International Journal of Human-Computer Studies 122 (2019) 184–195

Contents lists available at ScienceDirect

International Journal of Human-Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Supporting user-perceived usability benchmarking through a developed quantitative metric

Roberto Veral, José A. Macías*

Escuela Politécnica Superior. Universidad Autónoma de Madrid, Tomás y Valiente 11, 28049 Madrid, Spain

ARTICLE INFO

Keywords:
User perceived satisfaction
User Experience
Reaction Cards
User-centered development
Software metric, Software quality

ABSTRACT

Most user-centered assessment activities for ensuring usability are principally focused on performing formative evaluations, enrolling users to complete different tasks and thus obtaining indicators such as effectiveness and efficiency. However, when considering broader scenarios, such as in User Experience (UX) assessments, user perceived satisfaction (or perceived usability) is even much more relevant. There are different methods for measuring user perception, however most of them are mainly qualitative and based on individual assessments, providing little specific support to carry out comparisons–i.e., benchmarking on user-perceived usability. In this paper, we propose a quantitative metric to achieve comparative evaluations of usability perception based on Reaction Cards, a popular method for obtaining the user's subjective satisfaction in UX assessments. The metric was developed through an empirical study. Additionally, it has been validated with usability experts. Besides, we provide a supporting tool based on the developed metric, featuring a framework to store historical evaluations in order to obtain charts and benchmark levels for comparing perceived usability from different artifacts such as software products, applications categories, services, mockups, prototypes and so on. Furthermore, an evaluation involving usability professionals was achieved, providing satisfactory results to answer research questions, thus demonstrating the suitability of the approach proposed.

1. Introduction

Nowadays, usability can be considered as one of the principal software quality characteristics to assure (Sánchez and Macías, 2017). Most software products include interactive facilities that need to be tackled accordingly (Nan and Jun, 2016) depending on the different user roles (Bodker, 2009). Therefore, it is important to carry out usability activities throughout the development process (Seffah and Metzker, 2004), and not only at the end of it (Cayola and Macías, 2018). User testing is the most common practice used in user-centered design, as it facilitates the acquisition of information from final users expecting to utilize the software in a near future (Baldassarri et al., 2014). In fact, user testing comprises an important activity to measure effectiveness and efficiency of a given interactive software, exploring how the final user performs the elicited tasks and how long s/he takes to perform them (Wordle, 2017). However, these metrics report low or poor information concerning the user's overall satisfaction with respect to the product being developed, overlooking important concerns about how the user perceives other important characteristics such as usefulness, attractiveness and so on.

In addition, it is quite common in visual designs to consider the user's emotional response, which can be more transcending than the ability to complete tasks. This is an important key, since the user's satisfaction is commonly related to a valuable functionality, therefore the usefulness perceived by the user, as well as an attractive and pleasant design, encourage the user to utilize the software on a regular basis (Hawley, 2010).

When the user interacts with a software product, s/he perceives certain emotional responses that can be explored to evaluate the software's functional objectives. It is probed that the first 50 milliseconds predetermine the user's perception about the system (Hawley, 2010). As an example, when designing a website for a bank, it is important that the design conveys security and confidence to the user. If the system causes insecurity as a first impression, the user may be likely to reject the use of the system, as it does not transmit the necessary confidence, even when the functionality is appropriate and comprehensive. On the contrary, a reliable and convincing visual design can provoke a feeling of security and confidence on the user's perception, and therefore this design would be more likely to be used by her/him, although having a reduced functionally. In general, a first positive impression may cause functional failures to be overlooked, while a first negative impression may cause a correct functionality to be underestimated.

* Corresponding author.
E-mail address: j.macias@uam.es (J.A. Macías).

https://doi.org/10.1016/j.ijhcs.2018.09.012
Received 4 March 2018; Received in revised form 26 July 2018; Accepted 25 September 2018
Available online 26 September 2018

This way, the correct measurement of user perceived satisfaction (or perceived usability) is a key concern when evaluating the usability of a software product. However, measuring perceived satisfaction is not always a straightforward task. Most common and existing approaches are based on questionnaires, which usually provide well-defined constructs and psychometric values and are quantitatively measurable, in comparison to other methods such as open questions and interviews. However, questionnaires present certain difficulties to obtain all the desired information. In general, comprehensive questionnaires take longer time of completion, which can be costly in formative evaluations. Also, questionnaires should consider a trade-off between affirmative and negative statements (Travis, 2008). In addition, rating values can affect decision making if the result is ambiguous. In general, users sometimes tend to provide subjective opinions that, without proper guidance, can be difficult to interpret, leading to misunderstandings in usability evaluation. As a matter of fact, simpler visual evaluation methods can result more efficient, allowing to obtain a great deal of information in a short time (Barnum, 2010) rather than asking the user to fill in long surveys. Precisely, visual approaches are useful when measuring user perception in UX assessments (Aizpurua et al., 2016). However, most existing methods are principally based on qualitative values that need to be manually interpreted. This makes it difficult to compare and obtain formative measures that help determine the usability of a system in broader scenarios, as well as obtaining a benchmark level for summative usability evaluation involving different systems.

1.1. Research questions

Based on the aforementioned concerns, we propose the following general research questions that will be addressed and corroborated throughout the paper to conduct our research:

- RQ1: Is it possible to systematize the evaluation of user perceived usability transforming qualitative assessments into a quantitative metric for further analysis?
- RQ2: Is it possible to apply such a metric to establish a benchmark level and carry out comparative assessments of usability among different systems or software designs?
- RQ3: It is possible to develop a usable supporting tool, featuring the developed metric, to carry out such comparative usability assessments, also being useful to evaluators?

1.2. The contributed solution

In order to overcome the aforementioned drawbacks and provide answers to research questions, we have developed a quantitative metric based on the qualitative method Reaction Cards (RCs) (Benedek and Miner, 2002a, 2002b). RCs method enables the subjective evaluation of usability, specifically focusing on the measurement of user perceived satisfaction. This method helps elicit emotional responses that allow to capture subjective reactions of the user after an interactive session with a system or design. The method features a set of cards that represent descriptive adjectives (originally, of both positive and negative nature). This way, the user selects those that s/he thinks that they fit better with the achieved interaction. The evaluation of this method is mostly qualitative, so usability experts have to interpret the user's perception according to the cards that s/he has selected.

Firstly, in order to transform this qualitative measurement into a quantitative one, we have carried out a study, based on a user inquiry, to rate the positivity and negativity degree of each adjective in the RCs method. This way, 55 users participated in this inquiry, which helped us obtain a statistical metric that can be used to carry out comparative evaluations among different systems or designs. In addition, 10 usability experts helped validate the metric.

Furthermore, we have developed a tool called ASSURANCE (usAbility aSsessment SUpported by ReActioN-Cards Evaluations) that serves as a supporting tool for managing comparable measurements based on the metric developed. ASSURANCE provides different functionalities that help define evaluations for different projects, utilizing the quantitative metric to enable numerical comparison among different interactive software applications or designs. The tool stores all the evaluations made, and it manages an internal knowledge base that provides a benchmark level for the evaluations carried out. As a result, ASSURANCE features statistical data and charts for comparing different categories of interactive systems and designs, which can be useful to carry out both formative and summative evaluations throughout the development process. In addition, the tool has been tested in an evaluation with 16 real users, reporting satisfactory values concerning usefulness and overall satisfaction.

The rest of this paper is organized as follows. Section 2 presents related work. Section 3 describes the approach in detail, including the inquiry to calculate the metric, the validation and the supporting tool. Section 4 describes the user evaluation and the main results obtained. Finally, Section 5 reports on conclusions and future work.

2. Related work

Subjective usability evaluation has become a key issue in the field of UX (Barnum, 2010; Barnum and Palmer, 2010), where one of the principal concerns is to know the perception of the user (Hassenzahl and Tractinsky, 2006) in terms of emotions (Minge and Thüring, 2018) and satisfaction, rather than observing how the user accomplishes system tasks and controlling the time elapsed. Specifically, most common usability evaluation methods in UX include those applied to measure satisfaction through questionnaires (O'Brien et al., 2018; Berkman and Karahoca 2016; Lewis et al., 2015). Another approach is to use graphical symbols or words that represent the user's perception, needing the intervention of an expert to interpret the results of the UX evaluation.

As an example, Mojoleaf (Mojoleaf, 2017) is a tool that allows to evaluate a design (represented by an image shared among users) through a list of 30 words (15 positive and 15 negative) that is shown to users. As a result, the tool provides the percentage of positive and negative words selected by users and the frequency of each one. This approach does not allow introducing information or comments for further evaluation. Other tools (Spreadsheet, 2017) comprise a spreadsheet as support for evaluation. The results can be used to generate a tag cloud in Wordle (Wordle, 2017). Everything is carried out manually. Another popular tool is Lemtool (Huisman and Hout, 2010), which measures the emotional impact of a design. It allows to know which areas of a design produce certain emotions, providing the user with the possibility to select an area of the design, then indicate with an icon the emotions s/he perceives and adding a comment with the description. This approach is principally focused on specific visual elements, and only 8 emotional icons are considered to carry out evaluations that have to be further interpreted by usability experts.

Reaction Cards comprises a more comprehensive method. It was first introduced by Benedek and Miner (2002a, 2002b) and consists of a set of 118 cards, each one representing an adjective related to an emotion perceived by the user during the interaction with a system or design. After the interaction, cards are delivered to the user, and s/he is asked to select the cards whose adjectives better represent her/his experience during the interaction. Once selected, the user is asked to choose the five cards that s/he considers most representative, also asking her/his to explain the reason for the choice. Essentially, it is a qualitative method that allows to obtain the user's comments about her/his experience with the corresponding software system or design. Besides, as the user experiments a probed tendency to provide positive ratings rather than negative ones, the cards are divided into two different groups, so that 60% of cards are considered as positive and the remaining 40% are considered as negative (Barnum, 2010). One

advantage of using RCs is that the method provides quick emotional information from the user interaction. A typical test using RCs takes about five minutes, and it is principally carried out manually (with no supporting tool). Rather than asking users to fill in a long questionnaire, the RCs method allows to obtain a great deal of information in a short time (Barnum, 2010). However, this method also presents certain drawbacks, since the analysis is quite subjective and the results have to be further interpreted by usability experts. In addition, the high number of available cards and the number of required ones to be selected by users and further interpreted makes it necessary to think of improvements on the original method, introducing modification and, overall, proposing the utilization of a supporting tool as the one proposed in this paper.

Barnum and Palmer (2010) conducted studies related to RCs, exploring the method's ability to provide reliable information on perceived usability. Authors proposed a variation of the original method, in which the user selects 3, 4 or 5 cards from among a reduced set adapted to the system to be evaluated, and providing a comment to each card. In this study, it was obtained that the analysis of the selected adjectives provided a greater understanding to the UX. The metric used was the frequency of selection of each card, represented through tag clouds and radial charts, and comparing different systems in the same graph.

Merčun (2014) analyzed different ways of measuring and representing the results of RCs. He firstly considered, like Barnum and Palmer, the frequency of selection of individual cards by using tag clouds, also providing visual information based on bar graphs, which allow a more concrete analysis of the differences between diverse systems to evaluate. Later on, Merčun considered other metrics such as the number of selected cards (under the assumption that an interesting design could motivate the user to select a higher number of cards) and the number of positive and negative cards selected. A posterior study was based on analyzing joint results obtained from the cards, exploring the concept of dimension. However, the problem with RCs is that there is no exact definition of what different dimensions exist and how to measure them. This way, Merčun proposed to organize the adjectives into five dimensions: ease of use, usefulness, efficiency, appearance and implication, in order to compare different systems using a radial chart, and using such representation as a metric for perceived usability.

Li and Wang (2014) proposed a reduction in the number cards for the RCs method in order to be applied to the evaluation of mobile applications. This way, authors proposed a Likert-based user test to establish the suitability of the cards, grouping the words in a similar way to the dimensional classification proposed by Merčun.

Adikari et al. (2016) proposed a quantitative data analysis for comparing desirability assessment of two software products using RCs. This way, authors modified the original method by asking users to select the cards that best describe their interactive experience and then refine the selection assigning an order of importance for each card ranging from 1 to 5 (where 5 is the highest and 1 the lowest). Cards outside these five values are not considered in the analysis. A score for each card is calculated as the sum of the importance values in each evaluation. This is proposed as a metric to compare two different products, featuring a surface measure of overall performance based on radial graphs. This approach, although quantitative, involves individual ratings of both positive and negative adjectives, as well as later calculations for each product to evaluate, which makes it complex without the existence of a supporting tool or even when considering different products to compare.

Table 1 provides a summary of the related work and how our approach overcomes some of the drawbacks commented. This highlights some of the main contributions of our work.

As summarized in Table 1, related approaches present different drawbacks in some way. The first aspect to consider is that it is not evident how the classification of the cards is achieved in terms of positive and negative adjectives, as this is not a trivial matter. Some of the cards include adjectives that can be considered as polar opposites, so that the classification in terms of positive and negative adjectives can be easily done (e.g., attractive and unattractive). However, some other cards contain adjectives that cannot be considered as polar opposites (e.g., sophisticated or calm). Such adjectives are complex to be classified into positive or negative sets. This also means that the use of only two groups for classifying adjectives can result problematic as far as the conception of an evaluation metric is concerned. On the other hand, proposals considering usability dimensions in the analysis also present barriers associated to the classification of each card in different dimensions, as well as determining the exact number of dimensions to consider. In general, existing card-based approaches present limitations to establish a systematic method to compare user perception among different software systems or designs, as they are principally based on the expert's classification criteria or specific product-based calculations. In addition, most of them are manually performed, despite the high number of cards, considering frequencies of words or graphical representation that need to be manually analyzed and interpreted by experts. This complicates, to some extent, the possibility of obtaining a quantitative and product-independent metric to systematically carry out benchmark levels for different systems and provide straightforward and meaningful statistical results in formative and summative usability evaluations.

3. The proposal

The contribution of this research is twofold. On the one hand, we present a new metric for the RCs method overcoming the drawbacks commented in the previous sections. On the other hand, we also present a supporting tool, based on such a metric, to provide further analysis and comparison among different systems or designs by means of statistical data and charts.

The idea behind this new metric is to associate a generic score to each card that can be used to evaluate any product and create a benchmark level. This score has been obtained from a user inquiry, where users evaluated the adjectives related to each card through a Likert scale ranging from 1 (very negative) to 5 (very positive). This increases accuracy and independence in comparison to using only two categories (positive and negative) to score and classify each card or scoring each card according to the importance related to a specific product to evaluate. This also eliminates the subjectivity derived from the opinion of each user or expert individually, and it achieves a more accurate assessment with a valid justification in terms of the score associated to each card. The final score for a concrete evaluation is calculated once the user has carried out the final selection of cards, regardless of the number of cards selected. Each card is weighted accordingly, therefore the number of cards selected is not a factor, and thus all the cards can be considered for the evaluation, and not only the five most representative as in previous proposals.

On the other hand, with ASSURANCE it is possible to carry out perceived usability evaluations of different systems or designs based on the metric created, enabling the user to introduce qualitative comments for each card selected, thus allowing to carry out quantitative and qualitative evaluations to improve the analysis. Past registered evaluations are used as a knowledge base to create benchmark levels, enabling the possibility to compare different systems and designs using the same metric, therefore providing statistical significance. The tool also provides statistical charts to graphically compare different systems, which increases analytical capabilities with respect to other approaches.

3.1. A quantitative metric for Reaction Cards

In order to obtain a score for each adjective/card, and also analyze the degree of positivity or negativity of each adjective, a user inquiry was carried out.

Table. 1
Summary of the related work's main points and the strengths of the proposed approach.

Related work	Main points	Strengths of our proposal
Questionnaire-based approaches (O'Brien et al., 2018; Berkman and Karahoca 2016; Lewis et al., 2015).	They take longer time of completion, which is costly in formative evaluations. Rating values can affect decision making if the result is ambiguous. A trade-off between affirmative and negative statements should be considered. Further processing and understanding is required to carry out benchmarking.	Visual approach that facilitates the measurement of user perception in both formative and summative evaluations. Empirically evaluated to measure positive, negative and neutral adjectives quantitatively. Tool support to achieve benchmarking and visual representations easily.
Mojoleaf (Mojoleaf, 2017).	Reduced amount of adjectives classified into positive (15) and negative (15). It does not allow users to introduce further information or comments. Results have to be further interpreted by the expert.	Higher number of weighted adjectives, classified into positive, negative and neutral in different proportions and rated using a real-number scale. Supporting tool to enable a more complete and systematic evaluation and interpretation or results that allow the user to introduce comments for further analysis.
Spreadsheet (Spreadsheet, 2017) and Wordle (Wordle, 2017).	Manual evaluation process comprising a spreadsheet that could be used to generate tag cloud in Wordle. Complex benchmarking of different systems.	Assisted and systematic evaluation process through the elaborated metric and the supporting tool that automatically generates tags clouds, but also visual charts, to allow benchmarking in the same tool.
Lemtool (Huisman and Hout, 2010).	Focused on visual elements, comprising only 8 icons to carry out evaluations that have to be further interpreted by the expert.	Higher set of weighted adjectives that provide a precise and quantitative evaluation in a systematic way, reducing ambiguity in interpretations, and providing comparative charts automatically.
Reaction Cards. Original method by Benedek and Miner (2002a, 2002b).	User selects 5 cards out of 118 and explains the reason for the choice. Expert manually interprets this information in a qualitative way, which is subjected to subjectivity.	User selects as many cards as desired. Each card is weighted accordingly, therefore the number of cards selected is not a factor, and thus all the cards can be considered for the evaluation. The selection of cards systematically provides a score to quantitatively evaluate and compare different systems automatically using the supporting tool. This also reduces the ambiguity in later interpretations and easily facilitates further analysis.
Reaction Cards. Improvement by Barnum and Palmer (2010).	User select 3, 4 or 5 cards from among a reduced set adapted to the system to be evaluated. A metric was developed representing the frequency of the selection to compare different systems, with no specific weight for each adjective.	The final score for a concrete evaluation is calculated according to each weighted adjective, which results more accurate than the frequency of selection. This is regardless of the number of cards selected. Each card is weighted accordingly, therefore the number of cards selected is not a factor, and thus all the cards can be considered for the evaluation, and not only a reduced set of them.
Reactions Cards. Improvement by Merčun (2014).	Based on Barnum and Palmer work, a new metric was proposed—i.e., the number of selected cards, which is based on the assumption that an interesting design could motivate to select a higher number of cards. In addition, a different organization of adjectives, based on usability attributes, was proposed to represent the usability of different systems graphically, bringing more complexity to the original approach.	The metric provides a simpler way to measure each adjective individually, regardless of the number of cards selected, but considering a real-number scale and not specific dimensions of usability, which are difficult to obtain due to the ambiguity in interpreting the adjectives to create specific dimensions. The supporting tool provides a systematic benchmarking and more possibilities in terms of visualization and comparison.
Reaction Cards. Improvement by Li and Wang (2014).	A reduction in the number of cards was proposed for evaluating mobile applications. A Likert-based test was proposed to establish the suitability of cards grouped as proposed by Merčun, which implies a more complex dimensional classification.	Our metric is independent of the number of cards selected. Consequently, it can be also applied to mobile applications. Also, the metric and the supporting tool simplify the way benchmarking is achieved, with no necessity to specify dimensions that may be subject to certain ambiguity.
Reaction Cards. Improvement by Adikari et al. (2016).	Each card was assigned by an order of importance ranging from 1–5. A metric, as the sum of the importance values in each evaluation, was proposed to compare different products using radial graphs. Although quantitative, this involves individual ratings of both positive and negative adjectives, implying also further calculation for each product to evaluate.	A weighted set of adjectives, regardless of the number of card selected, provides different values not only for positive and negative adjectives but also for neutral ones. The supporting tool helps reduce the complexity of calculations and the representation of comparative results among different systems.

3.1.1. Method

The inquiry method comprised a questionnaire that was published in the web and filled in by participants. The questionnaire contained the 118 adjectives representing each card in the RCs method. Participants were asked to assess each adjective in a Likert scale, ranging from 1 to 5, where 1 means “very negative”, 2 means “negative”, 3 means “neutral”, 4 means “positive” and 5 means “very positive”. The cards were randomly presented to participants. We did not considered any specific counterbalancing design, as the number of both participants and random cards is high enough to avoid confounding factors and influencing the results. Once created, we distributed the link to a large set of participants to take part in the inquiry. The web questionnaire was available over a certain period of time. Once expired, we proceeded to analyze the results for each one of the 118 adjectives, using statistical analysis to establish criteria to elaborate the metric.

3.1.2. Research questions

In order to have meaningful results, we based on the RQ1 (related to the feasibility of obtaining a quantitative metric) to elaborate the following additional research questions:

- RQ1.1: Can the reliability of the results obtained be considered suitable to weigh each adjective?
- RQ1.2: Is it possible to establish a metric in view of the statistical results obtained?

3.1.3. Participants

The inquiry involved 55 persons completing the evaluation of the 118 cards on a voluntary basis. 28 were women (52%) and 26 were men (48%). Ages ranged from 18 to 66 years, being 69% of participants between the ages of 18 and 30. As for educational background, all of

Table. 2
List of 118 adjectives sorted by the mean value. Max, min, standard deviation, median, confidence interval (95% CI) and general category (positive, negative and neutral) values are also shown.

Adjective/Card	Mean	Max	Min	SD	Median	95% CI	Category
Undesirable	1.333	3	1	0.556	1	0.147	Negative
Poor quality	1.392	4	1	0.684	1	0.181	Negative
Not Secure	1.471	3	1	0.535	1	0.141	Negative
Stressful	1.471	3	1	0.535	1	0.141	Negative
Ineffective	1.510	4	1	0.654	1	0.173	Negative
Frustrating	1.529	3	1	0.596	2	0.157	Negative
Unapproachable	1.588	4	1	0.747	2	0.197	Negative
Annoying	1.608	3	1	0.583	2	0.154	Negative
Not valuable	1.667	4	1	0.779	2	0.206	Negative
Incomprehensible	1.686	5	1	0.852	2	0.225	Negative
Boring	1.725	5	1	0.755	2	0.200	Negative
Disruptive	1.745	5	1	0.833	2	0.220	Negative
Inconsistent	1.745	5	1	0.855	2	0.226	Negative
Uncontrollable	1.765	4	1	0.698	2	0.184	Negative
Hard to Use	1.784	3	1	0.671	2	0.177	Negative
Gets in the way	1.863	4	1	0.528	2	0.140	Negative
Unattractive	1.922	3	1	0.567	2	0.150	Negative
Confusing	1.961	3	1	0.538	2	0.142	Negative
Distracting	2.000	4	1	0.674	2	0.178	Negative
Slow	2.059	3	1	0.615	2	0.163	Negative
Unrefined	2.137	3	1	0.662	2	0.175	Negative
Disconnected	2.176	4	1	0.671	2	0.177	Negative
Dated	2.196	4	1	0.748	2	0.198	Negative
Dull	2.196	4	1	0.724	2	0.191	Negative
Intimidating	2.216	5	1	1.052	2	0.278	Negative
Irrelevant	2.255	4	1	0.883	2	0.233	Negative
Difficult	2.294	4	1	0.806	2	0.213	Negative
Impersonal	2.373	4	1	0.723	2	0.191	Negative
Fragile	2.431	4	1	0.757	3	0.200	Negative
Busy	2.490	4	1	0.735	3	0.194	Negative
Too Technical	2.510	5	1	0.735	2	0.194	Neutral
Old	2.549	5	1	0.848	3	0.224	Neutral
Rigid	2.569	5	1	0.779	3	0.206	Neutral
Overwhelming	2.608	5	1	1,136	2	0.300	Neutral
Unpredictable	2.627	5	1	1.016	3	0.268	Neutral
Complex	2.647	5	1	0.895	3	0.237	Neutral
Overbearing	2.647	5	1	1,063	2	0.281	Neutral
Simplistic	2.686	5	1	0.892	3	0.236	Neutral
Sterile	2.725	4	1	0.623	3	0.165	Neutral
Time-consuming	2.784	4	1	0.713	3	0.188	Neutral
Unconventional	2.863	5	1	0.832	3	0.220	Neutral
Predictable	2.902	5	1	0.759	3	0.201	Neutral
Ordinary	2.961	4	2	0.571	3	0.151	Neutral
Patronizing	3.059	5	1	0.882	3	0.233	Neutral
Connected	3.157	5	1	0.644	3	0.170	Neutral
Expected	3.392	5	2	0.677	3	0.179	Neutral
Controllable	3.490	5	1	0.871	4	0.230	Neutral
Usable	3.627	5	1	0.800	4	0.211	Positive
Personal	3.647	5	2	0.719	4	0.190	Positive
Trustworthy	3.647	5	2	0.817	4	0.216	Positive
Convenient	3.686	5	1	0.791	4	0.209	Positive
Helpful	3.686	5	1	0.817	4	0.216	Positive
Empowering	3.725	5	1	0.810	4	0.214	Positive
Sophisticated	3.725	5	1	0.871	4	0.230	Positive
Straight Forward	3.725	5	2	0.673	4	0.178	Positive
Cutting edge	3.745	5	1	0.888	4	0.235	Positive
Business-like	3.765	5	3	0.652	4	0.172	Positive
Relevant	3.784	5	2	0.687	4	0.181	Positive
Customizable	3.804	5	1	0.802	4	0.212	Positive
Essential	3.804	5	2	0.818	4	0.216	Positive
Meaningful	3.804	5	2	0.731	4	0.193	Positive
Fast	3.823	5	3	0.733	4	0.194	Positive
Effortless	3.843	5	2	0.654	4	0.173	Positive
Low Maintenance	3.877	5	2	0.632	4	0.252	Positive
Calm	3.882	5	2	0.740	4	0.195	Positive
Integrated	3.882	5	3	0.651	4	0.172	Positive
Accessible	3.902	5	2	0.562	4	0.148	Positive
Compelling	3.902	5	2	0.562	4	0.148	Positive
Energetic	3.902	5	2	0.832	4	0.220	Positive
Fresh	3.902	5	3	0.695	4	0.184	Positive
Approachable	3.961	5	3	0.571	4	0.151	Positive
Compatible	3.961	5	2	0.699	4	0.185	Positive
Powerful	3.961	5	2	0.644	4	0.170	Positive
Comfortable	4.000	5	2	0.738	4	0.195	Positive
Novel	4.000	5	2	0.674	4	0.178	Positive
Flexible	4.020	5	3	0.587	4	0.155	Positive
Appealing	4.039	5	2	0.713	4	0.189	Positive
Consistent	4.039	5	3	0.602	4	0.159	Positive
Advanced	4.059	5	1	0.750	4	0.198	Positive
Understandable	4.078	5	3	0.628	4	0.166	Positive
Desirable	4.098	5	2	0.628	4	0.167	Positive
Easy to use	4.118	5	3	0.623	4	0.165	Positive
Inspiring	4.118	5	3	0.623	4	0.165	Positive
Responsive	4.118	5	3	0.593	4	0.157	Positive
Clear	4.137	5	3	0.605	4	0.160	Positive
Collaborative	4.137	5	2	0.623	4	0.165	Positive
Stable	4.137	5	3	0.574	4	0.152	Positive
Enthusiastic	4.157	5	3	0.615	4	0.163	Positive
Intuitive	4.157	5	3	0.644	4	0.170	Positive
Inviting	4.177	5	3	0.596	4	0.157	Positive
Innovative	4.196	5	3	0.606	4	0.160	Positive
Comprehensive	4.216	5	3	0.671	4	0.177	Positive
Organized	4.216	5	3	0.553	4	0.146	Positive
Time-Saving	4.216	5	1	0.954	4	0.252	Positive
Stimulating	4.235	5	1	0.781	4	0.206	Positive
Familiar	4.255	5	3	0.713	4	0.188	Positive
Useful	4.255	5	3	0.571	4	0.151	Positive
Engaging	4.274	5	3	0.512	4	0.135	Positive
Entertaining	4.274	5	3	0.547	4	0.144	Positive
Attractive	4.294	5	3	0.493	4	0.130	Positive
Satisfying	4.294	5	3	0.586	4	0.155	Positive
Friendly	4.314	5	4	0.469	4	0.124	Positive
Optimistic	4.333	5	3	0.535	4	0.141	Positive
Professional	4.353	5	3	0.605	4	0.160	Positive
Clean	4.392	5	3	0.551	4	0.146	Positive
Confident	4.412	5	1	0.719	4	0.190	Positive
Valuable	4.412	5	3	0.618	4	0.163	Positive
Motivating	4.431	5	3	0.621	4	0.164	Positive
Effective	4.451	5	3	0.562	4	0.149	Positive
Efficient	4.451	5	3	0.562	4	0.149	Positive
Secure	4.451	5	3	0.623	4	0.165	Positive
Creative	4.471	5	3	0.533	4	0.141	Positive
Fun	4.471	5	3	0.564	4	0.149	Positive
Impressive	4.490	5	3	0.709	5	0.187	Positive
Exceptional	4.510	5	3	0.628	5	0.166	Positive
Exciting	4.510	5	4	0.500	5	0.132	Positive
Reliable	4.588	5	4	0.493	5	0.130	Positive
High quality	4.706	5	3	0.547	5	0.144	Positive

them had university education. As typical RC users involved in usability evaluations may include people with different background, we thought of opening the inquiry to a heterogeneous and random population sample.

3.1.4. Results and discussion

Table 2 shows the results obtained from the evaluation for each of the 118 adjectives. We considered a geometric mean in order to avoid skewed values due to outliers. As shown, mean scores under 2.5 represent adjectives evaluated as negatives, whereas scores over 3.5 represent adjectives evaluated as positive. More specifically, we obtained that 4% of cards were scored as “very negative”, 22% as “negative”, 15% as “neutral”, 56% as “positive” and 3% as “very positive”. In short, 59% of the adjectives were considered as positive and 26% as negative. The rest of adjectives (15%) were rated as neutral (those ranging between 2.5 and 3.5). These results are closer to the ones obtained by Benedek and Miner (2002a, 2002b), where authors roughly determined ratios of 60% and 40% for positive and negative adjectives, respectively.

The reliability of the measurement scale can be considered high according to the value obtained for the Cronbach's alpha ($\alpha = 0.867$). The Cronbach's alpha is commonly used as a measure of internal consistency, and it has been used in this case to study the average inter-

correlation of the measurements obtained in the user inquiry. A Cronbach's alpha of 0.7 or higher is considered acceptable in most research situations. In this case, the obtained value of 0.867 indicates that the items have a high internal consistency. In addition, it is worth mentioning that the 95% CI was less than 0.3 in all cases ($\bar{CI}_{95\%} = 0.184$). This implies a high reliability with respect the mean value calculated for each adjective, as there is 95% probability that the mean has a maximum margin of error of ± 0.3. This also provides reliability in terms of the sample size used for the objective pursued.

Due to that, the mean value can be considered as a good measure to compose the metric in order to rate each adjective/card i individually, ∀ i = 1...118:

$$score(i) = Mean(i) \quad (1)$$

3.1.5. Proposed metric

According to that, the metric to be applied in an evaluation with RCs method can be defined as:

$$metric = \frac{\sum_{i=1}^{n} score(i)}{\theta * n} * 100 \quad (2)$$

Where *score*(*i*) represents, according to Eq. (1), the scored obtained for the card *i*, ∀ *i* = 1...*n*. On the other hand, *n* represent the total number of cards selected in an evaluation, and θ represents the maximum number of classification categories used according to the Likert scale utilized to evaluate each card; in this case θ = 5. As shown in Eq. (2), the metric has been normalized in order to minimize the dependence with respect to the number of cards selected.

Statistical analysis and results helped corroborate research questions RQ1.1 (related to the reliability of the results to weigh each adjective) and RQ1.2 (related to the feasibility to establish the metric), concluding that the reliability of the results obtained can be considered suitable to weigh each adjective/card, being also possible to establish a metric in view of the results obtained.

The developed metric was validated by usability experts and verified with the help of a supporting tool. In addition, an evaluation with real users was carried out. All this will be detailed down below.

3.1.6. Expert validation

Once the metric was elaborated, we asked 10 usability experts of both our institution and partner organizations to analyze the appropriateness of the metric on a voluntary basis. They were 6 men and 4 women, with ages ranging from 26 to 40 (M = 33.5; SD = 2.15) and having advanced skills in software engineering and human-computer interaction, all being academics on both disciplines and industry professionals with background on software development and usability having more than five-year average of experience. This way, we propose the following research question that can be considered as part of RQ1 (related to the feasibility of obtaining a quantitative metric):

- RQ1.3 Can the metric developed be considered as appropriate and more advantageous, in terms of accomplishing comparisons, than the standard method?

To carry out the analysis, we asked each expert to evaluate three different web pages after a free walkthrough of about 5 minutes. The evaluation consisted in the following steps:

1) Evaluate the navigation experience using the standard Reaction Cards method (qualitative approach).
2) Evaluate the navigation experience using the proposed metric for the Reaction Cards method (quantitative approach).
3) Analyze and compare the evaluations obtained by answering the following two questions:
   a Degree of similarity between the results obtained with the qualitative and the quantitative method. Possible responses, in a Likert scale ranging from 1 to 5, were: 1—quite similar, 2—similar, 3—neutral, 4—different, 5—quite different.
   b Ease of comparison between the results obtained for the three web pages. Possible responses, in a Likert scale ranging from 1 to 3, were: 1—better with the developed metric, 2—similar in both cases, 3—better with the standard method.

Evaluations steps 1 and 2 were randomized in order to minimize the bias due to carryover effects. On the other hand, in order to facilitate the tasks, we provided each evaluator with a spreadsheet in order to easily obtain calculations.

In general, we obtained high agreement among experts. More specifically, 100% of experts considered that the degree of similarity between the results obtained with the qualitative and the quantitative method is quite similar (60%) or similar (40%). On the other hand, 100% of experts considered that the ease of comparison is better with the developed metric than with the qualitative method. Additional comments collected after the evaluation corroborated the results obtained. In a nutshell, most experts agreed that it was much easier to compare the results of the different evaluations at a glance using a systematic metric than reviewing the notes obtained through the standard method, which makes it difficult to achieve comparisons due to the subjectivity of unilaterally interpreting jointly the adjectives selected.

These findings provide an affirmative answer to RQ1.3 (related to the suitability of the metric to make comparisons), concluding that the metric provide appropriate results, similar to those obtained with the standard method, but being more advantageous when carrying out comparisons with different systems.

3.2. Supporting tool for the comparative evaluation of perceived usability

In order to verify the developed metric and provide support for its utilization, we have created ASSURANCE. This tool enables to manage evaluations and benchmarking between different systems or designs easily. More specifically, ASSURANCE is a responsive web-based tool, conceived as a support for usability and UX evaluators. With this tool, evaluators can create evaluation projects and get comparative results among different systems or designs using the RCs method.

The tool allows to perform usability tests in an interactive way. Thus, after an interactive session with the system or design to evaluate, a new evaluation can be added by asking the user to select the cards that better fit with her/his perception. The user can also insert comments about the selected cards. All the information is stored by the tool, this way the evaluator can consult such information to analyze it in detail later on. Fig. 1 depicts the use case diagram showing the main functionally and roles in ASSURANCE.

According to the information shown in Fig. 1, the tool includes the following functionality:

- User registration and login: This provides functionality concerning user management. In general, there are three roles for interacting with the tool. Registered evaluators can manage evaluation projects and perform/add evaluations. On the other hand, non-registered users can only consult statistical information about the evaluations made in the form of bar charts, box plots, etc. (some examples can be found in Figs. 5–7). Data privacy is ensured by providing only aggregate information obtained from the knowledge base that the system includes, which is updated with every evaluation. Finally, administrators can manage accounts and evaluation projects.
- Manage evaluation projects: This allows registered evaluators to create, consult and remove evaluation projects (see Figs. 2 and 3), as well as to create different categories of products to evaluate. This includes functionality to consult all the evaluations made by the responsible evaluator, as well as the comments introduced by the users during the evaluations. On the other hand, once the evaluator

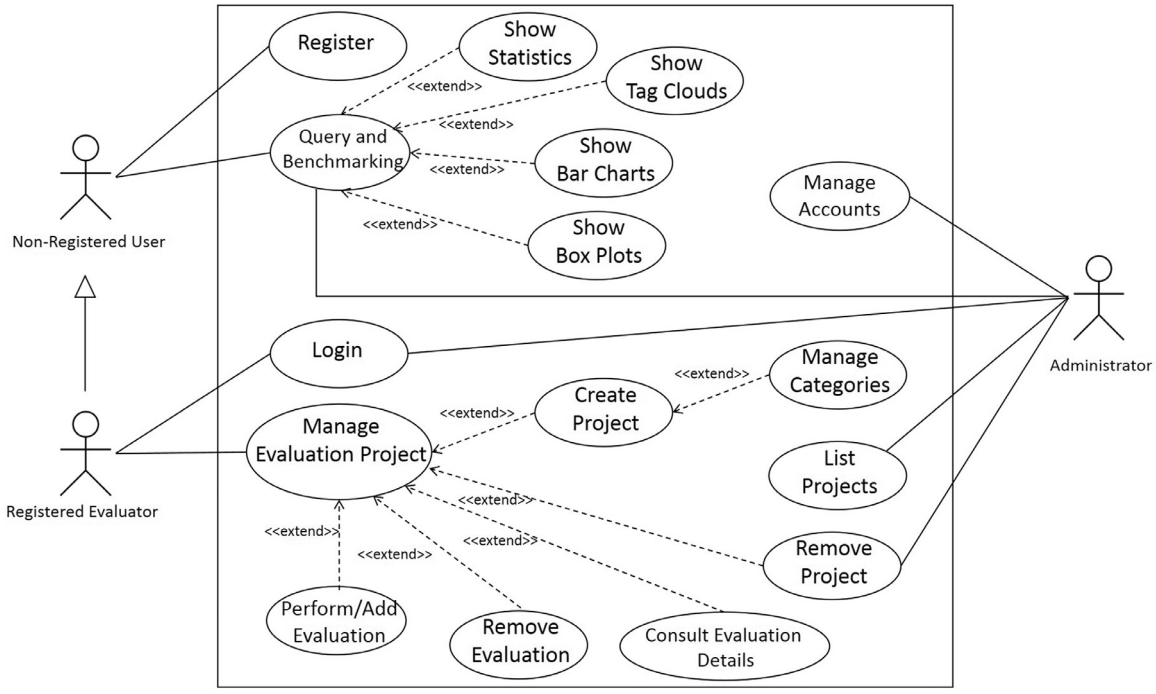

Fig. 1. Use case diagram representing the roles and functionality in ASSURANCE.

has setup a project, s/he can add a new evaluation about a categorized design or software product.

- Show statistic data and charts from evaluations: This provides functionality to show statistical information involving all the evaluations contained in the knowledge base. This way, non-registered user can compare different system categories, designs or specific software products at the same time, obtaining comparative representations through bar graphs, box plots, tag clouds, and descriptive statistics.

3.2.1. *Evaluation categories*

The first step in achieving an evaluation is to create a new project, introducing the corresponding information and the evaluation category. The information contained in the knowledge base of the system is hierarchically structured, featuring the following classification: Category -> Application -> Operating System -> Version. However, the evaluator can create new items within this hierarchy. Fig. 4 depicts an example of evaluation hierarchy created by previous evaluators, where there are two principal categories called “Web Browser” and

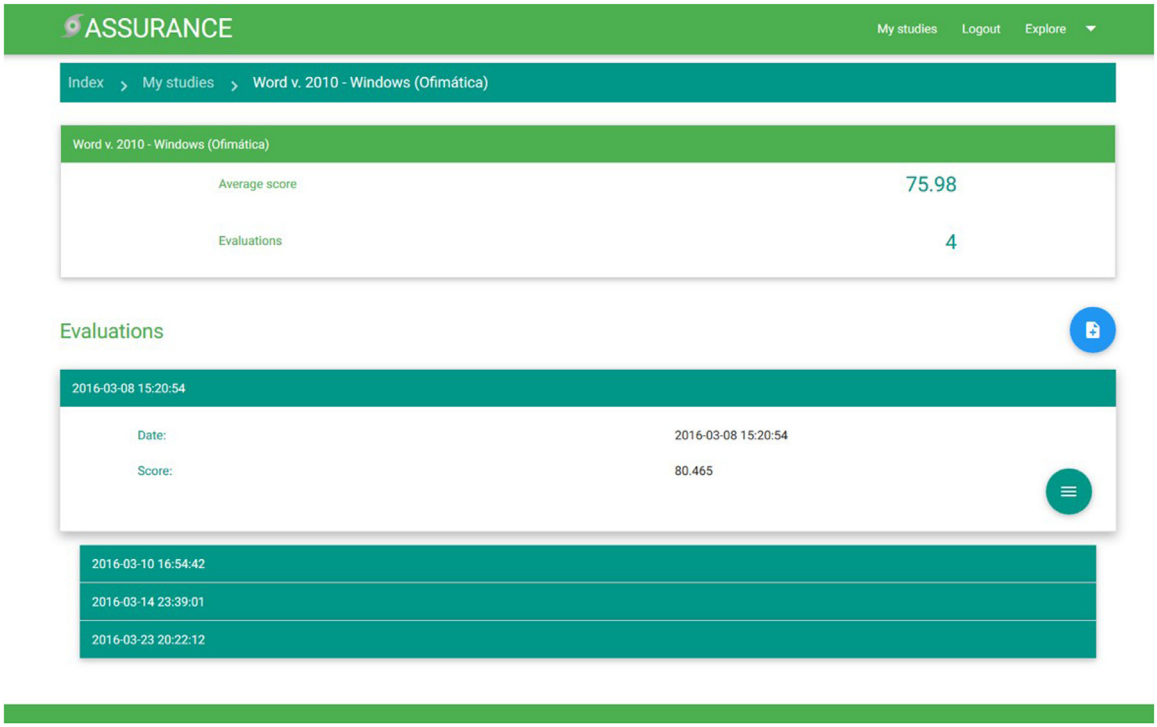

Fig. 2. An evaluation project in ASSURANCE for a specific application (Word 2010 for Windows), with the average score and the number of evaluations accomplished so far. One of the evaluations has been expanded to show further information.

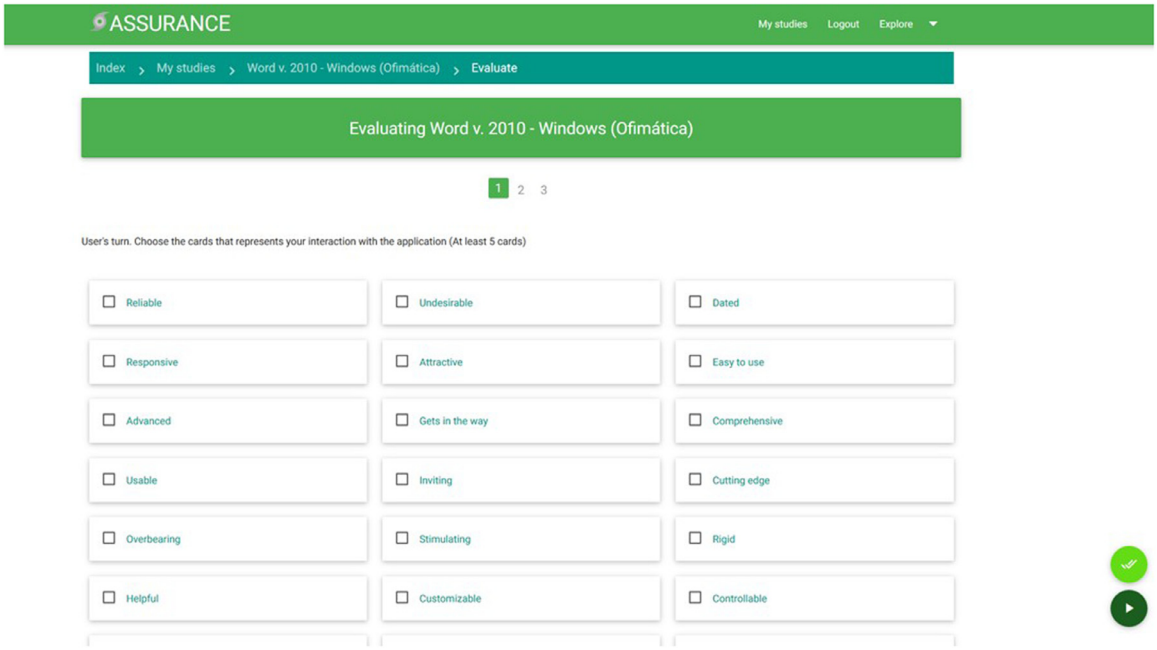

Fig. 3. A RCs evaluation in ASSURANCE for a specific application (Word 2010 for Windows). The evaluator is required to select the cards that better fit with her/his perception after an interactive session with the application.

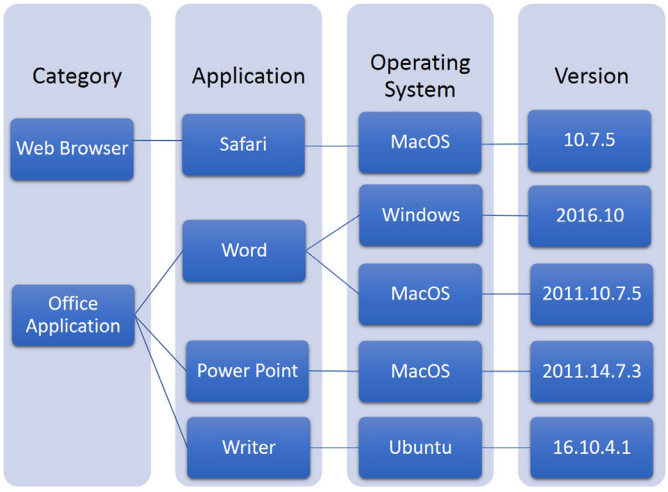

Fig. 4. An example of information hierarchy to carry out evaluations in ASSURANCE.

"Office Application", as well as different nodes representing specific applications such as Safari, Word and so on, for different operating systems and corresponding application versions. For example, if we may want to create a new usability evaluation involving Word 2016 for Windows 10, the navigation path through the hierarchy would be: Office Applications (Category) -> Word (Application) -> Windows (Operating System) -> 2016.10 (Version). Version refers to a combination of operating system and application versions, but evaluators can freely introduce any information depending on the desired aggregation criteria for establishing benchmark levels in final products. Similarly, the evaluator can create new categories in order to customize evaluations involving different applications, designs, and software families.

This hierarchical structure is also useful when showing statistical results, as an evaluator may want compare specific software applications (e.g., Word 2016 for Windows 10 and Word 2011 for Mac OS 10.7.5) or query comparative results for more general categories of applications (e.g., web navigators, office applications, all versions of Word, etc.).

3.2.2. *Visualizing statistical information and charts*

Any user (no login is required) can query aggregate information from past evaluations, which allows to compare different categories or applications (according to different operating systems and versions)

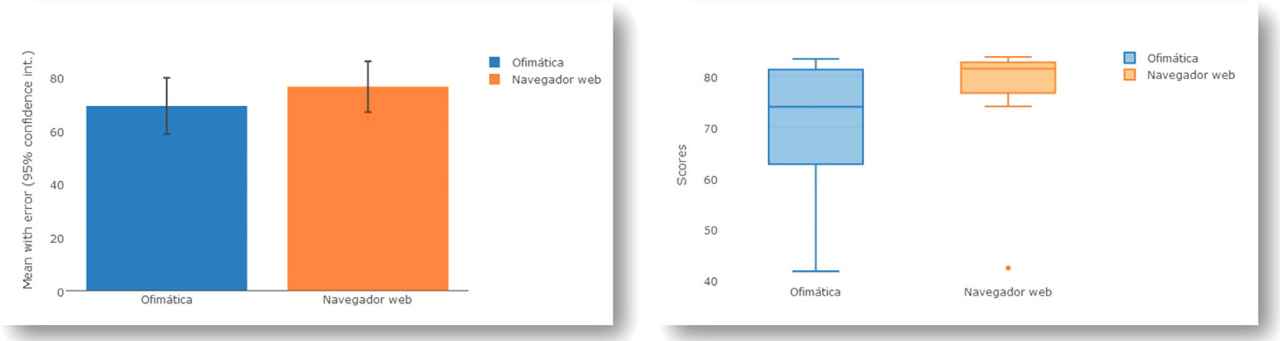

Fig. 5. Bar chart and box plot (from left to right) for comparing two different software applications categories: office application (*ofimática*) and web browser (*navegador web*).

with statistical evidence. Queried information is shown using graphical representations such as bar charts, box plots and tag clouds to show comparison of two or more systems or designs.

Fig. 5 shows, from left to right, examples of visualization in ASSURANCE. In this case, a category benchmarking including an overall comparison among office applications and web browsers has been selected. We have avoided to show a specific benchmarking of the commercial products shown in Fig. 4 in order to provide an example of visualization and not mislead the reader with a fictitious comparative overview, as the results shown mostly depend on the data collected in our knowledge base. After all, the benchmarking of products or categories, together with the resulting visualization, is similar regardless of the type of items (categories or applications) selected.

As shown in Fig. 5, bar charts include errors bars representing 95% confidence interval and mean values using the developed metric for all the evaluations concerning both application categories. By contrasts, box plots are based on quartiles that are best suited to study the dispersion in both application categories. As we can see in Fig. 5 (on the left), bar chart values inform that web browser category features a higher average score. On the other hand, the box plot (on the right) depicts how these scores are distributed, showing that web browser category presents a higher average score, as well as a higher median and quartile positions, in addition to having a more concentrated distribution than the office application category, which presents more variability between minimum and maximum values.

ASSURANCE also enables another representation to obtain detailed and individual information about the evaluations made. Fig. 6 presents, from left to right, individual statistical data concerning the same categories analyzed in Fig. 5: office application and web browser.

Additionally, ASSURANCE provides tag-cloud representations to depict individual information about the system, design or application category evaluated, showing the frequency of the cards selected in all the evaluations. As shown in Fig. 7, *comprehensive, dated, useful* and *advanced* are the most frequent cards appearing in office application evaluations (on the left), whereas *easy to use, efficient, useful* and *intuitive* are the most frequent cards appearing in web browser evaluations (on the right).

In summary, the new metric created and the supporting tool developed help answer affirmatively RQ1 (related to the feasibility of obtaining a quantitative metric) and RQ2 (related to the feasibility of applying the metric to establish benchmark levels and comparisons), and thus affirm that it is possible to systematize the evaluation of perceived usability by transforming qualitative assessments into a quantitative metric for further analysis. Besides, it is possible to apply such metric to establish benchmark levels and carry out comparative assessments of usability for different systems, software categories or designs, in a systematic way.

4. Proposal evaluation

Once the metric was verified with the construction of the supporting tool ASSURANCE and validated by experts, we proceeded to the validation of the tool by carrying through a user evaluation in order to analyze the usability and overall satisfaction with the proposal.

4.1. Evaluation method

The evaluation was carried out using ASSURANCE in a windows laptop computer, using Chrome® as web browser. This way, enrolled users interacted with the tool using the Thinking Aloud protocol (Boren and Ramey, 2000), that is, asking the user to speak aloud to register all her/his comments and behavior to be further analyzed later on.

This was a controlled evaluation carried out in a laptop having our tool installed. In general, we moved towards the user location to facilitate the evaluation, using always the same laptop. Users were provided with a short introduction to the context and objectives of the evaluation. Then, we asked users to carry out different tasks with the tool. After completing the tasks, we provided users with a satisfaction questionnaire in order to obtain different dimensions about her/his experience. We utilized the USE questionnaire (Lund, 2001), which includes 30 questions to be assessed through a Likert scale ranging from 1 (Strongly Disagree) to 7 (Strongly Agree). This questionnaire features a multidimensional assessment based on four different dimensions: *usefulness, ease of use, ease of learning* and overall *satisfaction*, providing reasonable psychometric values and a measure for usability through different well-defined constructs (Tractinsky, 2018). Once completed, we proceeded to analyze all the information and extract the corresponding conclusions to answer research questions.

4.2. Variables and research questions

In order to conduct the evaluation, the following dependent variables were considered:

- Quantitative variables:
  - Effectiveness: number of tasks successfully accomplished by users.
  - Efficiency: average time spent for users to complete each task.
  - Normalized values (0–100%) obtained from the USE questionnaire: usefulness, satisfaction, ease of use and learning.
- Qualitative variables:
  - User behavior and observations obtained from the Thinking Aloud protocol sessions.

In addition, we considered the following specific research questions, defined in terms of the above variables. These research questions are

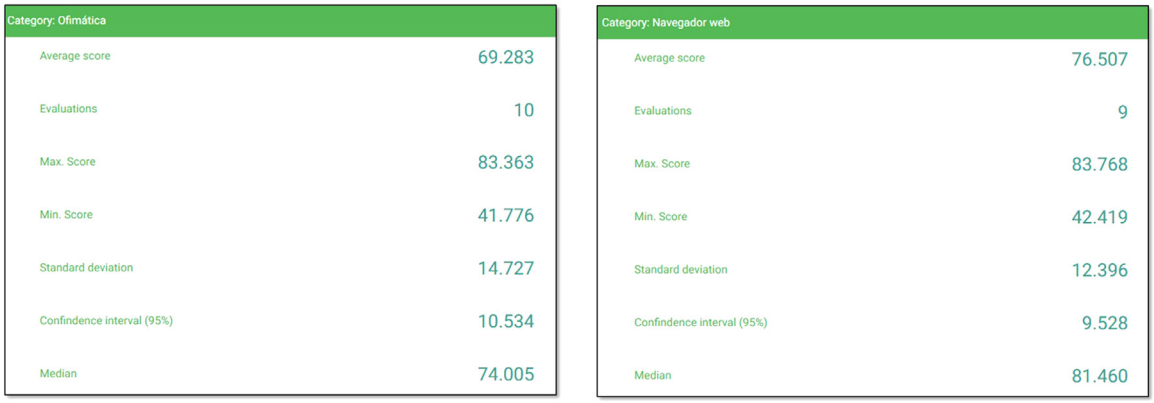

Fig. 6. Individual statistics for two software application categories (from left to right): office application (*ofimática*) and web browser (*navegador web*). Mean, number of evaluations, max and min scores, standard deviation, 95% confidence interval and median values are shown.

Fig. 7. Individual tag clouds for two application categories (from left to right): office application and web browser.

related to RQ3 (feasibility of developing a usable and useful tool featuring the developed metric) described in Section 1.1:

- RQ3.1: To what extent do the users perceive the approach as useful?
  - Validation criteria to be considered as useful: percentages over 75% for usefulness and satisfaction are expected.
- RQ3.2: Can the overall usability of the system be considered as suitable?
  - Validation criteria to be considered as suitable: minimal problems found during user interaction, percentages over 75% for usefulness, satisfaction, ease of use and learning, and suitable effectiveness and efficiency values in tasks assessment.

As stated, we established 75% as a final acceptance benchmark for most usability values. This is a positive benchmark level, higher than others used in usability measurement (Tullis and Albert, 2013; Sauro, 2010), to indicate agreement with respect to user satisfaction when responding to the different questions in a Likert scale; 1–7 for the case of the USE questionnaire. A normalized average measure of 75% represents a number between 5 and 6 (i.e., between agree and very agree), which can be considered as a high value for most usability dimensions.

4.3. Participants and tasks performed

For this evaluation, we enrolled 16 participants, 11 men and 5 women, aged between 22 and 34 (M = 24.08, SD = 3.74), all of them having a university degree related to information technology and working as UX and Software Engineering consultants, which represents the main target for the use of our proposal. We recruited all the participants from our institution and partner organizations, and they participated on a voluntary basis.

According to the binomial probability, commonly used to justify and establish the suitable number of users in usability studies (Nielsen and Landauer, 1993; Dix et al., 2004), we expect to identify problems that impact largely. In fact, a problem impact percentage among 30% − 60% implies problems affecting a great deal of users (i.e., coarse-grain errors), whereas reducing this figure to a more restrictive percentage (10% − 20%) helps find a higher number of problems, and more specifically those being more difficult to find—i.e., less obvious problems (Sauro, 2012). According to that, identifying problems that impact 17% or more users, with a 95% chance of observing them in the evaluation, allows to estimate the number of users to test that can be calculated as Log (1-0.95) / Log (1-0.17) ≈ 16 users. It is worth noting that the discovery rate can be considered as high (95%), and an impact percentage of 17% enables to find complex usability problems affecting a high number of users in most tested situations. This tradeoff would help find most important usability problems, so we think that a sample size of 16 is adequate given the typology of problems that we expect to observe according to the evaluation carried out (Tullis and Albert, 2013; Hwang and Salvendy, 2010; Faulkner, 2003).

All users were asked to perform five different tasks in order to explore both the tool and the results obtained with the metric:

- Registration and login (T_1): The idea is to start from the very beginning, thus we asked the user to first register in the system and then log in to have the role of evaluator.
- Add a new evaluation project (T_2): As an evaluator, the user had to create a new evaluation project, introducing the corresponding information and categories of the application that s/he wants to evaluate. This implies to modify the hierarchical structure and thus update the knowledge base of the system.
- Carry out an evaluation (T_3): According to the project just created, the user was asked to evaluate the selected application, playing the role of the final user, selecting the cards and introducing the corresponding comments.
- Consult the evaluation results (T_4): Once finished, the user was asked to check the evaluation created and the corresponding results.
- Carry out a comparative evaluation (T_5): We asked the user to compare the application evaluated with other existing one. This way, the user had to show statistical results and visualize bar char, box plot and tag clouds of each individual application.

Also, we asked users to analyze the results obtained, providing any comment about the coherence or possible expectations. Comments were registered through the Thinking Aloud protocol.

4.4. Analysis of results

All users successfully accomplished the tasks with no or minimal necessity of help (a couple of questions arose in T1, all related to the password's length), and in a short time, thus obtaining 100% effectiveness.

With respect to efficiency values, Table 3 shows the measures for all the tasks accomplished by users. It is worth highlighting that confidence interval values are lower than one minute in all cases, which indicates that the time measures are quite representative of what it would take a user to complete the tasks. T_4 took the shortest time, needing only few clicks to complete the task, whereas the time took to complete T_3 was the highest, requiring the user to carry out the selection of cards and add comments to them. Also, it can be noted the high deviation related to T_3, since this task's accomplishment heavily depends on the number of cards that each user selects and the time spent to introduce the corresponding evaluation comments, which can differ from one another. On the other hand, T_2, which consisted of adding a new evaluation project, is much more concise, and have the lowest deviation. T_4 presents the peculiarity of having the second higher deviation, depending on whether the user could find or not the shortcut to achieve the task faster. This situation was similar in T_5, where a large number of users ignored the top menu, thus consuming extra time to carry out the task. All in all, time values were measured together with

Table 3
Efficiency results in seconds obtained during the evaluation. Mean, min, max, SD, median and 95% confidence interval values are shown.

	T_1	T_2	T_3	T_4	T_5
Mean	66.22	35.92	183.72	14.80	26.96
Min	45.00	19.00	116.00	5.00	11.00
Max	87.00	65.00	348.00	75.00	70.00
SD	16.12	12.38	81.29	21.00	18.45
Median	66.00	37.00	169.00	13.00	27.50
CI (95%)	11.53	8.85	58.15	15.02	13.20

the application of the Thinking Aloud Protocol, and they are only useful to identify possible usability problems in the tasks proposed rather than achieving a strict performance measurement.

As for the results obtained from the analysis of the USE questionnaire, the overall mean value obtained for the four dimensions is 84% (SD = 0.74), which corroborates that in general the measure of satisfaction obtained is suitable enough, obtaining satisfactory ratings for ease of learning (90%), satisfaction (83%), ease of use (84%) and utility (80%).

On the other hand, there was not any critic situation or transcendental problem during the evaluation to report. Analyzing the comments obtained during the Thinking Aloud sessions, users appreciated the tool's tooltips to guide the interaction. In fact, any issue was solved with the help of the tooltips. Overall ratings have been mainly positive, as users considered the tool as flexible, intuitive, friendly and simple, as well as visually appealing.

According to the results obtained in the evaluation sessions, it is possible to corroborate the corresponding research questions:

- RQ3.1 (related to usefulness perception by users) can be answered in the affirmative, as users considered the approach as useful and satisfying. Average values obtained for utility and satisfaction were 80% and 83%, respectively, while the minimum value expected to validate this claim was 75%.
- RQ3.2 (related to the usability of the system) can be also answered in the affirmative according to the average value obtained for the four USE dimensions, which represents a good estimation of the overall usability of the approach. According to that, average value resulted in 84%. Besides, no remarkable problems were found during the user interaction. On the other hand, effectiveness and efficiency values can be considered high and, as average usability value is over 75% (minimum expected to validate the claim) for all dimensions, RQ3.2 can be also answered in the affirmative.

In summary, all the results obtained helped answer affirmatively RQ3 (feasibility of developing a usable and useful tool featuring the developed metric), and thus affirm that it is possible to develop a usable supporting tool, featuring the developed metric, to carry out comparative usability assessments, being also useful for evaluators.

5. Conclusions

Usability assessment has become an important issue today when developing software applications. Formative and summative usability evaluations (Lewis, 2014) provide significant findings that help designers take into consideration improvements in interactive software design. User testing is the most frequent form of usability evaluation to obtain the user's opinion about software at issue.

Most typical user tests are principally based on tasks performing, where the user is required to accomplish several tasks with the software, in a specific context, in order for the evaluator to obtain metrics concerning effectiveness and efficiency. However, as suggested by ISO 9241-11 (ISO, 2013), satisfaction is also an important concern when carry out usability evaluations. In fact, task performing is not sufficient when carrying out UX evaluations, where more detail about user perception is required in order to obtain other important metrics concerning utility, attractiveness, etc. In general, perceived usability evaluation (Hertzum, 2010) is a key issue when assessing usability through designs and early software developments, as visual appearance and perception have a high impact on the user, and this notably influence the way s/he will use and behave with the software (Bai et al., 2017; Nan and Jun, 2016; Nasoz et al., 2010; Macías and Paternò, 2007).

Although there exist different methods and techniques to evaluate user perception in UX assessments (Melo and Jorge, 2015), they are mostly qualitative and unsystematic, based on guidelines and manual analysis, and requiring a high endeavor from evaluators to interpret, analyze and compare the results obtained after each user evaluation.

The aim of this work is to overcome such drawbacks by concreting the following contributions:

- A quantitative metric for the systematic evaluation of perceived usability in Reaction Cards, a quick and qualitative evaluation method used to assess the user perception on different software applications or designs.
- A supporting tool, called ASSURANCE (usAbility aSsessment SUpported by ReActioN-Cards Evaluations), which exploits the developed metric to carry out evaluations and benchmarking, comparing the perceived usability of different software, application categories or designs, and allowing to obtain further analysis through statistical charts and data, as well as tag cloud to interpret the results.

Our approach is intended to implement the RCs method using its original conception but with several improvements that provide more flexibility in the way it can be applied. The selection or an arbitrary number of cards, instead of the five cards proposed by the original work, provides more flexibility to use our approach to evaluate different kinds of designs and software products. In addition, the existence of a metric based on weighted adjectives avoids to manually interpreting the cards, which may be subjected to ambiguity. These facilities provide a systematic way of measuring, benchmarking and visualizing information that can be arranged regardless of the number of cards used, providing an improved method for measuring perceived usability overall.

Besides, a user inquiry was carried out to provide empirical evidence of each adjective, in order to associate a numerical score to each card in the RCs method, obtaining high reliability and precision. This enabled to create a quantitative metric to study the level of satisfaction perceived by the user, using the metric to numerically and statistically compare different systems or categories, and allowing to establish benchmark levels and graphical comparisons to be further analyzed by usability engineers and software team members for decisions making. As a matter of fact, ASSURANCE features a knowledge base that aggregates data from all evaluations made in order to deal with different types of comparisons and increase statistical significance. On the other hand, the evaluation carried out with users has demonstrated the tool to have acceptable values of usability and overall satisfaction.

All these results helped give an answer to the stated research questions, therefore affirming that is possible to systematize the evaluation of perceived usability, transforming qualitative assessments into a quantitative metric that helps establish benchmarks levels and carry out comparative evaluations of usability for different systems or software designs. In addition, it has been demonstrated that it is possible to develop a usable supporting tool, featuring the developed metric, to carry out such comparative usability assessments, being also of utility to evaluators.

5.1. Limitations

One limitation of the work presented is that it has not been validated with any real concrete UX study to investigate its real impact on this field. Furthermore, the metric should be utilized in a larger number of evaluations in order to observe broader implications and impact. This limitation is mitigated with the verification and validation with both the expert evaluation and the construction of a supporting tool, which has provided acceptable levels of usability. All in all, the aim of this paper is to present the formal development and validation of the metric, as well as the method to carry out benchmarking, which will be utilized for future evaluations of user perceived satisfaction in UX studies.

On the other hand, some readers may wonder why we did not use ASSURANCE to evaluate ASSURANCE. This is not a limitation per se, but it could be an interesting discussion. All in all, the principal reason

to evaluate our tool with well-known usability methods is to fit standard usability assessments, providing a more complete an comprehensive evaluation including not only satisfaction but other common measures such as efficiency and effectiveness.

5.2. Future work

As for future work, we expect to use both the metric and tool in different UX studies. Also, we expect to analyze other different representation mechanisms by utilizing the proposed metric, in order to carry out further analysis and comparisons in our tool. Also, we expect to research other similar metrics in order to be integrated in the tool and thus improve analysis and decision-making in usability and UX assessments.

Acknowledgments

This work was partially supported by the Spanish Government [grant number TIN2014-52129-R]; and the Madrid Research Council [grant number S2013/ICE-2715].

References

Aizpurua, A., Harper, S., Vigo, M., 2016. Exploring the relationship between web accessibility and user experience. Int. J. of Hum. Comput. Stud 91, 13–23.
Adikari, S., McDonald, C., Campbell, J., 2016. Quantitative Analysis of Desirability in User Experience. In: Proceedings of 2016 Australasian Conference on Information Systems, ArXiv.
Bai, X., Cambazoglu, B.B., Gullo, F., Mantrach, A., Silvestri, F., 2017. Exploiting search history of users for news personalization. Inf. Sci 385–386, 125–137.
Baldassarri, S., Macías, J.A., Urquiza, J., 2014. Trending breakthroughs in human-computer interaction. J. Univers. Comput.Sci. 20, 941–1045.
Barnum, C., 2010. Usability Testing Essentials: Ready, Set...Test!. Morgan Kaufmann, 30 Corporate Drive, Suite 400, Burlington, MA 01803 USA.
Barnum, C., Palmer, L., 2010. More than a feeling: understanding the desirability factor in user experience. In: Proceedings of the CHI 2010 Conference on Human Factors in Computing Systems. New York. ACM.
Benedek, J., Miner, T., 2002a. Measuring Desirability: New Methods for Evaluating Desirability in a Usability Lab Setting. In: Proceedings of the 2002 Usability Professionals Association Conference.
Benedek, J., Miner, T., 2002b. Product Reaction Cards. Developed by and © 2002 Microsoft Corporation. All rights reserved, 2002.
Berkman, M.I., Karahoca, D., 2016. Re-assessing the usability metric for User Experience (UMUX) Scale. J. Usability Stud 11, 89–109.
Bodker, S., 2009. A human activity approach to user interfaces. Hum. Comput. Interact 4, 171–195.
Boren, T., Ramey, J., 2000. Thinking aloud: reconciling theory and practice. IEEE Trans. Prof. Commun 43, 261–278.
Cayola, L., Macías, J.A., 2018. Systematic guidance on usability methods in user-centered software development. Inf. Softw. Technol 97, 163–175.
Dix, A., Finlay, J.E., Abowd, G.D., Beale, R., 2004. Human-Computer Interaction. Prentice-Hall, Edinburgh Gate, Harlow, Essex CM20 2JE, England.
Faulkner, L., 2003. Beyond the five-user assumption: benefits of increased sample sizes in usability testing. Behav. Res. Methods 35, 379–383.
Hassenzahl, M, Tractinsky, N., 2006. User experience—a research agenda. Behav. Inf. Technol. 25, 91–97.
Hawley, M., 2010. Rapid Desirability Testing: A Case Study. http://www.uxmatters.com/mt/archives/2010/02/rapid-desirability-testing-a-case-study.php (accessed 26 July 2018).
Hertzum, M., 2010. Images of Usability. Int. J. Hum. Comput. Interact. 6, 567–600.
Huisman, G., Hout, M., 2010. The development of a graphical emotion measurement instrument using caricatured expressions: the LEMtool. In: Proceedings of the 2010 Conference on Emotion in HCI Designing for People.
Hwang, W., Salvendy, G., 2010. Number of people required for usability evaluation: the 10 ± 2 rule. Commun. ACM 53, 130–133.
ISO, 2013. International Standard ISO/IEC 9241-11. Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11: Guidance on.
Lewis, J.R., 2014. Usability: lessons learned … and yet to be learned. Int. J. Hum. Comput. Interact 9, 663–684.
Lewis, J.R., Utesch, B.S, Maher, D.E., 2015. Measuring Perceived Usability: The SUS, UMUX-LITE, and AltUsability. Int. J. Hum. Comput. Interact 31, 496–505.
Li, Y., Wang, X., 2014. Mobile interface studies about style description and influential factors. In: Proceedings of the ICMSE 2014 International Conference on Management Science and Engineering.
Lund, A.M., 2001. Measuring usability with the USE Questionnaire. Usability Interface 8, 3–6.
Macías, J.A., Paternò, F., 2007. Customization of Web Applications Through an Intelligent Environment Exploiting Logical Interface Descriptions. Interact. Comput 20, 29–47.
Melo, P., Jorge, L., 2015. Quantitative support for UX methods identification: how can multiple criteria decision making help? Universal Access Inf. Soc 14, 215–229.
Merčun, T., 2014. Evaluation of information visualization techniques: analysing user experience with reaction cards. In: Proceedings of the BELIV 2014 Conference on Beyond Time and Errors: Novel Evaluation Methods for Visualization. New York. ACM.
Minge, M., Thüring, M., 2018. Hedonic and pragmatic halo effects at early stages of user experience. Int. J. Hum. Comput. Stud 109, 13–25.
Mojoleaf, 2017. http://www.mojoleaf.com/ (accessed 4 March 2018).
Nan, Y., Jun, K., 2016. User experience with web browsing on small screens: Experimental investigations of mobile-page interface design and homepage design for news websites. Inf. Sci 330, 427–443.
Nasoz, F., Lisetti, C.L., Vasilakos, A.V., 2010. Affectively intelligent and adaptive car interfaces. Inf. Sci 180, 3817–3836.
Nielsen, J., Landauer, T.K., 1993. A Mathematical Model of the Finding of Usability Problems. In: Proceedings of INTERCHI 1993 Conference on Human Factors in Computing Systems. New York. ACM.
O'Brien, H., Cairns, P., Hall, M., 2018. A practical approach to measuring user engagement with the refined User Engagement Scale (UES) and New UES Short Form. Int. J. Hu. Comput. Stud. https://doi.org/10.1016/j.ijhcs.2018.01.004.
Sánchez, E., Macías, J.A., 2017. A set of prescribed activities for enhancing requirements engineering in the development of usable e-Government applications. Requirements Eng. https://doi.org/10.1007/s00766-017-0282-x.
Sauro, J., 2010. A Practical Guide to Measuring Usability. Sauro, North Charleston SC, United States.
Sauro, J., 2012. MeasuringU.. https://measuringu.com/ux-benchmarks/ (accessed 26 July 2018).
Seffah, A., Metzker, E., 2004. The obstacles and myths of usability and software engineering. Commun. ACM 12, 71–76.
Spreadsheet, 2017. Spreadsheet to randomise the words,. http://www.userfocus.co.uk/pdf/wordchoice.xls (accessed 26 July 2018).
Tractinsky, N., 2018. The usability construct: a dead end? Hum. Comput. Interact. 2, 131–177.
Travis, D., 2008. Measuring satisfaction: Beyond the usability questionnaire. http://www.userfocus.co.uk/articles/satisfaction.html (accessed 26 July 2018).
Tullis, T., Albert, W., 2013. Measuring the User Experience. Morgan Kaufmann, 225 Wyman Street, Waltham, MA, 02451 USA.
Wordle, 2017. http://www.wordle.net/ (accessed 26 July 2018).

Roberto Veral is a graduate in Computer Science and Engineering by the Universidad Autónoma de Madrid. He works as Big Data Developer and DevOps in Datio, getting in touch with cloud technologies, IaaS, PaaS, Docker, functional programming and Big Data architectures. He has been working in data processing and Big Data for the last year, also collaborating in interactive web projects as full-stack developer.

José A. Macías works as Associate Professor and Researcher in the Universidad Autónoma de Madrid, where he obtained his Ph.D. in Computer Science. He has been working on Human-Computer Interaction for more than 18 years, joining prestigious national and international HCI conferences as member of the program committee and review board, and collaborating in different national and international organizations, such as the Spanish HCI Association, where he serves as President, and the Spanish ACM-SIGCHI, where he participated as co-chair. Furthermore, José A. Macías appears as author on a great number of publications in prestigious HCI conferences, books and Journals.