

# Using a Quality Model to Evaluate User Interface Trustworthiness of e-Commerce Systems: Scoring Strategies and Preliminary Results

Andréia Casare<sup>1</sup><sup>a</sup>, Tania Basso<sup>1</sup><sup>b</sup>, Celmar Guimarães da Silva<sup>1</sup><sup>c</sup> and Regina Moraes<sup>1,2</sup><sup>d</sup>

<sup>1</sup>University of Campinas (UNICAMP), Limeira, Brazil

<sup>2</sup>University of Coimbra (UC), Coimbra, Portugal

**Keywords:** Trust, User Interface, Metric, Pilot Test.

**Abstract:** Trust in computational systems involves technical aspects and also aspects of human interaction. While technical aspects have been largely studied, we noted that there are few studies about human interaction regarding trust. In the e-commerce context, lack of consumer trust is a critical impediment to the success of e-commerce activities since users avoid systems they do not trust. In this paper, we present the results of a pilot test on a quality model to assess the trustworthiness of e-commerce systems based on user interface. The goal of the interface-based quality model is to complement the trustworthiness measurement of the whole system (i.e., complementing technical aspects measurements, such as security, connectivity, scalability, isolation, among others) and to help users to know if the e-commerce systems they are using are trustworthy. The pilot test was a means of evaluating the material to be used in a wider user interface test. In this test, we collected questionnaire answers and automatic measures, which were normalized to be inserted into our quality model. We also proposed a criteria to weight attribute scores in the model, according to the answers provided by users. Based on these results, the evaluation procedure and assets should be refined to better attend the purposes of the future assessment.


## 1 INTRODUCTION


It is a fact that individuals in societies interact with each other expecting consolidated relationships based on trust. This also happens in the digital environment, where the choice of whether to use a software product or a computing environment depends on the user's trust in the manufacturer or the perception of trust they have in the environment being used. However, different attributes are necessary to compose a computational environment that brings a perception of trust to the user (e.g. scalability, availability, QoS, robustness, security, privacy assurance, dependability, among others), since each layer of the environment to be represented relies on a set of attributes.


Online shopping has been flourishing exponentially during the last decade and is considered an excellent alternative for organizations to reach new customers. Due to the ability of reaching and attracting


consumers online, e-commerce websites play a vital role in online shopping, improving users satisfaction and for this reason, attracting the attention of marketing practitioners, society, as well as academics. In this context, the website interface plays a fundamental role in the proper functioning of the system as well as in the perception and satisfaction of the user, which leads him to trust the system being used. It has long been said that elements of human computer interface design have a significant influence on customer attitudes and perceptions of the trustworthiness of a supplier. Particularly, Roy, Dewit and Aubert (Roy et al., 2001) studied the impact of interface usability on trust in Web retailers and concluded that exists a strong relationship between interface quality and trust, highlighting the importance of some components of user interface quality and their implications.

In this direction, we argue that, although the perception of trustworthiness is quite subjective, if we identify measurable attributes that impact this perception, we can approximate the relative perception (benchmarking) by the composition of the measures of these attributes. This conviction motivated the proposal of a model (Casare et al., 2021) whose at-

<sup>a</sup> <https://orcid.org/0000-0002-8009-4929>

<sup>b</sup> <https://orcid.org/0000-0003-2467-2437>

<sup>c</sup> <https://orcid.org/0000-0001-6112-892X>

<sup>d</sup> <https://orcid.org/0000-0003-0678-4777>

tributes consider several aspects linked to the system interface, including from technical (e.g., computational infrastructure, storage space, services composition) to sociological aspects (e.g., company reputation, among others).

Based on the quality model presented in Casare et al. (Casare et al., 2021), we report a pilot test to assess the trustworthiness of an e-commerce website based on its interface. The evaluation was performed by 21 participants, whose answers help the authors to refine the evaluation process for a future application with a larger number of participants. The model being evaluated complements other similar models concerning infrastructure, data managing and services to reach, all together, the trustworthiness of the whole system.

Through the test with the users, it seeks to arouse feelings, reflections and changes in the behaviour of using the website (i.e., subjective feelings) that can be measured through objective measures (i.e., performance when loading the pages, control on the necessary functionalities, among others). The goal of the test with the users was to validate if the chosen attributes set, collected in the literature, is able to translate the perception of relative trust in using different websites.

The remainder of the work is organized as follows: Section 2 presents some background and related work; the way proposed to measure a system trustworthiness is presented in Section 3; experiments performed as a preliminary validation are presented in Section 4; finally, in Section 5 some conclusions and the suggestions for future work are presented.

## 2 BACKGROUND AND RELATED WORK

To the best of our knowledge, trustworthiness measurement from the perspective of the user experience (i.e., user perception based on interface) were not extensively studied up to now. Recently, Casare et al. (Casare et al., 2021) identified a set of user interface-based attributes that characterizes the perceived feeling of trust by the users and formalized a set of related trustworthiness metrics, based on usability, accessibility and user experience. Olsina et al. (Olsina et al., 2008) proposed an evaluation framework that allows saving values for concrete real-world measurement and evaluation projects. Their model is very similar to what we are proposing, that is, it uses software quality attributes, metrics, weights, aggregation, operators and the Logic Score of Preferences (LSP) technique. However, our model use attributes that impact user trust, and also calculates a final score that

can be used to choose the most trustworthy website (e.g., the one with best trustworthiness score).

Regarding usability measurement, Brooke (Brooke, 1996) proposed a set of usability metrics called SUS (System Usability Scale), which measures the efficiency, effectiveness, satisfaction in use, and ease of learning attributes. Seffah et al. (Seffah et al., 2006) proposed the QUIM (Quality in Use Integrated Measurement) model, which encompasses 10 usability attributes (with efficiency, effectiveness, satisfaction in use and ease of learning among them). Furthermore, standards proposed by ISO/IEC formalized some usability and accessibility attributes (e.g., ISO/IEC 25022 (ISO, 2016), which defines metrics for the quality of interaction between a user and a system).

Regarding accessibility measurement, Parmanto and Zeng (Parmanto and Zeng, 2005) proposed the WAB (Web Accessibility Barrier) metric. Based on the Web Content Accessibility Guidelines (WCAG) 1.0 checkpoints, it measures quantitatively the accessibility of web content. Song and Lai (Song and Lai, 2017) proposed a metric called Web Accessibility Experience Metric (WAEM) that matches the accessibility evaluation results with the user experience by pairwise comparisons between different websites. Also, some tools can be used for the assessment of accessibility, since they are based on WCAG guidelines (e.g., ASES<sup>1</sup> and Nibbler<sup>2</sup>).

## 3 MEASURING TRUSTWORTHINESS

Due to the complex nature of the human in business environment, assessing the interface trustworthiness is extremely subjective. However, by carefully identifying and evaluating all relevant measurable functional and non-functional characteristics that may influence trust on that service, its trustworthiness can be transformed into an objective notion. Considering the complex nature of trustworthiness, it is very unlikely that it can be scored based on only one characteristic in any scenario. More than that, it is very likely that several characteristics (i.e., attributes) from heterogeneous scales may compose the trustworthiness measurement and to score on a criteria it will be necessary to aggregate the values through a given procedure, which in turn is very likely to require the values to be expressed in the same units to operate with them. Quality Model (QM) is a reference model proposed in

<sup>1</sup><http://asesweb.governoeletronico.gov.br/ases/>

<sup>2</sup><https://nibbler.silktime.com/>

the ISO/IEC 25000 (SQuaRE) standard (IEC, 2005), whose structure formalizes the interpretation of measures and the relationship among them. It allows the representation of several attributes and the definition of how the measures should be aggregated, as well as what procedures have to be used to homogenize their values. It is possible to define one quality model for each attribute, and then, these different perspectives can be aggregated following a hierarchical structure. Furthermore, it allows the configuration of thresholds, weights and operators. The next subsections present two Quality Model representing e-commerce system components, such as the component for Interface and for the whole system.

### 3.1 Interface Trustworthiness Quality Model

Like any part of a software product, measuring interface quality is important because it helps to understand deficiencies and guides improvements in this field. The work of Casare et al. (Casare et al., 2021) presented 25 interface metrics formalized, as follows: 4 sub attributes composing *Learnability* (Easy of Learning, Navigation, Coherent Buttons and Coherent Menus); 4 composing *Efficiency* (General Flexibility, Environment Flexibility, Responsive, Performance); 4 composing *Perceivability* (Simple Screens, Colors and Fonts, Perception of System Status, Performance); 4 composing *Operability* (Back Button, Perceivable Focus, Broken Links, Affordable); 2 composing *Safety in Use* (Failure Handling, Rate of Failures); 5 composing *User Experience* (Company Information, Company Reputation, Privacy Policies, Customer Opinion and Padlock). Furthermore, *Satisfaction* and *Usefulness* are not composed of other metrics. More details about these metrics can be found in the work of Casare et al. (Casare et al., 2021).

Based on this interface-based metrics formalization, we designed an Interface Quality Model so that an interface trustworthiness score can be calculated based on the identified attributes. To preserve the readability, Figure 1 shows only the main composite attributes of the interface QM. It presents three levels of this QM and partially presents the fourth and fifth levels (in fact, we detailed only the *Efficiency* sub attributes in the fourth level and *Performance PageUp* in the fifth level). It is important to note that there is a common sub attribute between *Efficiency* and *Perceivability* (i.e., Performance), which means that this measurement is used to calculate the score of both composite attributes.

Due to the strong subjectivity of the interface attributes, the scores for the majority of the attributes are obtained through questionnaires, which are supposed to be answered after a test with the users (e.g., the users interact with the website to perform some usual functionalities and after answering the questionnaires). Only 4 attributes (Broken Links, Affordable, Performance Page Up and Responsive) are less subjective and can be measured through automatic tools. Leaf attributes represent metric definitions with associated scores based on the measures collected by the system monitoring process. They can be normalized (using the limit values NormalMin and NormalMax) ensuring that operators aggregate values at same scales and they are compared against thresholds (ThresholdMin and ThresholdMax) assuring that only relevant and valid values are considered. The values for each attribute  $i$  are influenced by an adjustable weight ( $W_i$ ), which specifies a preference over one or more attributes of the system, according to predefined requirements. For example, in the context of Figure 1, Usability ( $W_1=35\%$ ) and Accessibility ( $W_2=35\%$ ) have the same importance to compose the Interface Trustworthiness score, while User Experience has a bit less importance ( $W_3=30\%$ ) in this composition.

The final score is computed using the aggregation of the attribute values, starting from the leaf-level attributes towards the root one, using the Operators (OPn), which describe the relation between them. Different types of operators may be used to define the conditions under which composite attributes are aggregated, such as neutrality (combination of simultaneous satisfaction requirements with replaceability capability); simultaneity (all requirements must be satisfied); replaceability (used when one of the requirements has a higher priority replacing the remaining requirements). In Figure 1, *Environment Flexibility* is a *Efficiency* sub attribute, which, in turn, composes *Usability* attribute. This is a subjective sub attribute and it must be obtained by applying a questionnaire that must be answered by users. The questionnaire uses a 7-point Likert scale, with questions that helps to evaluate if the e-commerce website under test is flexible to be used in different browsers and devices. The measurement score is obtained by the weighted average of each question answered by all users. With the answers of the questionnaires, the weighted average is obtained considering the Likert scale (1 to 7) and the total of responses for each of these points ( $n_1$  – total responses as “Strongly Disagree” to  $n_7$  – total responses as “Strongly Agree”), after counting the answers of all participants. Then, the standard deviation must be evaluated allowing better analysis of the perception score.

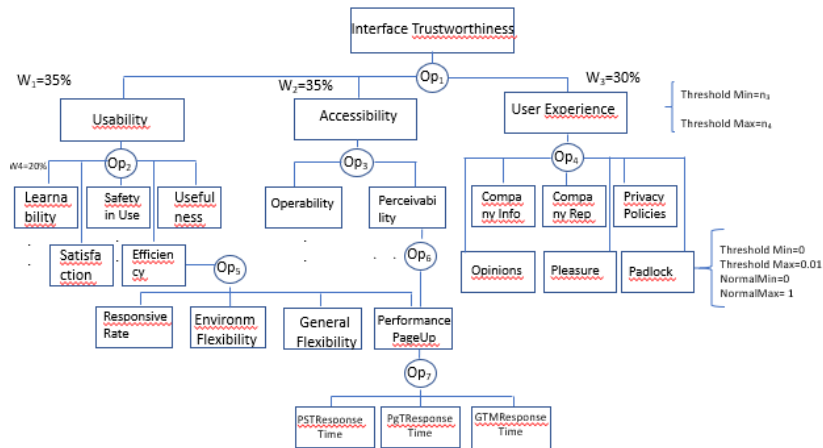


Figure 1: Interface Quality Model.

Expressions (1) and (2) present, respectively, the equations for calculating the weighted Average (Avg) and the Standard Deviation (SD) of the values informed by users considering the set of questions Q (e.g.,  $j(1), j(2), \dots, j(m)$ ) related to each attribute  $k$ . In these expressions,  $i$  is the value of the Likert Scale,  $n_{ij}$  is how many times the value  $i$  of the Likert Scale was pointed out (by all the participants) for each question  $j$  of the attribute  $k$ ,  $AVG_{attr_k}$  is the weighted average score considering all questions  $j$  belonging to the set of questions Q, and  $SD_{attr_k}$  is the standard deviation of the scores considering the same set of questions.

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

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

$$Score_{attr_k} = \frac{Score_{attr_k} - S_{min}}{S_{max} - S_{min}} \quad (3)$$

To generate the score for each attribute of the QM, transformation from the Likert Scale (1-7) to the interval score [0-1] of the  $AVG_{attr_k}$  must be done. Expression (3) shows the equation for calculating this score, where  $AVG_{attr_k}$  is the weighted average of attribute  $k$ ,  $S_{min}$  is the first value of the used Likert Scale (1) and  $S_{max}$  is the last value of the used Likert Scale (7).

Table 1 presents an example of these calculus for the sub attribute *Environment Flexibility* (EF) of an e-commerce website, measured through questionnaires, as a pilot test. Two respondents strongly agreed (Likert Scale 7) and one respondent agreed (Likert Scale 6) with question (i), and also two respondents strongly agreed and one respondent agreed with question (ii) (totaling six responses related to EF): (i) the website is

flexible to be used in different browsers; (ii) the website is flexible to be used in different devices (smartphones, tablets). As a result, the average ( $AVG_{attr_{EF}}$ ) is 6.677, the standard deviation ( $SD_{attr_{EF}}$ ) is 0.471 and the  $Score_{attr_{EF}}$  is 0.944, which indicates that, as the score value is close to 1, for this pilot test, the e-commerce website under test has a good level of environment flexibility.

Table 1: Sub Attribute Environment Flexibility - Website 1.

Environment Flexibility	
Likert Scale (1 to 7)	Total Responses
1	0
2	0
3	0
4	0
5	0
6	2
7	4
AVG	6.667
SD	0.471
Score	0.944

As we mentioned before, the interface Trustworthiness measurement is a complement to trustworthiness measurement of the whole system. The idea followed by this work is aligned with the interest of the Adaptive, Trustworthy, Manageable, Orchestrated, Secure, Privacy-assuring Hybrid, Ecosystem for Resilient cloud computing (ATMOSPHERE) project. ATMOSPHERE is an Europe-Brazil collaborative project that aims to propose solutions for federated clouds and our proposal complements the trustworthiness score with a user experience measurement. So, we present three other QMs that were defined in the scope of ATMOSPHERE project (Figure 2): *Infra Trustworthiness* (refers to available hardware and software resources), *Data Management Trustworthiness* (refers to data storage and retrieval) and *Trustworthy Data Processing Services (TDPS) Trustworthiness* (refers to services that are running to provide the expected results to the user). The de-



tails of these three QMs can be found in the ATMOSPHERE project website <sup>3</sup>. Figure 2 shows the System Trustworthiness QM that includes the Interface Trustworthiness (sub) QM.

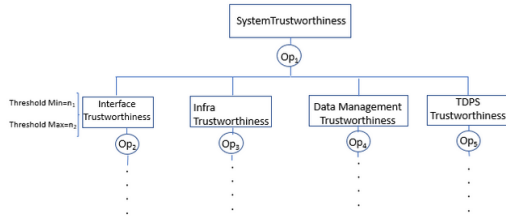


Figure 2: System Trustworthiness Model.

## 4 INTERFACE TRUSTWORTHINESS MEASUREMENT EXPERIMENTS

This section presents the experiments (pilot test) regarding the application of the Interface Trustworthiness Quality Model in order to calculate interface-based trustworthiness scores for e-commerce websites. The purpose of this pilot test is to refine the user testing process, that is, to improve the instructions on how to use the websites, improve the questions in the questionnaires and verify planned calculations on collected data, which is performed using the metrics defined in the work of Casare et al. (Casare et al., 2021). Although the main goal is not to report test results or conclusions (given that it is too early to conclude anything based on the measurements obtained with few users), we intended to test from the users interaction to complete calculations to better understand the weak points of the whole process. The first task of all the participants was to fulfill a Free and Informed Consent Form (Termo de Consentimento Livre e Esclarecido - TCLE, in Portuguese) to meet the requirements of the research ethics board.

### 4.1 Profile of Participants

Twenty one people between 21 and 52 years old performed the tests to evaluate three e-commerce websites: one of a world-renowned e-commerce, one of a famous Brazilian e-commerce and one of a famous Brazilian product. The participants (12 male and 9 female) answered questionnaires about their experi-

ence using these websites. Regarding their professional performance, 70% are not from the Information Technology domain and have never worked with user interface or computer systems; 22% already worked with computer systems, and 8% currently work with computer systems. The users were divided into two groups: 12 participants performed the test on the three websites and answered one questionnaire for each evaluated website, with questions about learnability, satisfaction, usefulness, privacy policies, among others; the other 9 participants performed the test using one website and different devices (smartphones, tablets or laptops) and browsers (Chrome, Firefox, Safari), and answered the questionnaire composed by questions about general flexibility and environment flexibility. The questionnaires and respective responses can be seen in more detail in the site <sup>4</sup>.

### 4.2 Results and Discussions

Based on the answers obtained through the questionnaires, the weighted average, standard deviation and score were calculated for each attribute represented in the Interface QM. Figure 3 presents these calculus for the three evaluated e-commerce websites. For the results presentation, we have decided not to mention the company and the e-commerce website to assure neutrality and because usually these companies do not allow the publication of evaluation results. This way, they are referred to in the rest of this paper as e-commerce 1, e-commerce 2 and e-commerce 3, with no particular order. In Figure 3, for e-commerce 1, the *Environment Flexibility* attribute presented the best values for weighted average (AVG = 6.667), standard deviation (SD = 0.471) and score (0.944). For e-commerce 2, the *Customer Opinions* attribute presented the best values for average (6.583), standard deviation (0.862) and score (0.931). For e-commerce 3, the best values were presented for the *Company Info* attribute (average 5.500, standard deviation 1.708 and score 0.750).

As mentioned before, some attributes are less subjective and can be evaluated using automatic tools. So, the scores for Responsive Rate, Performance Page Up, Affordable Rate and Broken Links attributes of the Interface QM were calculated based on the results of automatic tools. Regarding the Responsive Rate attribute, the Mobile Friendly Test tool <sup>5</sup> is the only stable tool identified to obtain its metric. In this case, this metric should be 0 (non responsive) or 1 (responsive). All the e-commerce websites used in the experiment are considered ready for mobile devices (scored

<sup>3</sup>[https://www.atmosphere-eubrazil.eu/sites/default/files/D3.6-Trustworthiness\\_Measurement\\_and\\_Analysis\\_Services\\_Implementation.docx.pdf](https://www.atmosphere-eubrazil.eu/sites/default/files/D3.6-Trustworthiness_Measurement_and_Analysis_Services_Implementation.docx.pdf)

<sup>4</sup><https://wordpress.ft.unicamp.br/seis/relatorios/>

<sup>5</sup><https://search.google.com/test/mobile-friendly>

Trust attributes	e-commerce 1			e-commerce 2			e-commerce 3		
	AVG	SD	Score	AVG	SD	Score	AVG	SD	Score
Learnability	5.486	1.922	0.748	6.139	1.294	0.856	3.431	2.254	0.405
Safety in use	4.600	1.855	0.600	5.217	1.529	0.703	3.617	1.817	0.436
User Satisfaction	5.917	1.498	0.819	6.167	1.434	0.861	2.875	2.297	0.313
Usefulness	5.583	1.622	0.764	6.056	1.026	0.843	2.889	2.208	0.315
General Flexibility	4.875	1.922	0.646	4.208	2.179	0.535	3.792	2.398	0.465
Environment									
Flexibility	6.667	0.471	0.944	5.833	1.462	0.806	5.333	5.556	0.722
Perceivability	6.135	1.497	0.856	6.010	1.396	0.835	4.177	2.305	0.530
Operability	6.333	1.106	0.889	5.500	1.732	0.750	4.417	1.935	0.569
Company info	5.667	2.134	0.778	6.500	0.866	0.917	5.500	1.708	0.750
Company reputation	5.125	1.900	0.688	6.167	1.179	0.861	3.667	1.724	0.037
Privacy policies	5.250	2.126	0.708	6.042	1.399	0.840	3.208	2.121	0.187
Customer opinions	6.250	1.639	0.875	6.583	0.862	0.931	4.833	2.034	0.639
Pleasure	6.083	1.256	0.847	5.500	1.848	0.750	3.167	2.115	0.361

Figure 3: Average (AVG), Standard Deviation (SD) and Score calculation for e-commerce, based on questionnaires.

as 1) based on Mobile Friendly Test tool (i.e., they are responsive). Furthermore, Figures 4, 5 and 6 shows, respectively, the results for Performance Page Up, Affordable Rate and Broken Links attributes.

Website	Automatic tools			Score
	PageSpeed	PingDom	GTMetrix	
	AVG	AVG	AVG	
e-commerce 1	88.55	95.55	91	0.917
e-commerce 2	76.77	91.77	76.44	0.817
e-commerce 3	29.22	83.55	33.11	0.486

Figure 4: Performance Page Up measurements and score calculation for e-commerce, based on automatic tools.

Website	Automatic tools			Score
	ASES %	Nibbler %	Access Monitor %	
e-commerce 1	90.45	85	46	0.738
e-commerce 2	89.72	not rated	48	0.689
e-commerce 3	93.39	82	64	0.798

Figure 5: Affordable Rate measurements and score calculation for e-commerce, based on automatic tools.

In Figure 4, the Performance Page Up score is calculated using the average of the measurements provided by the automatic tools (Page Speed, PingDom and GTMetrix). The same calculation (i.e., average) is performed to obtain the Affordable Rate score (5), which uses the Ases, Nibbler and Access Monitor tools. The Broken Link score is calculated based on the maximum rate obtained by any of the tools, i. e., it is calculated as MAX(Dead Link Checker (broken links / total links), Xenu’s Link (broken links / total links)). More details about the metrics for calculating scores based on automatic tools can be found in the work of Casare et al. (Casare et al., 2020).

After calculating the scores for the leaf attributes (i.e., the attributes already presented, whose calculus were obtained through questionnaires or automatic tools), these values are used to calculate the scores of their respective composite attributes in the Interface QM. To do this, it is necessary to use the weights for

Website	Automatic tools						Score
	Dead Link Checker		Xenu Link		Screaming Frog		
	Broken links	Total	Broken links	Total	Broken links	Total	
e-commerce 1	8	607	0	1	1406	69444	0.020
e-commerce 2	0	1	1	1	33	165743	0.000
e-commerce 3	145	2000	31	4272	38	6395	0.073

Figure 6: Broken links measurements and score calculation for e-commerce, based on automatic tools.

each composite attribute. These weights are shown in Figure 7 and the way how they were obtained is explained in the next subsection.

Definition of attributes weight in QM		
Usability	Learnability	38%
	Safety in use	19%
	User Satisfaction	15%
	Usefulness	16%
	Efficiency	12%
Accessibility	Perceivability	82%
	Operability	18%
User Experience	Company info	15%
	Company reputation	12%
	Privacy policies	23%
	Customer opinions	17%
	Pleasure	11%
	Padlock	22%

Figure 7: Weights of Sub Attribute of QM.

### 4.3 Weighting the Attributes

Besides calculating the attribute scores, it was possible, analyzing the questionnaires responses, to determine the weights for each composite attribute in the Interface QM. These weights were defined according to the participants’ perception, i.e., the attributes that received the highest score (7) are considered more significant for measuring trust Expression (4) presents the equation for calculating the weight for each composite attribute. The Weight<sub>j</sub> is the importance of the attribute j for its parent attribute. It is important to mention that, if two attributes have the same amount of respondents who scored it with the highest score (7), the second highest score (6) will be used to determine which one is the most important; then the third highest score (5) will be used, and so on so forth. However, the weight for both attributes will be the same.

$$Weight_j = \frac{n7_j}{\sum_{j=1}^n n7_j} \tag{4}$$

Figure 7 presents the weight of each composite attribute of Interface QM that was collected with questionnaires. These weights were calculated with Expression (6), and were distributed guided by the QM structure. According to the experiments, the most important Usability sub attribute is Learnability, with 38%; for Accessibility is Perceivability, with 82% and for User Experience is Privacy Policies, with 23%.

These weights must be considered to complete the Interface QM.

#### 4.4 Calculating the Trustworthiness Scores for the e-Commerce Websites

In this subsection we present the process to calculate the scores for composite attributes. For sake of simplicity, we explain, as example, the calculation of some attributes of the third level of the Interface QM (e.g. Efficiency), one attribute of the second level (Usability) and the first level, i.e., the root attribute in the QM (Interface Trustworthiness). The remaining composite attributes, which are not explained here, follow the same calculation process.

The Efficiency attribute is a composition of Responsive Rate, Environment Flexibility, General Flexibility and Performance Page Up. Environment Flexibility and General Flexibility attributes were evaluated through the questionnaires. Their weights were defined based on the Expression (6) and got a rate of 11% and 15% respectively. To reach the full rate (100%), it was assigned, respectively, a weight of 37% for Responsive Rate and Performance Page Up attributes. Therefore, as an example, the calculation of Efficiency score is:

$$\text{ScoreEfficiency} = (\text{ResponsiveRate} * W + \text{EnvironmentFlexibility} * W + \text{GeneralFlexibility} * W + \text{PerformancePageUp} * W), \text{ i.e., } \text{ScoreEfficiency} = (1 * 0.37 + 0.944 * 0.11 + 0.646 * 0.15 + 0.917 * 0.37) = 0.910.$$

Following the Interface QM, the next attribute to be calculated is Usability, which is composed by Learnability, Satisfaction, Safety in use, Efficiency and Usefulness attributes. The same reasoning applies to Accessibility and User Experience attributes. Finally, the Interface Trustworthiness score, which is composed by Usability, Accessibility and User Experience is calculated. Figure 8 presents the Usability, Accessibility, User Experience and Interface Trustworthiness scores of the Interface QM.

	e-commerce 1	e-commerce 2	e-commerce 3
ScoreEfficiency	0.910	0.841	0.959
ScoreUsability	0.753	0.824	0.449
ScorePerceivability	0.878	0.836	0.542
ScoreOperability	0.890	0.881	0.790
ScoreAccessibility	0.880	0.844	0.587
ScoreUserExperience	0.824	0.895	0.528
ScoreInterfaceTrust	0.819	0.852	0.521

Figure 8: Trustworthiness score calculation for e-commerce websites.

Analyzing the scores obtained in the pilot test we have some evidence that e-commerce 2 had the

best Interface Trustworthiness score (0.852), followed by e-commerce 1 (0.819) and e-commerce 3 (0.521), which presents the worst trustworthiness. The worst score is the e-commerce 3 Usability attribute (0.449). The e-commerce 2 User Experience score is the best one. The e-commerce 2 also presents the best usability among the three websites and e-commerce 1 is the most accessible of them.

Although these results may provide some evidence, they are not conclusive results, since the test was carried out with few users and aimed to improve the process. However, the pilot test reached the expected goals once we are able to identify problems in some steps of the methodology and fix them before the test with a wider number of users.

Firstly, the participants reported that knowing the post-test questionnaire before starting the pilot test helped them to have more attention to some details of the interface and the task that had to be performed during the test on the website. Aware of this, we have improved our test guideline to suggest that the participant read the post-test questionnaire before interacting with the website.

Some participants reported that the option “not applicable” was missing in some questions, such as ones related to website failures. If there is no failure, how should it be scored? The questionnaire was analyzed and this option was added in the questions about Fault Handling and Broken Links, plus an observation in the instruction to select the option “4” (neutral score) in case of doubt in choosing the answer.

During the results calculation step, the lack of information about “Start and end time” of the test in each analyzed website was detected. In addition to being interesting to measure the test effort, it is necessary to calculate the failure rate, which is one of the attributes in the model and it was completely forgotten. The information now is being required in the questionnaire and we added an alert in the guideline to highlight the importance of this information.

At the beginning we were in doubt about the usefulness of performing the whole evaluation process, as the results with few people would not be reliable enough for any strong conclusion. Fortunately, we persisted in completing the Quality Model with all collected metrics and calculated the results (all the scores). In doing so, we realized that the weights of the metrics collected by the automatic tools were overvalued. This was happening because the value to complete the total percentage (100%), taking into account the other attributes of the same group, was being attributed to this weight automatically. To solve the problem, a question was added for each automatic tool about the importance perceived by the participant

related to the automatic attribute.

## 5 CONCLUSIONS AND FUTURE WORK

This work presents a solution to support user interface measurement and analysis, which can help the computation of trustworthiness scores. The approach was evaluated during a pilot test whose results are also presented. Twenty one metrics were obtained based on the answers of questionnaires and four metrics with automatic tools evaluation. The work is part of a wider proposal, in which several metrics were defined, validated and combined following a methodology toward trustworthiness score calculation.

The interface trustworthiness score should translate the relative user's perception when using online applications. It complements other technical trustworthiness scores (such as Infrastructure, Data Management and Data Processing Services) toward the System Trustworthiness Score, which will allow users to compare (benchmarking) and choose systems that present a high level of trust. It is important to emphasize that the proposal is not to predict the website trustworthiness, but rather to offer a mechanism for evaluating the website trustworthiness aiming to choose, among the possible websites available for the task to be done, the one with the highest level of trust. Through the use case composed by 3 e-commerce websites, it was possible to conclude that the approach is feasible and can be applied to e-commerce websites. Moreover, it was possible to observe the importance of the proposed mechanism (i.e., the Interface Quality Model) to obtain the score, as well as the equation to calculate the weight of each sub attribute, as it balances the results based on the importance of the attributes.

The problems identified during the pilot test were fixed for the more complete test, as follows: (i) in the test guidelines, a suggestion to read the post-test questionnaire before accessing the website was added and also a highlight on the importance of filling the time of test start and end on each website; (ii) the option "Not Applicable" was added to some questions plus a remark linking option "4" when no answer is adequate; (iii) the start and end time were added to the post-test questionnaire; (iv) a question was added for each metric collected by the automatic tools, to catch the participant perception about their importance.

Future work includes the use of the testing procedure with a larger number of participants and defining appropriate statistical models (for example, PLS - Partial Least Squares and Cronbach's alpha), to assess

the reliability of the measures, their consistency and the homogeneity of the items in the scale, helping to identify the best set of attributes to consider (that is, the most reliable set of measures).

## ACKNOWLEDGEMENTS

This work is supported by the ATMOSPHERE project (<https://www.atmosphere-eubrazil.eu/> - Horizon 2020 No 777154 - MCTIC/RNP), ADVANCE project (<http://advance-rise.eu/> - Horizon 2020-MSCA-RISE No 2018-823788) and CAPES, Finance code 001.

## REFERENCES

- Brooke, J. (1996). Sus: a "quick and dirty" usability. *Usability evaluation in industry*, 189.
- Casare, A., Basso, T., and Moraes, R. (2020). Trust metrics to measure website user experience. In *The 13th International Conference on Advances in Computer-Human Interactions*, pages 1–8.
- Casare, A., Silva, C., Basso, T., and Moraes, R. (2021). Towards usability interface trustworthiness in e-commerce systems. In *15th International Conference on Interfaces and Human Computer Interaction*, pages 1–8.
- IEC, I. (2005). Software Product Quality Requirements and Evaluation - SQUARE. User guide, ISO/IEC.
- ISO (2016). Systems and software engineering - systems and software quality requirements and evaluation (square) - measurement of quality in use (ISO/IEC).
- 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.
- 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.
- Roy, M. C., Dewit, O., and Aubert, B. A. (2001). The impact of interface usability on trust in web retailers. *Internet research: Electronic Networking Applications and Policy*, 11(5):388–398.
- Seffah, A., Donyae, M., Kline, R. B., and Padda, H. K. (2006). Usability measurement and metrics: A consolidated model. *Software quality journal*, 14(2):159–178.
- Song, L. and Lai, H. (2017). Identifying factors affecting customer satisfaction in online shopping. In *Proc. of the 4th Multidisciplinary International Social Networks Conference*, pages 1–12.