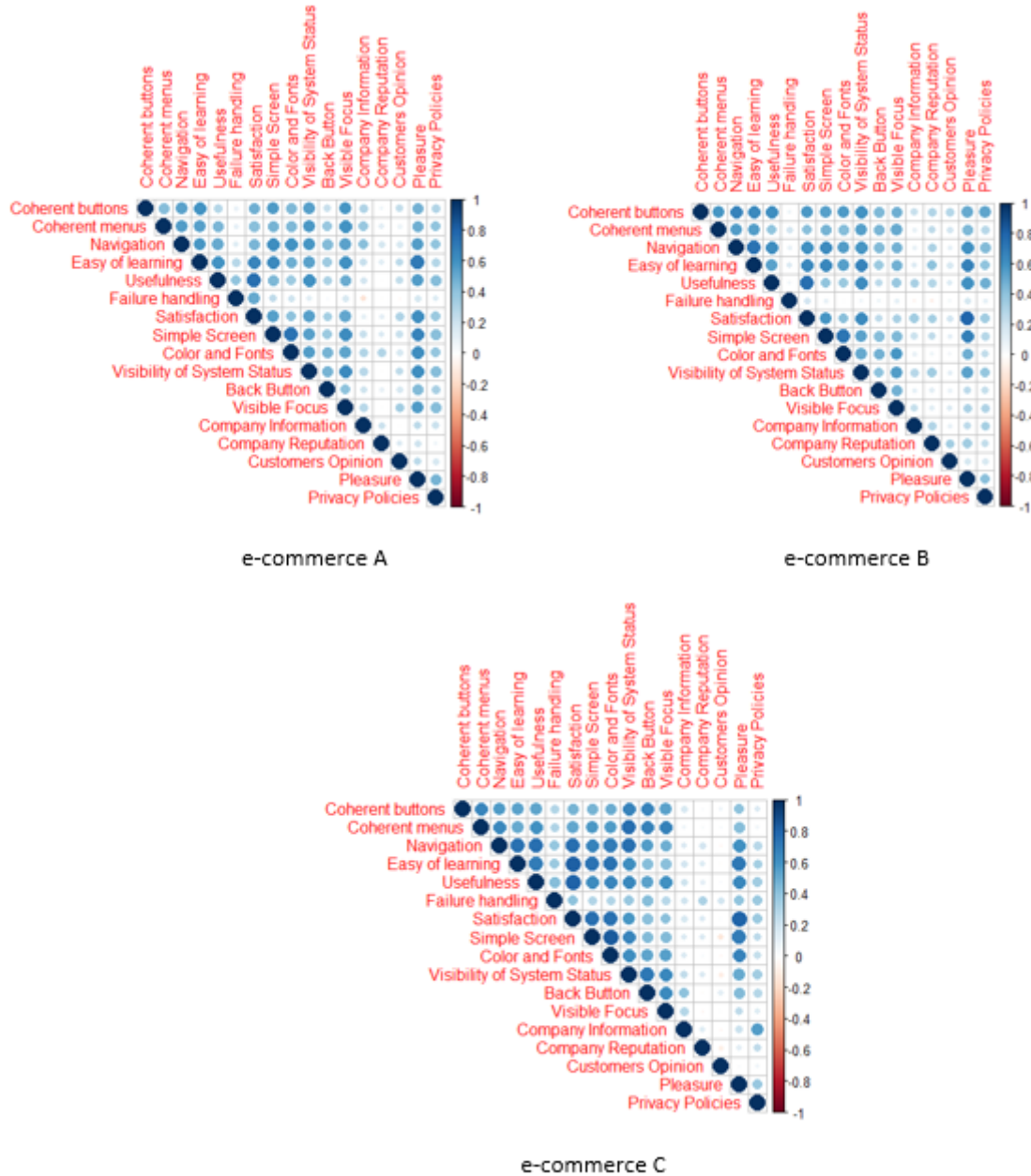


Figure 28. Correlation Matrix - Trustworthiness Attributes Experiment 2

menu than the one presented on website A. The search for the product is also simple and brings exactly the product that the user wants (Figure 30).

Website C does not have a menu by category or department; it only has a horizontal menu and pictures with links in the body of the page. When you click on a category, several subcategories appear. The navigation option confuses the search, not showing an objective navigation (Figure 31).

The search on the main page brings all related products; that is, it does not return exactly what was typed in the search field. We suggest that website C improves its navigation by creating vertical menus with categories and subcategories of products, and in the search for the product, bring only the product that was chosen. This facilitates the usability of the site, making learning easier.

In the user experience part, we analyzed the sub-attributes **Company Information** (Company name, **Privacy Policy**.

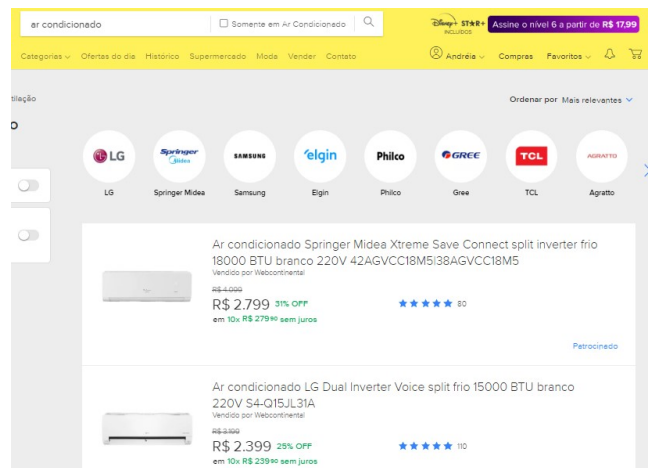


Figure 29. Searched product - website A

Sub-attribute: Company Information (Company name,



Figure 30. Searched product - website B

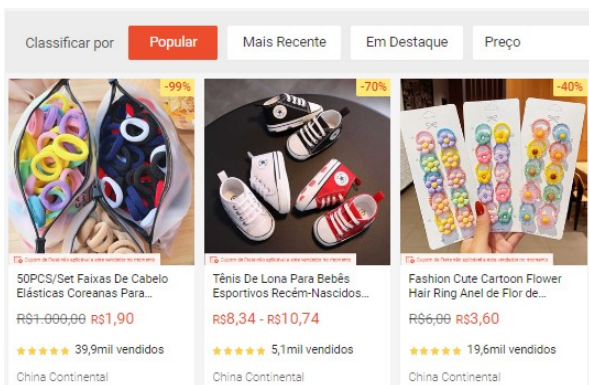


Figure 31. Horizontal Menu - website C

address, e-mail, and telephone are present on the website).

On websites A and B, it was verified that there is no information about telephone or email contact. On the other hand, website C does not contain an address, telephone number, or email for contact. We suggest that on the three websites, this information be added on the first page (address, e-mail, and telephone), as it contributes to the user having more confidence in the website being used.

Sub-attribute: Privacy Policy (The website informs about the use of cookies / privacy policies). As mentioned before, the Privacy Policy sub-attribute automatically contributes to a better user experience, increasing the trust score.

The three websites evaluated in this second experiment have a link to the privacy policy at the end of the first page for the user to consult and find out how their data will be used and treated. However, website C was worse evaluated in this regard by the research participants. We suggest that the website use a larger font and highlight the link to the privacy policy information.

5.4 Methodology Validation - Bank Websites

The goal of the last experiment is to verify if the methodology can be applied to another system context. Although still focused on web interfaces, bank websites give us an idea of how generic the methodology can be. Banking systems require a higher degree of dependability, security and privacy, and the idea is to verify what the impact of these characteristics is on the applicability of the methodology, as well as on the users' perception of trust. The details of this experiment

can be found in a previous work of the same group [Casare *et al.*, 2022a]. We included here only the most important details to complement the findings and conclusions.

Table 20 presents the *Usability*, *Accessibility*, *User Experience* and *Interface Trustworthiness* scores of the Interface QM, which are calculated using the same expressions presented before (Expression 4).

Table 20. Trustworthiness score obtained for on-line bank websites

Scores	Bank 1	Bank 2	Bank 3
Usability	0.573	0.704	0.661
Accessibility	0.563	0.583	0.549
UserExperience	0.782	0.833	0.757
InterfaceTrustworthiness	0.632	0.706	0.650

When these scores were computed, Bank 2 had the highest Interface Trustworthiness score (0.706), followed by Bank 3 (0.650) and Bank 1 (0.632), which had the lowest trustworthiness. Bank 2 really had the highest score for all three level-2 attributes (Usability, Accessibility, and User Experience), albeit somewhat higher in some cases. Bank 3 *Accessibility* attribute has the lowest score (0.549), whereas Bank 2 *User Experience* has the highest score (0.833).

Through the use case of three online bank websites, it was possible to see that the approach is applicable to online systems that require a higher level of dependability. A result to be highlighted is that we can observe that the weights of the attributes vary a bit, mainly the ones related to safety in use (which include failures). We considered that this occurs due to the business context, which is diverse than on the previous websites (in Experiments 1 and 2), making the difference of importance of some attributes. The proposed mechanism (i.e., the Interface Quality Model) is evident in its importance, as it allows developers to analyze the shortcomings of each feature and enhance the interface depending on the unique requirements. For example, all three websites have ratings lower than 0.8. The interface designers must then concentrate on all of the Usability sub characteristics, as they all have scores below 0.7, with efficiency providing a particularly low score (0.24). Even Accessibility and User Experience had sub-attributes with poor scores, showing the necessity to modify the interface in order to achieve a more efficient, pleasurable, and high-quality system.

6 Conclusions and Future Work

This work presented the validation of a model (Interface Quality Model) to support user interface quality measurement and analysis, with a particular focus on the impact of the user interface in the trustworthiness of a system (and also in its trustworthiness score).

The proposed model comprises a set of metrics related to a given website under analysis. It defines twenty-two metrics based on the answers of questionnaires by users, and four metrics obtained by the use of automatic tools. These metrics are grouped in a hierarchical way, and scores are computed for each node of this hierarchy. The highest level contains the main score of this model (the interface trustworthiness score), and the second level contains the three main divisions

of our model, related to usability, accessibility, and user experience. Besides, the set of computed scores of the model indicates how each attribute (and sub-attribute) is being properly considered in the evaluated website; therefore, the worst scores may be an starting point for a user interface team to improve the website under assessment.

The model was evaluated in three experiments. This paper focused on two of them, related to the assessment of two sets of e-commerce websites. The scores computed by the model in each experiment for each website are directly related to the votes given by participants regarding how trustworthy these websites are. Some interface problems related to attributes (or sub-attributes) with low scores were also highlighted and discussed. Besides, the paper presented a third experiment, in which the model was used to evaluate bank websites. In this case, it was noted that the quality of the websites is more homogeneous, and the scores informed by the model also reflected this perception. These results provide evidences that the use of this model can help identify potential problems related to the user interface that affect the trustworthiness of an e-commerce system, and can also be used (with adaptations) to evaluate other types of systems.

A new revision of the set of attributes will be done, updating the new publications in the literature, and then a more extensive validation with users is planned, which should present more complete results in future work. Also, the evaluation of new versions of tools for calculating objective attributes is planned, considering the low accuracy that we observed during the experiments. If necessary, we will consider using paid tools if we cannot achieve better convergence between the measurements obtained by the set of tools focusing on the same attribute. It is also intended to continue building the trustworthiness calculation and visualization tool by making available routines to register Quality Models and save them in a database. The idea is to use the MYSQL DBMS, as it is free software like the others used in building the prototype of this tool.

Another goal is to compose this interface quality model with other models that have been developed by other researchers for infrastructure and data storage in order to obtain a score that reflects all aspects of a system. However, this can only be done within the scope of the website owner, as it is necessary to know the details of the implementation and deployment of the system under analysis to compose the values of the more technical models.

Declarations

Acknowledgements

This work was financed by CAPES - Coord. de Aperfeiçoamento de Pessoal de Nível Superior - Brasil, finance code 001. Also, it was funded by the FCT - Foundation for Science and Technology, within the scope of CISUC R&D Unit - UIDB/00326/2020 and ADVANCE projects (<http://advance-rise.eu>).

Authors' Contributions

Casare Andreia participates in the investigation, artefacts' creation and formalization, methodology application in experiments, analy-

sis and interface improvements and results statistical analysis. Silva Celmar participates in the solution's discussion and metrics definition, investigation, result analysis, and supervision. Moraes Regina participates in the conceptualization, project administration, formal description, investigation, and supervision. All authors wrote part of original draft, read and approved the final manuscript.

Availability of data and materials

The datasets generated and/or analysed during the current study are available in the research group's website.

References

- Al-Azzawi, G., Miskon, S., Abdullah, N., and Mat Ali, N. (2021). Factors influencing customers' trust in ecommerce during covid-19 pandemic. In *7th International Conference on Research and Innovation in Information Systems -ICRIIS*. DOI: <https://doi.org/10.1109/ICRIIS53035.2021.9617021>.
- Albert, B. (2012). How quick are we to judge? a case study of trust and web site design. <https://pt.slideshare.net/NYTechCouncil/>, Accessed: 16 January 2024.
- Aljazzaf, Z. M., Perry, M., and Capretz, M. A. (2010). Online trust: Definition and principles. In *2010 Fifth International Multi-conference on Computing in the Global Information Technology*, pages 163–168. IEEE. DOI: <https://doi.org/10.1109/ICCGI.2010.17>.
- Ang, L. and Lee, B.-C. (2000). Influencing perceptions of trustworthiness in internet commerce: A rational choice framework. In *Fifth COLLECTer Conference on Electronic Commerce*, pages 1–12. COLLECTeR.
- Attar, R., Shanmugam, M., and Hajli, N. (2020). Investigating the antecedents of e-commerce satisfaction in social commerce context. *British Food Journal*, 123:849–868. DOI: <https://doi.org/10.1108/BFJ-08-2020-0755>.
- Bangor, A., Kortum, P., and Miller, J. (2009). Determining what individual sus scores mean: Adding an adjective rating scale. *Journal of usability studies*, 4(3):114–123.
- Basso, A., Goldberg, D., Greenspan, S., and Weimer, D. (2001). First impressions: Emotional and cognitive factors underlying judgments of trust e-commerce. In *Proceedings of the 3rd ACM Conference on Electronic Commerce*, pages 137–143. DOI: <https://doi.org/10.1145/501158.501173>.
- Bauer, P. C. (2019). Conceptualizing trust and trustworthiness. Technical report, Revised version of working paper published in: Political Concepts Working Paper Series.
- Bonisoli, L. and Castillo Leyva, K. (2022). Regulatory beliefs and trust: Adaptation of the technological acceptance model to e-commerce during the covid-19 pandemic in ecuador. *Innovar*, 32:135–149. DOI: <https://doi.org/10.15446/innovar.v32n86.104666>.
- Brooke, J. (1996). Sus: a “quick and dirty” usability. *Usability evaluation in industry*, 189(3):189–194.
- Casare, A., Basso, T., and Moraes, R. (2020). Trust metrics to measure website user experience. In *The 13th Int. Conf. on Advances in Computer-Human Interactions*, pages 1–8.

- Casare, A., da Silva, C. G., Basso, T., and Moraes, R. (2021). Towards usability interface trustworthiness in e-commerce systems. In *15th Int. Conf. on Interfaces and Human Computer Interaction*, pages 1–8.
- Casare, A., Silva, C. G., and Moraes, R. (2022a). Do dependable systems need good user interfaces? In *Proceedings of the 11th Latin-American Symposium on Dependable Computing*, pages 21–28. DOI: <https://doi.org/10.1145/3569902.3569905>.
- Casare, A. R., Basso, T., da Silva, C. G., and Moraes, R. (2022b). Using a quality model to evaluate user interface trustworthiness of e-commerce systems: Scoring strategies and preliminary results. In *VISIGRAPP (2: HUCAPP)*, pages 209–216. DOI: <https://doi.org/10.5220/0010889700003124>.
- Chen, J. and Dibb, S. (2010). Consumer trust in the online retail context: Exploring the antecedents and consequences. *Psychology & Marketing*, 27(4):323–346. DOI: <https://doi.org/10.1002/mar.20334>.
- Cho, J.-H., Chan, K., and Adali, S. (2015). A survey on trust modeling. *ACM Computing Surveys (CSUR)*, 48(2):1–40. DOI: <https://doi.org/10.1145/2815595>.
- Cho, J.-H., Xu, S., Hurley, P. M., Mackay, M., Benjamin, T., and Beaumont, M. (2019). Stram: Measuring the trustworthiness of computer-based systems. *ACM Computing Surveys (CSUR)*, 51(6):1–47. DOI: <https://doi.org/10.1145/3277666>.
- Dujmovic, J. J. (2007). Continuous preference logic for system evaluation. *IEEE Transactions on fuzzy systems*, 15(6):1082–1099. DOI: <https://doi.org/10.1109/TFUZZ.2007.902041>.
- Filho, G. K., Guerino, G., and Valentim, N. (2022). A systematic mapping study on usability and user experience evaluation of multi-touch systems. In *Anais do XXI Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais*. SBC. DOI: <https://doi.org/10.1145/3554364.3559131>.
- Finstad, K. (2010). The usability metric for user experience. *Interacting with computers*, 22(5):323–327. DOI: <https://doi.org/10.1016/j.intcom.2010.04.004>.
- Friginal, J., Martínez, M., De Andrés, D., and Ruiz, J.-C. (2016). Multi-criteria analysis of measures in benchmarking: Dependability benchmarking as a case study. *Journal of Systems and Software*, 111:105–118. DOI: <https://doi.org/10.1016/j.jss.2015.08.052>.
- Gao, X., Ma, Y., and Zhou, W. (2021). The trustworthiness measurement model of component-based software based on the subjective and objective weight allocation method. In *2021 IEEE 21st International Conference on Software Quality, Reliability and Security Companion (QRS-C)*, pages 478–486. IEEE. DOI: <https://doi.org/10.1109/QRS-C55045.2021.00076>.
- Gol Mohammadi, N., Paulus, S., Bishr, M., Metzger, A., Könnecke, H., Hartenstein, S., Weyer, T., and Pohl, K. (2014). Trustworthiness attributes and metrics for engineering trusted internet-based software systems. In *Cloud Computing and Services Science: Third International Conference, CLOSER 2013, Aachen, Germany, May 8-10, 2013, Revised Selected Papers 3*, pages 19–35. Springer. DOI: https://doi.org/10.1007/978-3-319-11561-0_2.
- Granollers, T. (2018). Usability evaluation with heuristics: new proposal from integrating two trusted sources. In *Design, User Experience, and Usability: Theory and Practice: 7th International Conference, DUXU 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15-20, 2018, Proceedings, Part I 7*, pages 396–405. Springer. DOI: https://doi.org/10.1007/978-3-319-91797-9_28.
- Guerino, G. and Valentim, N. (2020). Usability and user experience evaluation of conversational systems: A systematic mapping study. In *Anais do XXXIV Simpósio Brasileiro de Engenharia de Software*. SBC. DOI: <https://doi.org/10.1145/3422392.3422421>.
- Habib, S., Hamadneh, N. N., and Hassan, A. (2022). The relationship between digital marketing, customer engagement, and purchase intention via ott platforms. *Journal of Mathematics, Special Issue - Analysis of Financial Problems Based on Mathematical Models*, 2022. DOI: <https://doi.org/10.1155/2022/5327626>.
- Hendradjaya, B. and Praptini, R. (2015). A proposal for a quality model for e-government website. In *2015 International Conference on Data and Software Engineering (ICoDSE)*, pages 19–24. IEEE. DOI: <https://doi.org/10.1109/ICODSE.2015.7436965>.
- Henry, S., participants of the Education, and (EOWG), O. W. G. (2023). Introduction to web accessibility. <https://www.w3.org/WAI/fundamentals/accessibility-intro/#context>, Accessed: 16 January 2024.
- Hussain, F. K. and Chang, E. (2007). An overview of the interpretations of trust and reputation. In *The Third Advanced International Conference on Telecommunications (AICT'07)*, pages 30–30. IEEE. DOI: <https://doi.org/10.1109/AICT.2007.11>.
- Hussin, A. R. C., Macaulay, L., and Keeling, K. (2007). The importance ranking of trust attributes in e-commerce website. *PACIS 2007 Proceedings*, page 99.
- ISO (2014). International organization for standardization. when the world agrees (ISO/IEC).
- ISO (2016a). Systems and software engineering - systems and software quality requirements and evaluation (square) - measurement of quality in use (ISO/IEC).
- ISO (2016b). Systems and software engineering - systems and software quality requirements and evaluation (square) - measurement of system and software product quality (ISO/IEC).
- ISO (2019). Ergonomics of human system interaction — part 210: Human-centred design for interactive systems.
- ITA Publishing (2024). Impact of covid pandemic on ecommerce. <https://www.trade.gov/impact-covid-pandemic-ecommerce>, 16 January 2024.
- Jiménez, D. L., Dittmar, E. C., and Portillo, J. P. V. (2021). The use of trust seals in european and latin american commercial transactions. *Journal of Open Innovation: Technology, Market, and Complexity*, 7(2):150. DOI: <https://doi.org/10.3390/joitmc7020150>.
- John Joseph, A. J. and Mariappan, M. (2018). A

- novel trust-scoring system using trustability coefficient of variation for identification of secure agent platforms. *PloS one*, 13(8):e0201600. DOI: <https://doi.org/10.1371/journal.pone.0201600>.
- Lew, P., Olsina, L., and Zhang, L. (2010). Integrating quality, quality in use, actual usability and user experience. In *2010 6th Central and Eastern European Software Engineering Conference (CEE-SECR)*, pages 117–123. IEEE. DOI: <https://doi.org/10.1109/CEE-SECR.2010.5783161>.
- Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, 7(1):57–78. DOI: <https://doi.org/10.1080/10447319509526110>.
- Lewis, J. R., Utesch, B. S., and Maher, D. E. (2013). Umux-lite: when there's no time for the sus. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2099–2102. DOI: <https://doi.org/10.1145/2470654.2481287>.
- Lin, H. X., Choong, Y.-Y., and Salvendy, G. (1997). A proposed index of usability: a method for comparing the relative usability of different software systems. *Behaviour & information technology*, 16(4-5):267–277. DOI: <https://doi.org/10.1080/014492997119833>.
- Lund, A. M. (2001). Measuring usability with the use questionnaire. *Usability interface*, 8(2):3–6.
- Malhotra, R. (2016). *Empirical research in software engineering: concepts, analysis, and applications*. CRC press.
- Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of management review*, 20(3):709–734. DOI: <https://doi.org/10.5465/amr.1995.9508080335>.
- McKibbin, W. and Fernando, R. (2021). The global macroeconomic impacts of covid-19: Seven scenarios. *Asian Economic Papers*, 20(2):1–30. DOI: https://doi.org/10.1162/asep_a_00796.
- McKnight, D. H., Choudhury, V., and Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information systems research*, 13(3):334–359. DOI: <https://doi.org/10.1287/isre.13.3.334.81>.
- Mohammadi, N. G., Paulus, S., Bishr, M., Metzger, A., Koennecke, H., Hartenstein, S., and Pohl, K. (2013). An analysis of software quality attributes and their contribution to trustworthiness. In *CLOSER*, pages 542–552. DOI: <https://doi.org/10.5220/0004502705420552>.
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., and Group, P. (2009). Reprint—preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *Physical therapy*, 89(9):873–880.
- Mouratidis, H. and Cofta, P. (2010). Practitioner's challenges in designing trust into online systems. *Journal of theoretical and applied electronic commerce research*, 5(3):65–77. DOI: <https://doi.org/10.4067/S0718-18762010000300007>.
- Nielsen, J. (2024). Usability 101: Introduction to usability. <https://www.nngroup.com/articles/usability-101-introduction-to-usability/>, Accessed: 16 January 2024.
- Nietzio, A., Strobbe, C., and Velleman, E. (2008). The unified web evaluation methodology (uwem) 1.2 for wcag 1.0. In *International Conference on Computers for Handicapped Persons*, pages 394–401. Springer. DOI: https://doi.org/10.1007/978-3-540-70540-6_57.
- Norman, D. and Nielsen, J. (2024). The definition of user experience (ux). <https://www.nngroup.com/articles/definition-user-experience/>, Accessed: 16 January 2024.
- OECD (2017). *OECD Guidelines on Measuring Trust*. OECD. <https://www.oecd-ilibrary.org/content/publication/9789264278219-en>, Accessed: 16 January 2024.. DOI: <https://doi.org/10.1787/9789264278219-en>.
- Olsina, L., Papa, F., and Molina, H. (2008). Ontological support for a measurement and evaluation framework. *International Journal of Intelligent Systems*, 23(12):1282–1300. DOI: <https://doi.org/10.1002/int.20320>.
- Parlakkilä, A., Äzmez, M., and Mertoälu, S. (2020). Digital transformation of e-commerce: How did covid-19 affect customers' online shopping behaviors? *Journal of Business in The Digital Age*, 3:117–122. DOI: <https://doi.org/10.46238/jobda.823955>.
- Parmanto, B. and Zeng, X. (2005). Metric for web accessibility evaluation. *Journal of the American Society for Information Science and Technology*, 56(13):1394–1404. DOI: <https://doi.org/10.1002/asi.20233>.
- Patterson, P. G., Cowley, E., and Prasongsukarn, K. (2006). Service failure recovery: The moderating impact of individual-level cultural value orientation on perceptions of justice. *International Journal of Research in Marketing*, 23(3):263–277. DOI: <https://doi.org/10.1016/j.ijresmar.2006.02.004>.
- Perlman, G. (1997). Practical heuristics for usability evaluation. In *CHI '97 Extended Abstracts on Human Factors in Computing Systems*.
- Ramadhan, B. A. and Iqbal, B. M. (2018). User experience evaluation on the cryptocurrency website by trust aspect. In *2018 International Conference on Intelligent Informatics and Biomedical Sciences (ICIBMS)*, volume 3, pages 274–279. IEEE. DOI: <https://doi.org/10.1109/ICIBMS.2018.8550019>.
- Sauro, J. and Kindlund, E. (2005). A method to standardize usability metrics into a single score. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 401–409. DOI: <https://doi.org/10.1145/1054972.1055028>.
- Sauro, Jeff (2024). A practical guide to the system usability scale. <https://measuringu.com/product/suspack/>, 17 January 2024.
- Seffah, A., Donyaee, M., Kline, R. B., and Padda, H. K. (2006). Usability measurement and metrics: A consolidated model. *Software quality journal*, 14:159–178. DOI: <https://doi.org/10.1007/s11219-006-7600-8>.
- Shankar, V., Urban, G. L., and Sultan, F. (2002). Online trust: a stakeholder perspective, concepts, implications, and future directions. *The Journal of strategic information systems*, 11(3-4):325–344. DOI: [https://doi.org/10.1016/S0963-8687\(02\)00022-7](https://doi.org/10.1016/S0963-8687(02)00022-7).

- Sharma, G. and Lijuan, W. (2015). The effects of on-line service quality of e-commerce websites on user satisfaction. *The electronic library*, 33(3):468–485. DOI: <https://doi.org/10.1108/EL-10-2013-0193>.
- Sommerville, I. (2011). *Engenharia de software*. Pearson Prentice Hall.
- Song, S., Bu, J., Shen, C., Artmeier, A., Yu, Z., and Zhou, Q. (2018). Reliability aware web accessibility experience metric. In *Proceedings of the 15th International Web for All Conference*, pages 1–4. DOI: <https://doi.org/10.1145/3192714.3192836>.
- Song, S., Wang, C., Li, L., Yu, Z., Lin, X., and Bu, J. (2017). Waem: a web accessibility evaluation metric based on partial user experience order. In *Proceedings of the 14th International Web for All Conference*, pages 1–4. DOI: <https://doi.org/10.1145/3058555.3058576>.
- Tao, H., Chen, Y., and Pang, J. (2015). A software trustworthiness measure based on the decompositions of trustworthiness attributes and its validation. In *Industrial Engineering, Management Science and Applications 2015*, pages 981–990. Springer. DOI: https://doi.org/10.1007/978-3-662-47200-2_102.
- Tao, H. and Zhao, J. (2018). Source codes oriented software trustworthiness measure based on validation. *Mathematical Problems in Engineering*, 2018:1–10. DOI: <https://doi.org/10.1155/2018/6982821>.
- Tsuda, N., Washizaki, H., Honda, K., Nakai, H., Fukazawa, Y., Azuma, M., Komiyama, T., Nakano, T., Suzuki, H., Morita, S., et al. (2019). Wsqf: Comprehensive software quality evaluation framework and benchmark based on square. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, pages 312–321. IEEE. DOI: <https://doi.org/10.1109/ICSE-SEIP.2019.00045>.
- Veral, R. and Macías, J. A. (2019). Supporting user-perceived usability benchmarking through a developed quantitative metric. *International Journal of Human-Computer Studies*, 122:184–195. DOI: <https://doi.org/10.1016/j.ijhcs.2018.09.012>.
- Vigo, M., Arrue, M., Brajnik, G., Lomuscio, R., and Abascal, J. (2007). Quantitative metrics for measuring web accessibility. In *Proceedings of the 2007 international cross-disciplinary conference on Web accessibility (W4A)*, pages 99–107. DOI: <https://doi.org/10.1145/1243441.1243465>.
- Wang, Y. D. and Emurian, H. H. (2005). An overview of online trust: Concepts, elements, and implications. *Computers in human behavior*, 21(1):105–125. DOI: <https://doi.org/10.1016/j.chb.2003.11.008>.