Hindawi Publishing Corporation
Advances in Human-Computer Interaction
Volume 2016, Article ID 4089520, 16 pages
http://dx.doi.org/10.1155/2016/4089520

Research Article

Combinations of Methods for Collaborative Evaluation of the Usability of Interactive Software Systems

Andrés Solano,[1] César A. Collazos,[2] Cristian Rusu,[3] and Habib M. Fardoun[4]

[1] *Operations and Systems Department, Universidad Autonoma de Occidente, Cali, Colombia*
[2] *Systems Department, Universidad del Cauca, Popayán, Colombia*
[3] *School of Informatics Engineering, Pontificia Universidad Catolica de Valparaiso, Valparaíso, Chile*
[4] *Informatics Department, King Abdulaziz University, Jeddah, Saudi Arabia*

Correspondence should be addressed to Andrés Solano; andresfsolano@gmail.com

Received 3 February 2016; Revised 1 June 2016; Accepted 12 July 2016

Academic Editor: Francesco Bellotti

Usability is a fundamental quality characteristic for the success of an interactive system. It is a concept that includes a set of metrics and methods in order to obtain easy-to-learn and easy-to-use systems. Usability Evaluation Methods, UEM, are quite diverse; their application depends on variables such as costs, time availability, and human resources. A large number of UEM can be employed to assess interactive software systems, but questions arise when deciding which method and/or combination of methods gives more (relevant) information. We propose *Collaborative Usability Evaluation Methods, CUEM*, following the principles defined by the Collaboration Engineering. This paper analyzes a set of CUEM conducted on different interactive software systems. It proposes combinations of CUEM that provide more complete and comprehensive information about the usability of interactive software systems than those evaluation methods conducted independently.

1. Introduction

Currently, the increasingly common tendency is to work collaboratively among people to achieve a common goal. The work is organized into teams and each member interacts with the rest of the group for better productivity [1, 2]. By integrating aspects of collaborative work at a given process, the goal is not only to improve communication but also to achieve greater participation and commitment among members of a group working around a common activity, which leads to the better quality of the finished product [3].

Now, the number of interactive systems is continuously increasing. In this way, usability plays an important role, as the systems must enable users to achieve their goals with effectiveness, efficiency, and satisfaction, always bearing in mind that such systems should be understandable and easy to use. Focusing on the context of design and evaluation of user interfaces, the usability evaluation process is not immune to this trend of working collaboratively. Historically, the discipline of Human-Computer Interaction (HCI) recognizes the need for multidisciplinary teams that allow for a more appropriate assessment. Thus, with the aim of contributing to the traditionally defined process, the *Methodology for the Development of Collaborative Processes* [4] has been used to obtain *Collaborative Usability Evaluation Methods (CUEM)*.

Chilana et al. [5], Woolrych et al. [6], and Hartson et al. [7] have analyzed and implemented several methods to evaluate usability satisfaction degree in different interactive systems. However, the detailed information about the process within these works, such as deliverables, requirements, and roles, is insufficiently defined. Thus, in [8], the collaborative specification of a set of evaluation methods is proposed (and developed following the principles defined in the Collaboration Engineering [4]). This provides a sequence of well-defined activities, collaborative processes (in which several people from different areas of expertise are involved, whom may be geographically distributed), deliverables, participants in the evaluation process, and specification of the communication process (using *thinklets* [9]) between participants.

This is to provide documentation on how to perform usability collaborative evaluations of interactive systems.

The collaboration between practitioners of usability during the evaluation process is of significant importance. The meeting of multiple evaluators in the identification and analysis of usability issues has proved useful in improving the rigor and reliability of the identification of issues [3]. That is, the collaborative work among evaluators increases the likelihood that the vast majority of real issues are found and the consistency in the analysis of results is improved. Also, the inclusion of collaborative work processes related to the usability evaluation can see some benefits, such as [3] obtaining results richer in content, identifying a greater number of usability issues, improving reliability and preventing biased results due to the perspective of a single person, and generating better proposals for redesign and quality assurance. In that sense, linking the application of the Usability Evaluation Methods (UEM) to several people with different expertise and experience proves to be considerably useful [10].

The classical UEM, that allow measuring the implementation of this attribute in a certain system and under certain factors, are quite diverse [11]; their realization depends on variables such as cost, availability of time, and human resources. Thus, the problem arises when deciding which of the existing UEM (inspection and/or test) or which combination of these is suitable for evaluating the usability of interactive systems, so that the greatest amount of relevant information can be obtained, considering reasonable times and obtaining important issues, among other factors. Thus, the research question is how to combine UEM to maximize the potential of identifying relevant information.

Given the above, this work focuses on the study of a set of CUEM on various interactive systems. The areas of application to be used as experimental basis are interactive digital television (iTV), transactional web, and mobile applications. Therefore, this paper presents some CUEM combinations proposed that could provide more complete information on the usability of interactive software systems than those evaluation methods conducted independently. The expectation is that such combinations can be also used to evaluate the usability of interactive systems in other application areas.

Section 2 presents related works. Section 3 describes the *Methodology for the Development of Collaborative Processes*. Section 4 presents the UEM and interactive systems under study. Section 5 describes the application of the CUEM in each of the application areas. Then, Section 6 presents the description of the metrics considered in Section 7, in which the results obtained from the application of the CUEM are analyzed. The CUEM combinations are presented in Section 8; then, Section 9 presents the evaluation of one of these. Finally, Section 10 presents some conclusions and future work.

2. Related Works

Ferré [12] foreground is that software development is increasing recognition of usability as a key factor for the success of software product. However, the UEM that allow us to achieve the desired level of usability in the software product are not regularly applied in an integrated manner in the development process. Software engineering and HCI have disparate approaches to software development, suggesting a major obstacle to the integration of evaluation methods in the overall development [12]. The author's proposal is to integrate UEM in the traditional development process. Thus, developers can know where they can fit activities and UEM in their development process. This research does not propose possible combinations of UEM and integrates collaborative processes in methods; however, it is an important reference for identifying that UEM are useful at different stages of the development cycle of software system (relevant aspect to propose combinations of UEM).

Otaiza et al. [13] have studied a set of UEM on transactional web applications, comparing its characteristics and generating a methodological evaluation proposal. This methodology consists of three combinations of UEM (inspection and testing), depending on the objectives of the evaluation. This research analyzes the importance of combining evaluation methods; however, it does not consider any aspect of Engineering Collaboration in the implementation of the methods. Similarly, other researchers such us Gray and Salzman [14] compare UEM, but these methods not include collaborative processes.

In [8], the researcher has studied a set of UEM in the context of interactive television applications, in order to obtain evaluation methodological proposal. This research proposes combinations of CUEM, which have been designed collaboratively by following the *Methodology for the Development of Collaborative Processes* [4]. Such combinations were considered as a starting point in this paper. Additionally, this work has allowed preliminarily validating the methodology used in the collaborative design of evaluation methods.

Considering the state of the art system, there are few studies that review the evaluation of interactive software systems taking into account the combination of UEM that integrate collaborative processes. Related work supports the idea that integrating collaborative work in activities (which are part of the UEM) contributes to increased productivity in the implementation of methods, which can mean obtaining a greater amount of relevant information [3].

3. Collaborative Design Evaluation Methods

The *Methodology for the Development of Collaborative Processes* [4] has been used to get the *Collaborative Usability Evaluation Methods (CUEM)* [8]. This methodology allows obtaining the collaborative specification of a process [4], in this case, a usability evaluation method. The methodology is composed of the following phases [4]: task diagnosis, task assessment, activity decomposition, task thinklet match, design documentation, and design validation. The methodology allows generating and structuring collaborative processes from identifying recurrent tasks/activities and/or highlighting them. Thus, the specified activities collaboratively promote communication, coordination, and negotiation in order to increase productivity while such activities are performed. The procedure in each phase is as follows.

Phase 1: Task Diagnosis. At this phase, a detailed description of the process (Usability Evaluation Methods) is made. The description includes information about deliverables, requirements, participants, and other relevant data about the process.

Phase 2: Task Assessment. The activities of the studied process are identified and sequenced.

Phase 3: Activity Decomposition. The activities that will be performed in a collaborative way are defined at this phase. One or more *collaboration patterns* are associated with each activity.

Phase 4: Task Thinklet Match. The relationships between *thinklets* and collaborative activities are defined at this phase. Identified *thinklets* should be adapted to resources, the group itself, and even the abilities of the people involved in the execution of the process.

Phase 5: Design Documentation. Based on the information gathered from the previous phases, the elements defined in the Collaboration Engineering are generated [4]: *Facilitation Process Model (FPM)* and *Detailed Agenda*. These documents present the information related to the designed collaborative process.

Phase 6: *Design Validation*. The collaborative process specification is validated. The methodology offers the following ways of validation [4]: pilot testing, walkthrough, simulation, and discussion with colleagues.

4. Usability Evaluation Methods and Interactive Systems under Study

4.1. Selecting Evaluation Methods. Since there are a number of methods to evaluate the usability of interactive systems, it is necessary to select a reduced set of methods to study. To do this, a doctoral work [12] has been used as reference, in which a *total valuation of utility* of the UEM has been performed, based on criteria such as applicability, need for training, representation, and contribution versus effort. Additionally, in the process of selecting the UEM, the strengths and weaknesses of different methods were considered, as well as experiences and recommendations in related research, such as [7, 14–16]. Thus, the selected UEM (of inspection and testing) are as follows:

(1) Heuristic evaluation: it is a usability engineering method for finding the usability problems in a user interface design so that they can be attended to as part of an iterative design process. Heuristic evaluation involves having a small set of evaluators who examine the interface and judge its compliance with recognized usability principles (the "heuristics") [17].

(2) Cognitive walkthrough: it is a usability evaluation method in which one or more evaluators work through a series of tasks and ask a set of questions from the perspective of the user. The focus of the cognitive walkthrough is on understanding the system's learnability for new users [18].

(3) Formal experiments: these are controlled and measurable experiments with test users. Users perform the requested tasks on the system while the evaluators observed the interaction. All necessary information is stored for later analysis (video with the actions and reactions of the user). Thus, it is possible to perform statistical analysis of user actions, considering the time involved and error rate, among others [2].

(4) Constructive interaction: in this method, two users interact together, discovering the characteristics of the system under evaluation, while they verbalize their impressions, like a conversation. Users establish a communication and natural interaction while discovering the system, not limited to a specific list of tasks [19].

(5) Coaching method: in this method, the evaluator (or coach) guides the user in the "right direction while the system is used"; the user can ask the evaluator what he considers necessary and the evaluator must resolve doubts [2].

(6) Interviews: in this method, a conversation between the evaluator and users of the interactive system is established. The interview does not study the user interface itself, but the user reviews this. The interview is conducted in order to obtain information about the users experience with the system, their impressions, preferences, and so forth [16].

(7) Questionnaires: this is a set of questions about the system, which is given by the evaluator to users to obtain conclusions from their answers. The questionnaires try to obtain qualitative and quantitative information about the user experience [16].

Once the UEM are selected under study, the *Methodology for the Development of Collaborative Processes* was used to obtain *Collaborative Usability Evaluation Methods (CUEM)*. Some lessons learned from this process and the notation used in CUEM were already published in [20, 21].

4.2. Selecting Interactive Systems as Case Studies. In order to collect as much information related to the object of target application areas as possible, interactive systems (for each application area) were selected based on the following criteria: appropriateness of the system, availability, and representative tasks.

Interactive Digital Television. The selected applications of iTV are *electronic program guide (EPG)*, *board*, and *chat*. These applications have been developed in the Laboratory of Digital Television of the Universidad del Cauca of Colombia. These are transmitted through the technological standard DVB (Digital Video Broadcasting) (available at https://www.dvb.org/; last accessed on June, 2013), which was adopted in Colombia by the National Television Commission (available at http://www.antv.gov.co/; last accessed on June, 2013) in the year 2008, and which, furthermore, follows the MHP (Multimedia Home Platform) (available at http://www.mhp.org/; last accessed on June, 2013) specification. This means that applications can be displayed on a TV (and not on other devices, such as phones, tablets, etc.) by using a STB (Set-Top-Box), device that allows adapting the digital signal. These applications have been selected because they have many features and a higher level of navigability than other available applications, so it is possible to perform more representative tasks.

Transactional Web. The selected transactional website is *Booking.com* (available at the following URL: http://www.booking.com/), which is a (final) system that offers hotel booking services and operates internationally. This system is available for free, is easily accessible, and provides an appropriate amount of functionality with a good level of navigability. Booking.com was selected because a hotel reservation transaction is a representative example, so the results of the evaluations could be generalized to other systems belonging to the transactional web area.

Mobile Applications. The selected application is Dropbox in its free version for mobile devices, which corresponds to a *Cloud Mobile* application [22]. Dropbox is available in free and paid versions, each of which has varied options, it is of easy access and provides a number of features with a good level of navigability. Dropbox is a file storage system on the Internet for the purpose of performing a backup of them, plus they can be synchronized between multiple devices and shared with other people.

5. Application of the CUEM

5.1. Interactive Digital Television (iTV). The application of CUEM (inspection and test) was subject to the following conditions:

(i) The evaluated iTV applications correspond to functional prototypes at an advanced stage of development.

(ii) The *representative of the organization* provided all the necessary information and was observant of the activities of the *supervising evaluator*. The *representative of the organization* and *supervising evaluator* are defined roles in the collaborative specification of the UEM.

(iii) In the CUEM, the group of evaluators basically consisted of researchers of different topics related to the field of usability and/or iTV.

(iv) Users who participated in the test methods are between 22 and 29 years, with medium/high experience in using information technologies and low experience using iTV applications.

(v) The methods of inspection and testing were performed at the Laboratory of iTV of the University of Cauca, which is easily accessible for the participants of this process. In addition, the laboratory offers optimal conditions (lighting and furniture) for conducting individual and group activities.

(vi) The person responsible for conducting the user testing was the *supervising evaluator*, so other evaluators did not participate in the process of observing the users' actions in the evaluated applications.

(vii) The user tests were recorded and then distributed to the evaluators, which helped in further analysis since no information is lost.

(viii) The hardware devices used in the tests were a 32-inch television, a remote control, and a STB (Set-Top-Box).

5.1.1. Inspection Methods. The heuristic evaluation was performed by a set of 5 evaluators who inspected the interface design of applications based on a number of specific principles for iTV applications [23]. By using this method, 24 usability issues were detected. In general, the level of criticality of the issues is high, a number of issues (16 of 24) were rated, on average, with ratings higher than 6 (on a scale of 0 to 8), and 8 of the 24 issues detected were rated with grades under 6. This may have occurred because the evaluated applications are not completely finished; they are functional prototypes that are in an advanced stage of development. The cognitive walkthrough was also performed by a set of 5 evaluators who conducted a series of tasks in order to check whether the interfaces are suitable for users. This method uncovered 20 usability issues directly related to the ease of learning and use in the applications under study.

5.1.2. Test Methods. In formal experiments, the list of tasks related to the characteristics under study was prepared based on the critical issues identified in the heuristic evaluation. Questionnaires were performed after carrying out the formal experiments. With the application of the questionnaires, no usability issues were identified; however, a number of statistical calculations were obtained based on user responses, related to their subjective satisfaction. The questionnaires allowed obtaining very encouraging results regarding the subjective satisfaction of users with the evaluated iTV applications. The averages do not exceed grade 4 (on a scale of 1 to 5), so it may be said that, in general, users were dissatisfied with the control over the applications.

In the constructive interaction, pairs of users freely explored iTV applications while exchanging their impressions out loud. Due to the nature of the method, 6 interactions were performed. Interviews were conducted after the constructive interactions, in order to obtain information on their perception of various aspects of the iTV applications. From the results of the interviews, it can be stated that the users considered the interaction with the applications to be unfriendly. The insufficient information available and poor navigation throughout the applications were the issues identified in the heuristic evaluation, and that the interviews later confirmed. Finally, the coaching method was performed in which users were led (with use of a preset scenario) while performing a number of tasks. Table 1 presents a summary of the application of the CUEM in the iTV area.

Table 1: Summary of the application of the CUEM in the iTV area.

CUEM	Number of evaluators	Number of users	Number of issues	Number of confirmed critical issues	Number of issues not detected in the heuristic evaluation
Heuristic evaluation	5	Not applicable	24	16	Not applicable
Cognitive walkthrough	5	Not applicable	20	7	9
Formal experiments	4	8	16	8	6
Questionnaires	4	8	0	0	0
Constructive interaction	4	12	25	16	5
Interviews	4	12	12	5	4
Coaching method	4	8	15	11	2

Table 2: Summary of the application of the CUEM in the transactional web area.

CUEM	Number of evaluators	Number of users	Number of issues	Number of confirmed critical issues	Number of issues not detected in the heuristic evaluation
Heuristic evaluation	5	Not applicable	36	6	Not applicable
Cognitive walkthrough	4	Not applicable	21	6	7
Formal experiments	4	10	24	6	11
Questionnaires	4	10	0	0	0
Constructive interaction	4	10	29	6	7
Interviews	4	10	14	5	3
Coaching method	4	8	25	6	9

5.2. Transactional Web. The application of the CUEM was subject to the following conditions:

(i) The evaluated transactional website corresponds to a final system.

(ii) The role of *representative of the organization* was assumed by one of the authors of this paper.

(iii) The heuristics used as defined in [24] are specific to evaluating transactional web applications.

(iv) Inspection methods were conducted in the workplace of the evaluators since these were distributed geographically.

(v) Users who participated in the test methods have the following profile: skills in the use of information technologies, low experience in using backup systems, and aged 21 to 29 years.

(vi) The place where the tests were performed (Universidad del Cauca) is easily accessible to users. This provides appropriate conditions (lighting, furniture, and Internet) for performing the tests.

(vii) The user tests were recorded using the software tool MORAE (software tool designed to ease the process of usability evaluation and data analysis; the characteristics of the tool are described in the following URL: https://www.techsmith.com/morae.html), which facilitated the subsequent analysis of interactions since no information is lost. The recordings were distributed among evaluators so that, if possible, they could analyze the same number of records.

(viii) In the different CUEM, the collaborative activities were conducted virtually since the evaluators were distributed geographically.

(ix) Responsible for conducting user testing was the *supervising evaluator*, so other evaluators did not participate in the process of observing the actions of users on the evaluated transactional website.

(x) The hardware device (personal computer) used by users during the execution of test methods has the following specifications: Dell XPS L421X, Intel® Core™ i5-3317U CPU @ 1.70 GHz, RAM 4 GB, anf 64-bit operating system.

(xi) The software used for the execution of test methods is as follows: Windows 7 Home Premium OS and Google Chrome browser.

Table 2 presents a summary of the application of the CUEM in the transactional web area.

Table 3: Summary of the application of the CUEM in the area of mobile applications.

CUEM	Number of evaluators	Number of users	Number of issues	Number of confirmed critical issues	Number of issues not detected in the heuristic evaluation
Heuristic evaluation	5	Not applicable	17	7	Not applicable
Cognitive walkthrough	4	Not applicable	8	4	1
Formal experiments	4	10	10	4	3
Questionnaires	4	10	0	0	0
Constructive interaction	4	10	16	7	2
Interviews	4	10	7	5	0
Coaching method	4	8	13	6	3

5.3. Mobile Applications. The application of the CUEM was subject to the following conditions:

(i) The evaluated mobile application corresponds to a final system.

(ii) The role of *representative of the organization* was assumed by one of the authors of this paper.

(iii) In the methods of evaluation (inspection and test), the group of evaluators consisted of 4-5 people. In general, the profile of the evaluators meets one or more of the following aspects: high experience in the use of mobile devices with touch technology, medium/high experience in heuristic evaluations (only in the case of heuristic evaluation), and researchers of various issues related to usability, knowledge of the basic features of a mobile application, and experience in the design and application of mobile applications.

(iv) For heuristic evaluation, the heuristics used, defined in [25], are specific to evaluating mobile applications supported on devices with touch technology.

(v) Inspection methods were performed on site (work environment) of each evaluator. For this reason, each evaluation was performed in different conditions related to the hardware device (Smartphone) used, (Android, iOS, Windows Phone) OS, connectivity (data network or Wi-Fi), and lighting.

(vi) In developing the CUEM, it was suggested to the evaluators and users to be at rest to avoid distractions and difficulties when entering data on the mobile device.

(vii) In the test methods, software tool was not used (installed on the device) to record user actions in the mobile application. The tool used for recording user actions was the close observation by the *supervising evaluator*. It is noteworthy that the tests were recorded by video camera to record comments, impressions, and attitudes of users as well as the discussions established with *the supervising evaluator*. The recordings were distributed among evaluators, so that, if possible, all would analyze the same amount of information.

(viii) Collaborative activities were conducted virtually since the evaluators were distributed geographically.

(ix) The place where the tests were performed (University of Cauca-Colombia) is easily accessible to users. This provides appropriate conditions (lighting, furniture, and Internet) for performing the tests.

(x) Responsible for conducting user testing was the *supervising evaluator*, so other evaluators did not participate in the process of observing the actions of users on the evaluated mobile application.

(xi) In the test methods, the mobile device (Smartphone) used by users was a Samsung Galaxy S4.

Table 3 presents a summary of the application of the CUEM in the area of mobile applications.

After running the CUEM on the various interactive systems, the results were analyzed. For the analysis of results, a set of metrics was defined, which are described in the following section.

6. Metrics Description

6.1. Identifying Metrics. For the analysis of the results obtained in the execution of the CUEM, it is necessary to define a set of metrics to objectively measure the results. To do this, after a process of observation and revision of the literature, a series of metrics was obtained which were grouped in the following characteristics: *detection of usability issues*, *human resource*, *equipment*, *time,* and *tasks.*

After defining the preliminary set of metrics, a survey was developed to identify, based on experience and knowledge of an expert group, the most relevant metrics to perform the analysis of results. The survey was developed using the SUS (*System Usability Scale*) system [26], so that each question has five answer choices. Thus, a consensus was carried out among 11 participants with experience in the usability evaluation of interactive systems (that perform at least 3 assessments per year).

6.2. Metrics Selection. Once the survey results were collected and processed (including averages and standard deviation),

Table 4: Description of the metrics to consider in the analysis of the CUEM.

Metric	Description	Interpretation
	Characteristic: detection of usability issues	
Total number of identified issues (TII)	This metric is the total amount of usability issues identified in the evaluated system.	The more usability issues are identified; the metric value is closer to 1.
Number of critical issues (NCI)	This metric corresponds to the number of critical issues identified in the evaluated system.	The more critical issues are identified; the metric value is closer to 1.
Number of frequent issues (NFI)	This metric corresponds to the number of frequent issues identified in the evaluated system.	The more frequent issues are detected; the metric value is closer to 1.
	Characteristic: time	
Time required to complete the planning phase (TPP)	This metric corresponds to the time taken to perform the activities that constitute the *planning phase*.	The less time used to perform the activities of the *planning stage*, the better. It must be defined: TPPc = 1 – TPP, to define relation: less (time) – more (value).
Time required to complete the implementation phase (TIP)	This metric corresponds to the time taken to perform the activities that constitute the *implementation phase*.	The less time used to perform the activities of the *implementation phase*, the better. It must be defined: TIPc = 1 – TIP, to define relation: less (time) – more (value).
Time required to complete the analysis of results phase (TAP)	This metric is the time taken to complete the *analysis of results phase*.	The less time used to analyze the results, the better. It must be defined: TAPc = 1 – TAP, to define relation: less (time) – more (value).

the most relevant metrics were selected as their highest averages. The selected metrics belong to the following characteristics: *detection of usability issues, human resources, and time*. However, in the analysis between the CUEM, the metrics generated by native methods must be taken into account, so the *human resource* metrics characteristic is not considered as criteria for discriminating between CUEM under study. The reason is that these metrics are not related to the method itself, but a test session in which this is used. For example, the number of people involved in the execution of a method (metric *quantity involved*) should not be a criterion for comparing among several CUEM because this would assign a value to a metric that is not generated by the method itself. Table 4 presents the description of the selected metrics.

The selected metrics correspond to base (or direct) measures, and this indicates that they do not depend on any other measure [26]. Metrics which belong to the characteristic of *detection of usability issues* are associated with a type of *absolute scale* [27], since there is only one possible way of measuring: counting; while the metrics of the *time* characteristic are associated with a type of *scale ratio* [27], which has a fixed point of reference: zero (no value may be less than zero). However, once measured, the metric values are not between 0 and 1 (exceeding 1), meaning a table must be used for "*normalization*" to bring them to a range of values between 0 and 1. After normalizing the values, the metrics generate a real number between 0 and 1. Thus, metrics provide positive evidence if the values are close to 1. It is noteworthy that, for metrics whose "good" values are close to zero (in the case of *time*), it would be necessary to perform a calculation like $V_c = 1 - V$. so when value (V) of the metric is closer to zero, complementary value (V_c) is closer to 1 so that all the metrics can be brought to values with a positive (or increasing) direction.

7. Analysis of Results

Based on the application of the CUEM on application areas under study, and considering the evaluation conditions under which these were executed, we have the following comparative analysis of results. Firstly, the metrics that belong to the characteristic of detection of usability issues are analyzed, and, secondly, those related to the characteristic of *time* are analyzed. Regarding the measurement methods, this did not include software tools to automate them. To obtain the measurement metrics, a manual count was performed from an observational monitoring during the application of the UEM (recordings, guide documents, etc.).

7.1. Detection of Usability Issues.

The values of the TII, NCI, and NFI metrics are not between 0 and 1, so *normalization tables* were used (not presented in the paper due to extension restrictions). In the case of the TII metric for each application area, the normalization table takes as a reference value the number of issues identified in the heuristic evaluation. Similarly, normalization tables for NCI and NFI metrics use as reference values the amount of critical and frequent issues identified in the heuristic evaluation. Below the analysis of the TII metric is presented.

7.1.1. Total Number of Identified Issues (TII).

Table 5 presents the normalized measurements of the TII metric according to the application areas under study.

According to Table 5, the heuristic evaluation and constructive interaction helped identify the largest number of

Table 5: Measurements of the TII metric by application area.

CUEM	Normalized TII by application area		
	iTV	Transactional web	Mobile applications
Heuristic evaluation	**0,80**	**0,80**	**0,80**
Cognitive walkthrough	0,65	0,50	0,35
Formal experiments	0,50	0,50	0,50
Questionnaires	0,05	0,05	0,05
Constructive interaction	**0,95**	**0,65**	**0,80**
Interviews	0,35	0,35	0,35
Coaching method	0,50	**0,65**	**0,65**

usability issues in the interactive systems studied, since, in these methods, the systems are evaluated in a global manner. That is, the evaluators (in the heuristic evaluation) and the pairs of users (constructive interaction) freely explore all the features offered by the system, significantly increasing the number of detected issues. The results of Table 5 confirm the good references from the heuristic evaluation, so this method is appropriate to evaluate different interactive systems. However, it should be noted that it is highly necessary to have a set of specific heuristics for evaluating interactive systems.

Regarding the constructive interaction, it is noteworthy that, in each application, area it detected (on average) a percentage higher or equal to 80% of the issues identified in the heuristic evaluation. These positive results could be caused by this method of testing, being carried out more naturally by the pair of users as they verbalize their impressions together. Based on the above, the constructive interaction is appropriate for the overall evaluation of interactive systems.

On the other hand, there are the other methods: coaching, formal experiments, and cognitive walkthrough, which use a list of specific tasks to perform in the system; for this reason, the number of detected issues is minimized (compared to methods that do an overall evaluation) as these correspond only to the features that have been evaluated. Among the aforementioned methods, the coaching method identified the largest number of issues in two areas of application under study. The positive results obtained in the coaching method are due to the fact that the evaluator (or coach) can control the course of the test, in order to discover the information requirements of the users in the target system, considering the critical usability issues identified previously in performed inspection methods (in this case the heuristic evaluation). In general, the coaching method allowed identifying the most issues among those using a list of specific tasks, which is why this is a candidate for forming one of the CUEM combinations.

The methods, constructive interaction and coaching, are possibly carried out in a context that is not realistic, since, for example, in an iTV or mobile applications setting, two users do not interact simultaneously with an application (via remote or mobile device). However, they are methods that work well and allow collecting good data on subjective perceptions of users, since they feel more confident to express their views out loud. In that sense, in different application areas than those studied here, these methods could obtain more appropriate results compared to other "classical" methods on the subject of usability evaluation.

Regarding the methods of inquiry, interviews and questionnaires, these did not identify a significant number of issues; moreover, through the questionnaires, usability issues were not detected. This is because in these methods the questions are directed to information on the subjective satisfaction of the users; thus, the number of usability issues identified is significantly reduced. However, these methods of interrogation are a good complement to other test methods for additional information. For example, in the case of the questionnaires from consolidated statistical results (calculated based on the responses of users), it is possible to obtain information on the subjective satisfaction of users.

In summary, according to the results of the metrics associated with the characteristic of *detection of usability issues*, the methods of constructive interaction and coaching, despite having a considerable degree of subjectivity, show positive results over formal experiments, which may indicate that appropriate results (feedback) are obtained with evaluation methods that promote direct interaction with users.

7.2. Time.

The values of the TPP, TIP, and TAP metrics are not between 0 and 1, so a normalization table is used to bring them to values between 0 and 1. Also, the complement must be calculated to define the relationship: less (time) – more (value). Below, the analysis of the TAP metric is presented.

7.2.1. Time Required to Complete the Stage of Analysis of Results (TAP).

Table 6 presents the normalized measures of the TAP metric according to application areas under study.

According to Table 6, the interviews, questionnaires, and cognitive walkthrough required the least amount of time for result analysis. Therefore, the average time spent by the evaluators analyzing and interpreting the data collected (such as recordings, consolidated results of the questionnaires, and annotations in guide documents) is relatively low. Regarding the interviews in the transactional web and mobile applications areas, it is worth mentioning that the short time taken in the analysis of results was not necessarily related to the characteristics of the method but to the limited availability of the evaluators to participate in the suggested collaborative activities.

The interviews and questionnaires allow obtaining additional information (qualitative and quantitative) to the

Table 6: Measurements of the TAP metric by application area.

CUEM	Normalized TAP by application area		
	iTV	Transactional web	Mobile applications
Heuristic evaluation	0,50	0,78	0,83
Cognitive walkthrough	**0,80**	**0,92**	**0,91**
Formal experiments	0,40	0,36	0,29
Questionnaires	**0,90**	0,85	**0,91**
Constructive interaction	0,70	0,85	0,67
Interviews	**0,90**	**0,92**	**0,99**
Coaching method	0,60	0,78	**0,91**

execution of other test methods; however, these interrogation methods can identify a limited number of issues because they are focused on information about the subjective satisfaction of users (experience with the system, perspectives, impressions preferences, etc.). However, given the positive results in the metrics associated with the characteristic of *time*, the interrogation methods (interviews and questionnaires) are well suited to complement the performance of other test methods, because their preparation, execution, and analysis do not demand a significant amount of time.

The time spent on analysis activities is subject to the amount of information collected and amount of usability issues identified. In this regard, if a significant number of issues are identified, it takes evaluators longer to analyze each one. Given the above, other methods with positive results regarding the TAP metrics are the coaching and constructive interaction methods, since the time of data analysis (test records) is directly related to the duration of the records (evaluators need to visualize the user interaction with the system and then make the respective analysis and interpretation of the actions). This corresponds to the methods mentioned above which also yielded positive results in the TIP metric.

Furthermore, the methods that require more time when analyzing the results (due to their low grades, as shown in Table 6) are heuristic evaluation and formal experiments. On one hand, heuristic evaluations used a considerable period of time in which the evaluators made their contributions based on the analysis of information (ranked in criticality, severity, and frequency); they identified positive elements of the systems evaluated, among other activities. On the other hand, during the formal experiments (slowest method in the stage of analysis of results), evaluators analyzed the actions performed by users in each of the proposed tasks, the time taken in achieving each task, and cases of failure, among other information, which required each evaluator to devote a significant amount of time.

Regarding the testing methods, the time devoted to the analysis of results depends on the number of users participating in the evaluations. Moreover, the average time spent by the evaluators depends largely on the way the information is distributed among them. On the other hand, the results obtained in this metric might indicate that collaborative work contributes in a good way to the time required analyzing the information collected in the application of the CUEM. This is because the information is distributed among several people (group of evaluators) which reduces the time and effort spent by the head of the evaluation; additionally, the analysis of information would not be limited to the perception of just one person. Furthermore, it could be estimated that combinations of proposed *thinklets* allow obtaining positive experiences in the time required to obtain a series of contributions by the evaluators.

The times involved in each method can be determinant to choose one of these over the other, especially when time is scarce, which occurs in most cases. Therefore, the results of the metrics associated with the characteristic of *time* will support the identification of the CUEM, which will form the following combinations of methods.

8. Proposing CUEM Combinations

8.1. Discussion. The implementation and analysis of results of the CUEM made in the application areas under study (iTV, transactional web, and mobile applications), as well as comparative analysis between them, have allowed to propose a series of combinations of evaluation methods (for inspection and testing).

The most common way to do usability evaluations is to combine inspection methods with test methods, depending on the scenario presented [7, 13]. In this regard, combinations of CUEM must include at least one of the methods of inspection and one of the usability testing methods.

First, regarding inspection methods, the heuristic evaluation identifies a number of usability issues through inspections of the evaluated system [28]; their ability to find issues at different levels (major and minor) and that breach various usability principles support this inclusion. Regarding the cognitive walkthrough, considering the characteristic of *detection of usability issues* and *time*, this did not present remarkable qualifications in any of the two characteristics except in the TAP metric, since it employed (on average) the third least amount of time in the analysis of results. Based on the above, the positive results regarding the *detection of usability issues* favor the heuristic evaluation. For this reason, this inspection method will be part of the combinations as the first of the methods to be performed.

The heuristic evaluation stands out for its ability to help find problems, while its disadvantage is the influence of the system domain [29]. Evaluators can have high experience in this type of evaluation, but if they do not know the business rules or do not have a set of specific heuristics, several usability issues may not be identified. Another important advantage of the heuristic evaluation is its ease of execution compared to other inspection methods [30]; however, a major disadvantage is the time required for execution and analysis of results. In this work, just 2 inspection methods were considered, but the literature raises the capabilities of the heuristic evaluation over other methods, besides being the most used inspection method for detecting problems [3].

Second, regarding the testing methods, the evaluation methodology must include at least one of the methods that perform direct user interaction with the system, that is, formal experiments, constructive interaction, or coaching. This aspect will be discussed later on.

Among the test methods, which do not consider direct interaction with the system, are interviews and questionnaires. These interrogation methods allowed identifying problems related to the design, navigation, and other aspects of the systems evaluated, in addition to information related to the subjective satisfaction of users after interacting with the system. This last feature is key when proposing possible combinations of CUEM, because such interrogation methods work to complement test methods that perform direct interaction with the system. In addition to the above, interviews and questionnaires are appropriate to complementing the execution of other test methods as their preparation, execution, and analysis do not require a significant amount of time.

Considering the above, the discussion will focus on defining the intermediate test methods to be performed to get as much relevant information about the usability of the evaluated interactive system as possible (considering reasonable times and obtaining significant issues). This is because it has been established that the initial method is the heuristic evaluation and the final methods are the interviews and/or questionnaires.

In each of the application areas under study, the formal experiments, constructive interaction, and coaching method have allowed to confirm, in general, critical issues identified in the heuristic evaluation, in addition to finding some other problems (see Tables 1, 2, and 3). However, there are some marked differences like the following. The tests conducted in different application areas indicate that the constructive interaction found more usability issues than other test methods, and it also allowed to empirically confirm the most critical issues identified in the heuristic evaluations. Now, formal experiments and the coaching method also achieved the same as the constructive interaction, but on a smaller scale. Between the two methods mentioned above, the coaching method identified a greater number of issues in two application areas under study (transactional web and mobile applications). In general terms, the coaching method identified the most issues between those that used a list of specific tasks. Therefore, it would be appropriate to use this method when the objective is to evaluate a number of specific tasks or functions of a system.

Regarding the characteristic of *time* for the 3 test methods mentioned above, the results obtained indicate that the coaching and constructive interaction require less time than formal experiments. Constructive interaction does not need task design for completion, while formal experiments and the coaching method themselves do need it, which increases the time spent in the preparation of such methods. Now, formal experiments demand more time in the *stage of analysis of results* mainly due to the statistical studies performed and the analysis of user actions on each task, among other activities. Thus, constructive interaction and coaching method stand out between the methods which include direct interaction between the user and the evaluated system.

Another important factor when proposing combinations of methods consists of the scope of the evaluation, that is, the system functionalities to be evaluated. Constructive interaction is not limited to a pair of users to focus on specific features of the system; on the contrary, it allows a complete analysis of these, for which many more usability issues are detected than in formal experiments and the coaching method. Clearly, the same does not happen with formal experiments and the coaching method. The list of tasks associated with these methods restricts the interaction to the parts where the design tasks suggest, which is not a disadvantage but obviously does not allow a full (global) evaluation of the system,but of a set of specific areas or features. It is noteworthy that the list of tasks associated with the coaching method is more flexible than in formal experiments, since the evaluator (or coach) can control the course of the test depending on how the user completes the proposed tasks.

Based on the arguments presented above, 3 combinations of CUEM have been proposed which can be useful, depending on the objectives of the evaluation. Such combinations are presented below.

8.2. CUEM Combinations

8.2.1. Global Evaluation: High Detection of Issues. This combination is focused on analyzing a system completely and includes the following methods: *heuristic evaluation + constructive interaction + interviews*. This combination is expected to work correctly when a global type analysis is required, in which a number of usability issues will be identified, both by the evaluators and by the analysis of the interaction of representative users. The interviews, as a method of supplementary interrogation, will allow additional/supplemental information about user perceptions regarding the target system, which would also be possible to confirm critical issues identified by the two methods previously conducted.

In this evaluation, the constructive interaction has within its advantages the ability to obtain results involving few (6 or more) representative users. This is based on the expression of the impressions of users while performing the interaction. Working the users in pairs increases the fluidity of comments; therefore, a greater number of usability problems can be identified. An important factor is also the time, as the constructive interaction requires less time than formal experiments for preparation and analysis of results. Moreover, constructive interaction identifies the reasons/causes of the issues, which is a major advantage as it helps in confirming critical issues identified in the heuristic evaluation.

Table 7: Software components of an interactive system.

General group	Specific components
Content type	Forms, tables, lists, dates, times, numeric values, currency signs.
Information	Images, news, graphics, text, formatting, URL, abbreviations, audio, nomenclatures, colors, icons.
Data management	Transmission of information, registration form, login form, information updates, data validation, recovery and/or backup information.
Search	Search form, search results.
Navigation area	Pages, titles, cursor, shortcuts.
Emergency exits	Without associated components.

Table 8: Hardware components of an interactive system.

General group	Specific components
Operating system	Without associated components.
Browser	Without associated components.
Input/output devices	Printer, digital certificate, electronic card, mouse, screen, keyboard, microphone, scanner, kinect.
Technical assistance	Icons, assistance hardware.
Audio	Volume.
Indicators	Without associated components.
Help and documentation	Without associated components.

8.2.2. Specific Evaluation: Time Reduction. This combination is directed to evaluate certain scenarios or functionalities of a system, and it includes the following methods: *heuristic evaluation + coaching method + questionnaires*. This combination, the only one evaluated between the 3 combinations proposed, turns out to be very useful to evaluate specific functionalities because the information obtained by the coaching method allows detecting issues at those points where the user requests help/information to the evaluator (coach). That is, at the points where there is communication between the user and coach, it is very likely that there is a need for information in the system. As for the evaluation of this combination, the *specific evaluation* obtained more appropriate results than a series of methods proposed by experts, given a specific scenario of evaluation (see Section 9).

In this combination of CUEM, the heuristic evaluation identifies design issues or details of presentation that can impede the progress of users to perform a task. Via the coaching method, it is possible to identify the differences between the conceptual model of the system and the mental model of users. In this method, the test users perform the tasks requested following their mental model, which, in many cases, generates differences due to improper modelling of the system, which prevents or hinders users from performing the tasks.

8.2.3. Evaluation Focused on Specific Tasks: No Time Restrictions. This combination aims to analyze specific tasks of an interactive system, and it includes the following methods: *cognitive walkthrough + formal experiments + questionnaires*. In this combination, the three methods provide important features, but it is the formal experiments that make a difference to the *specific evaluation*. The formal experiments allow efficient analysis of the tasks of interest. They have a good level of objectivity and adequately complement the questionnaires (pretest and posttest), which also have good objectivity and allow for quantitative information. In this way, when the 3 methods that make this combination are performed, it is estimated that the information obtained on the usability of the tasks would be completely objective. However, this combination should be used when the available time is high.

By implementing this combination of CUEM, accurate information about the usability of a set of specific tasks will be obtained, because inspections will be performed by expert evaluators, statistical analysis can be performed based on user actions through formal experiments, and, finally, information will be obtained on subjective perceptions (qualitative and quantitative) of representative users.

8.3. Applicability of the CUEM Combinations. The combinations of CUEM proposals can be carried out in virtually any stage of the development cycle of an interactive system, although it would probably better suited in the early stages, when a functional prototype (not necessarily a final version) permits testing with actual users. It is suggested that the evaluated system has some degree of progress or functionality to enable experts to evaluate it better and obtain more complete results.

On the other hand, according to usability evaluations conducted in the areas of iTV, transactional web, and mobile applications, a number of components susceptible to a usability evaluation have been identified. Based on related work [31], Tables 7 and 8 present a set of software and hardware components that may be in different types of interactive systems. In this regard, the combinations of methods may be used to evaluate the usability of interactive systems that include any of the hardware and software components listed in the tables.

The hardware interfaces, like any physical device, allow us to interact with them. Examples include elevator panels and ATMs. It should be noted that these physical systems are complete interactive systems with a large set of components and a software part. Based on the above, the hardware interfaces are classified into [31] operating system, browser, input/output devices, technical assistance, audio, indicators, and help and documentation. Thus, each category is divided into several components (see Table 8) to allow complete classification of an interactive hardware system in order to facilitate its evaluation.

These component classifications are not mutually exclusive; they are a complementary classification due to the hardware and software components of a system which may be in a concrete interactive system.

9. Preliminary Evaluation of the CUEM Combination: Specific Evaluation

The *specific evaluation* combination was evaluated considering social networks as an area of application. First, a survey was conducted to identify, based on the experience and knowledge of a group of experts, the most appropriate combination of UEM to use in a specific scenario of evaluation. Then, secondly, the CUEM comprising the combination suggested by the experts were executed. Thirdly, the *specific evaluation* combination was performed according to the scenario of evaluation. Finally, in fourth place, the results of evaluations were compared to determine if the proposed combination allows obtaining better results than the methods suggested by the experts.

9.1. Expert Consensus. Table 9 presents the percentage of votes obtained on the consulted UEM. The survey was filled out by 13 participants, who have high-level expertise about usability evaluation of interactive systems and user-centered design.

According to the results shown in Table 9, the UEM with higher percentages of votes ar: cognitive walkthrough, think-aloud, and questionnaires. Therefore, these methods were executed and then the results were compared with those obtained in the *specific evaluation* combination.

Table 9: Percentage of votes.

Evaluation method	Percentage (%) of votes
Inspection methods	
Heuristic evaluation	46
Cognitive walkthrough	**61**
Pluralist walkthrough	7
Inspection standards	15
Analysis of actions	23
Test methods	
Formal experiments	15
Think-aloud	**61**
Constructive interaction	0
Recording of use	30
Coaching method	7
Measure of performance	7
Retrospective test	15
Questionnaires	**46**
Interviews	30

9.2. CUEM Application Summary. Table 10 presents, in short, information about the application of the CUEM (suggested by experts and those that comprise the *specific evaluation*) in social networks area.

9.3. Comparative Analysis of Results

9.3.1. Detection of Usability Issues. Table 11 presents the measurements of the metrics associated with the characteristic *detection of usability issues*.

First, regarding the inspection methods implemented, the heuristic evaluation identified a greater number of issues (30) compared to the cognitive walkthrough (18). Due to the nature of the heuristic evaluation to evaluate different aspects of a system, 7 problems were identified that are not entirely related to the analyzed functionalities. Thus, considering the scenario of evaluation, the heuristic evaluation identified a total of 23 problems. Similarly, in the cognitive walkthrough, 2 issues were detected that are not related to the evaluated functionalities, so it is considered that this method identified a total of 16 usability issues. In this sense, regarding the TII metric, the heuristic evaluation allowed the detection of 7 additional issues compared to the cognitive walkthrough.

The cognitive walkthrough is highly subjected to the experience of the evaluators as these require knowledge and skills to detect learning issues associated with the functionalities or tasks evaluated. In contrast, the heuristic evaluation is easier to perform because it is subject to the heuristics used and detail on elements (subheuristics) that are part of the checklist. Based on the above, the heuristic evaluation would allow obtaining more complete information (and in an easier way) on the usability of an interactive software system compared to the cognitive walkthrough. Thus, the evaluation shows that the inspection method (heuristic evaluation) included in the combination *specific evaluation* allows obtaining appropriate results when it is focused on a set of specific tasks.

Second, regarding the testing methods, the coaching method identified a greater number of issues (highest value in the TII metric) compared to the think-aloud method. However, regarding the NCI and NFI metrics, the difference between these two methods is minimal. This suggests that both methods work properly confirming the system's critical issues previously identified by an inspection method. In this regard, other aspects are analyzed as shown below.

What the think-aloud and coaching methods have in common is that the user speaks their impressions during the interaction with the system through a list of previously designed tasks. Now, the evaluation process confirmed that the main problem of thinking aloud is that the user's verbalizations significantly interfere in the normal use of the system. Additionally, this method relies heavily on the spontaneity of the user to express all of their impressions while performing the suggested tasks. Thus, the coaching method overcomes

Table 10: Summary of the application of CUEM in the area of social networks.

Combination	CUEM	Number of evaluators	Number of users	Number of issues	Number of confirmed critical issues	Number of issues not detected in the heuristic evaluation
Suggested by experts	Cognitive walkthrough	4	Not applicable	18	11	9
	Think-aloud	4	10	13	8	0
	Questionnaires	4	10	0	0	0
Specific evaluation	Heuristic evaluation	5	Not applicable	30	11	Not applicable
	Coaching method	4	10	21	9	1
	Questionnaires	4	10	0	0	0

Table 11: Measurements of the metrics associated with the characteristic *detection of usability issues*.

Combination	CUEM	TII	NCI	NFI
Suggested by experts	Cognitive walkthrough	18 (16)	11	13
	Think-aloud	13	8	10
	Questionnaires	0	0	0
Specific evaluation	Heuristic evaluation	30 (23)	11	13
	Coaching method	21	9	11
	Questionnaires	0	0	0

these disadvantages, since a more "normal" conversation with the user is set during application. However, the drawback is that, generally, it is not two users who are interacting with the system. The positive point is that the feedback obtained by the coaching method poses a greater detection of issues, since the evaluator (who assumes the role of the coach) has greater control over the development of the tests, emphasizing on tasks that could not be completed successfully.

Considering the above, it is estimated that the feedback obtained with the coaching method is of significant relevance compared to records obtained by the think-aloud method. In that sense, the evaluation shows that the test method (coaching) included in the *specific evaluation* combination would allow the analysis and more appropriate interpretations of the actions and impressions of users, which could identify a greater amount of usability issues.

From the inspection methods implemented, it is noted that the heuristic evaluation can be performed at a low cost, depending on the number of "expert" evaluators performing the process (3 to 5 as suggested to detect most usability issues). For this reason, the cost would be lower than other CUEM, which require end users to be performed. On the other hand, regarding test methods, it is important to note that the coaching method got one of the lowest percentages of votes according to Table 9, while think-aloud obtained the highest percentage of votes. This indicates that the coaching method is not commonly used in practice, so the obtained results prove the adequate function of this method and promote its use.

Finally, in third place, the questionnaires did not detect usability issues; however, they obtained additional information on the subjective satisfaction of the users regarding the use and tasks of the system. This reveals that this interrogation method works as a good complement for the test methods to capture quantitative and qualitative data.

9.3.2. Time. Regarding the characteristic of *time* in the area of social networks, the collaborative activities that make up the executed CUEM were made virtually (using shared documents in Google Docs) because the evaluators were distributed geographically. Therefore, contributions by the evaluators were collected over a considerable period of time (4 days on average).

Table 12 presents an approximation of the time spent (in minutes) by group participants to perform the activities that comprise the stages of planning, implementation, and analysis of results of the executed CUEM.

Overall, regarding the methods of inspection, the cognitive walkthrough used less time than the heuristic evaluation. For example, the TPP is the metric that shows the greatest difference between these two methods, since planning the heuristic evaluation includes some activities that require a little more time (close-up of the evaluated system, selection of appropriate specific heuristics, and development of the guide document for the evaluators). Based on the above, the cognitive walkthrough should be selected if the time for evaluation is limited; however, the time taken for planning the heuristic evaluation could be reduced because the current literature has proposed a number of specific heuristics to evaluate different interactive systems (using the methodology defined in [32]), such as virtual worlds [33], transactional web applications [24], and grid computing applications [34]. Thus, if the system to evaluate corresponds to any of the above, it is estimated that planning is less delayed and the possibility of identifying a greater number of usability issues increases (no need for high-level experts).

Table 12: Measurements of metrics associated with the characteristic *time*.

Combination	CUEM	TPP	TIP	TAP
Suggested by experts	Cognitive walkthrough	219	90	60
	Think-aloud	151	34	92
	Questionnaires	82	10	62
Specific evaluation	Heuristic evaluation	304	123	96
	Coaching method	180	31	83
	Questionnaires	79	10	60

Now, regarding the testing methods, the TPP metric presents the largest difference between the coaching method and think-aloud method, benefiting the latter. This is mainly because, in the coaching method (as defined in the collaborative specification for this method), the *supervising evaluator* (which assumes the role of the coach) plans a specific scenario for evaluation and becomes familiar with the sequence of actions associated with the tasks to be performed by the user, then asking about those actions which are performed during the test. On the other hand, regarding the TAP metric, in the coaching method, the analysis of a test session would require less time because the evaluators focus on the actions that the user could not complete in achieving a task.

As for the questionnaires, as an interrogation method complementary to the test methods, these used the least amount of time during the stages, which form the evaluation process (planning, implementation, and analysis of results). In general, the preparation of questionnaires consists of only a few activities, as they have the information immediately, with which it is possible to quickly analyze user feedback.

10. Conclusions and Future Work

The *Collaborative Usability Evaluation Methods (CUEM)* attempt to strengthen collaboration between the different members of a group; that is, these methods promote communication, coordination, and cooperation in order to increase the productivity of the evaluators. The collaboration allows group members to unite intellectual efforts to find a common goal, which in this particular case is to evaluate the usability of interactive systems more accurately.

The conduction of a process designed collaboratively increases the possibility of more complete and rich in content results, compared to a process that does not include aspects of collaborative work. Through this research, a set of appropriate CUEM was identified to conduct in different application areas (interactive digital television, web), which can be performed in the traditional way (as defined), obtaining significant results. However, the conduction of these methods designed collaboratively allows more appropriate results regarding the number of identified usability issues and time spent on analysis of information, this considering the advantages of collaborative work.

In this work, a set of *CUEM* were performed on 3 areas of application, presenting the analysis of the results obtained based on a number of metrics. It is important to note that there is no "best method" and all have strengths and weaknesses and are focused to assess specific aspects of usability, for that reason, according to [12], combining them is the most appropriate procedure. Therefore, we propose three CUEM combinations according to different evaluation scenarios. This is a starting point that is considered relevant because, as stated in [10], there are a variety of UEM but are necessary studies that compare and analyze their efficiency in different areas and application contexts, from the viewpoint of *formative usability*.

On one hand, traditional (noncollaborative) UEM do not define clearly and specifically roles and responsibilities for the different actors involved in the evaluation process. In this research, the CUEM provide a sequence of well-defined activities, specification deliverables, description of the different participants in the evaluation process, and specification of the communication process between participants. On the other hand, the documentation (guidelines) about how to conduct collaborative usability evaluations of interactive systems is scarce. The results we present may help usability practitioners and/or persons responsible for the usability evaluation process of interactive software systems.

The CUEM combinations would cover the most critical points for the measurement of usability with an acceptable level of accuracy. First, they include at least an inspection method and a testing method. Second, they include methods which perform quantitative and qualitative, objective and subjective, and global and specific evaluation of the system. Therefore, it is possible to say that they are covering all the necessary factors to extensively evaluate the usability of an interactive software system.

In the first two proposed combinations (global and specific evaluation), the methods, constructive interaction and coaching, are possibly carried out in a context that is not realistic, since, for example, in an atmosphere of iTV or mobile applications, two users would not simultaneously interact with an application (via remote or mobile device). However, they are methods that work well and allow collecting good data on subjective perceptions of users, since they feel more confident to express their views out loud. Whereupon, in different application areas than those studied in this research, these methods could obtain more appropriate results compared to other "classical" methods on the topic of usability evaluation (i.e., think-aloud).

Findings from the evaluation process with the combination of CUEM are as follows: *specific evaluation* has been successful in the sense that it allowed the detection of a greater number of issues regarding a combination suggested by people with experience in usability evaluations of interactive systems. The test method (coaching), which is part of the *specific evaluation*, has confirmed a good number of the critical issues identified by the inspection method (heuristic evaluation). How that information is obtained is different compared to other test methods; the important thing is that it has complied with expectations. Now, with respect to the time factor, the *specific evaluation* combination employs more time in the *planning stage* than the combination suggested by experts; however, the time spent in the execution and analysis of results did not show a significant difference.

In the metrics related to the characteristic of *time*, it is debatable whether a "small" value is a positive fact, as it could doubt the quality of the deliverables generated in each activity. Similarly, the performance speed of the stages of an evaluation method (planning, implementation, and analysis of results) is not a value in itself without ensuring a minimum quality in the results derived from them. For this reason, the *supervising evaluator* (role defined in the collaborative specification of the CUEM) is responsible for ensuring a certain degree of quality in the results/deliverables obtained.

The combinations of CUEM were proposed considering the characteristics of *detection of usability issues* and *time*, in the evaluation methods executed. A set of metrics corresponding to each of these characteristics was studied to observe the behavior of the CUEM studied in different application areas. As for future work, if it is desired to propose additional combinations of CUEM considering other factors/characteristics, it would be convenient to implement a process like the one conducted in this case, which basically consists of studying the behavior of the CUEM based on the new factor defined in different application areas. On the other hand, it would be interesting to have one or more metrics associated with the impact of collaboration in the process of using the analyzed evaluation methods.

Given that only one combination of CUEM was evaluated (*specific evaluation*), as future work, the combinations *global evaluation* and *evaluation focused on specific tasks* are yet to be evaluated through specific case studies. Additionally, we intend to experimentally validate the three combinations in other application areas.

In this work, the *usability* facet has been considered as the core. We would like to extend the scope of the research, switching from "combinations of methods for collaborative evaluation of the usability" to "combinations of methods for collaborative evaluation of the user experience." Thus, for further work, it would be appropriate to include (or combine) elements of other facets, such as emotional, multiculturalism, and playability, to the collaborative specification of the UEM.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

This work has been supported by the IDIS (Research and Development in Software Engineering) research group of the Universidad del Cauca (Colombia) and GITI (Research Group in Telematics and Applied Informatics) research group of the Universidad Autonoma de Occidente (Colombia). This work has been partially supported by projects "*UsabiliTV-Framework para la evaluación de la usabilidad de aplicaciones en entornos de Televisión Digital Interactiva*" (COLCIENCIAS, code 110352128462), "*propuesta metodológica para la evaluación colaborativa de la usabilidad de sistemas software interactivos*" (Research Division, Universidad Autonoma de Occidente, Colombia) and "*convocatoria para el apoyo a proyectos con Asia 2014*" (COLCIENCIAS, Colombia).

References

[1] C. A. Ellis, S. J. Gibbs, and G. L. Rein, "Groupware some issues and experiences," *Communications of the ACM*, vol. 34, no. 1, pp. 38–58, 1991.

[2] T. Granollers, *MPIu+a una metodología que integra la ingeniería del software, la interacción persona-ordenador y la accesibilidad en el contexto de equipos de desarrollo multidisciplinares [M.S. thesis]*, Departamento de Sistemas Informáticos, Universidad de Lleida, Lleida, Spain, 2007.

[3] A. Følstad, E. L.-C. Law, and K. Hornbæk, "Analysis in practical usability evaluation: a survey study," in *Proceedings of the 30th ACM Conference on Human Factors in Computing Systems (CHI '12)*, pp. 2127–2136, May 2012.

[4] G. Kolfschoten and G.-J. D. Vreede, "The collaboration engineering approach for designing collaboration processes," in *Proceedings of the International Conference on Groupware: Design, Implementation and Use*, Bariloche, Argentina, September 2007.

[5] P. K. Chilana, J. O. Wobbrock, and A. J. Ko, "Understanding usability practices in complex domains," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*, pp. 2337–2346, ACM, Atlanta, Ga, USA, April 2010.

[6] A. Woolrych, K. Hornbæk, E. Frøkjær, and G. Cockton, "Ingredients and meals rather than recipes: a proposal for research that does not treat usability evaluation methods as indivisible wholes," *International Journal of Human-Computer Interaction*, vol. 27, no. 10, pp. 940–970, 2011.

[7] H. R. Hartson, T. S. Andre, and R. C. Williges, "Criteria for evaluating usability evaluation methods," *International Journal of Human-Computer Interaction*, vol. 15, no. 1, pp. 145–181, 2003.

[8] A. Solano, *Metodología para la evaluación colaborativa de la usabilidad de sistemas software interactivos [Ph.D. thesis]*, Departamento de Sistemas, Universidad del Cauca, Popayán, Colombia, 2015.

[9] G. L. Kolfschoten, R. O. Briggs, and G. Vreede, "Definitions in collaboration engineering," in *Proceedings of the 39th Hawaii International Conference on System Sciences*, Kauia, Hawaii, USA, 2006.

[10] J. R. Lewis, "Usability: lessons learned... and yet to be learned," *International Journal of Human-Computer Interaction*, vol. 30, no. 9, pp. 663–684, 2014.

[11] T. Hollingsed and D. G. Novick, "Usability inspection methods after 15 years of research and practice," in *Proceedings of the 25th ACM International Conference on Design of Communication (SIGDOC '07)*, pp. 249–255, October 2007.

[12] X. Ferré, *Marco de integración de la usabilidad en el proceso de desarrollo software [Ph.D. thesis]*, Lenguajes y Sistemas Informáticos e Ingeniería del Software, Universidad Politécnica de Madrid, Madrid, Spain, 2005.

[13] R. Otaiza, C. Rusu, and S. Roncagliolo, "Evaluating the usability of transactional web sites," in *Proceedings of the 3rd International Conference on Advances in Computer-Human Interactions*

(ACHI '10), pp. 32–37, Saint Maarten, The Netherlands, February 2010.

[14] W. D. Gray and M. C. Salzman, "Damaged merchandise? a review of experiments that compare usability evaluation methods," *Human-Computer Interaction*, vol. 13, no. 3, pp. 203–261, 1998.

[15] W. O. Galitz, *The Essential Guide to User Interface Design: An Introduction to GUI Design Principles and Techniques*, Wiley Computer Pub, 2002.

[16] J. Nielsen, *Usability Engineering*, Morgan Kaufmann, Boston, Mass, USA, 1993.

[17] J. Nielsen, "Usability inspection methods," in *Proceedings of the Conference Companion on Human Factors in Computing Systems*, pp. 413–414, 1994.

[18] J. Rieman, M. Franzke, and D. Redmiles, "Usability evaluation with the cognitive walkthrough," in *Proceedings of the Conference on Human Factors in Computing Systems*, pp. 387–388, Denver, Colo, USA, May 1995.

[19] C. O'Malley, S. Draper, and M. Riley, "Constructive interaction: a method for studying user-computer-user interaction," in *Proceedings of the Conference on Human-Computer Interaction*, pp. 1–5, 1984.

[20] A. Solano, T. Granollers, C. A. Collazos, and J. L. Arciniegas, "Experiencias en la especificación colaborativa de métodos de evaluación de usabilidad," in *Proceedings of the 14th Congreso Internacional de Interacción Persona-Ordenador (Interacción '13)*, Madrid, Spain, September 2013.

[21] A. Solano, T. Granollers, C. A. Collazos, and C. Rusu, "Proposing formal notation for modeling collaborative processes extending HAMSTERS notation," in *Proceedings of the World Conference on Information Systems and Technologies (WorldCIST '14)*, pp. 257–266, Madeira, Portugal, 2014.

[22] M. Mehta, I. Ajmera, and R. Jondhale, *Mobile Cloud Computing*, International journal of Electronics, 2013.

[23] A. Solano, C. Rusu, C. A. Collazos, and J. Arciniegas, "Evaluating interactive digital television applications through usability heuristics," *Ingeniare. Revista chilena de ingeniería*, vol. 21, no. 1, pp. 16–29, 2013.

[24] D. Quiñones, C. Rusu, and S. Roncagliolo, "Redefining usability heuristics for transactional web applications," in *Proceedings of the 11th International Conference on Information Technology: New Generations (ITNG '14)*, pp. 260–265, IEEE, April 2014.

[25] R. Inostroza, C. Rusu, S. Roncagliolo, C. Jiménez, and V. Rusu, "Usability heuristics for touchscreen-based mobile devices," in *Proceedings of the 9th International Conference on Information Technology (ITNG '12)*, pp. 662–667, April 2012.

[26] T. Tullis and B. Albert, *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*, Morgan Kaufmann, Boston, Mass, USA, 2nd edition, 2013.

[27] M. Piattini, F. García, J. Garzás, and M. Genero, "Medición y estimación del software: técnicas y métodos para mejorar la calidad y productividad del software," RA-MA Editorial, pp. 121–127, 2008.

[28] M. Hertzum, R. Molich, and N. E. Jacobsen, "What you get is what you see: revisiting the evaluator effect in usability tests," *Behaviour & Information Technology*, vol. 33, no. 2, pp. 144–162, 2014.

[29] L. Masip, T. Granollers, and M. Oliva, "A heuristic evaluation experiment to validate the new set of usability heuristics," in *Proceedings of the 8th International Conference on Information Technology: New Generations (ITNG '11)*, pp. 429–434, April 2011.

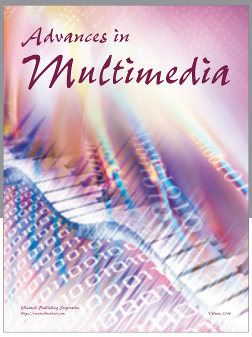
[30] M. González, L. Masip, A. Granollers, and M. Oliva, "Quantitative analysis in a heuristic evaluation experiment," *Advances in Engineering Software*, vol. 40, no. 12, pp. 1271–1278, 2009.

[31] L. Masip, M. Oliva, and T. Granollers, "Classification of interactive system components enables planning heuristic evaluation easier," in *Design, User Experience, and Usability. Theory, Methods, Tools and Practice*, pp. 478–486, Springr, Berlin, Germany, 2011.

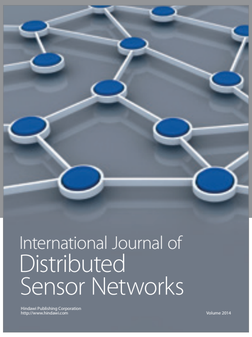
[32] C. Rusu, S. Roncagliolo, V. Rusu, and C. Collazos, "A methodology to establish usability heuristics," in *Proceedings of the 4th International Conference on Advances in Computer-Human Interactions (ACHI '11)*, Gosier, France, 2011.

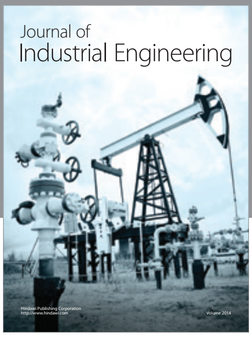
[33] C. Rusu, R. Muñoz, S. Roncagliolo, S. Rudloff, V. Rusu, and A. Figueroa, "Usability heuristics for virtual worlds," in *Proceedings of the 3rd International Conference on Advances in Future Internet (AFIN '11)*, pp. 16–19, Nice, France, August 2011.

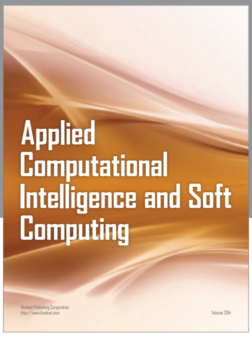
[34] C. Rusu, S. Roncagliolo, G. Tapia, D. Hayvar, V. Rusu, and D. Gorgan, "Usability heuristics for grid computing applications," in *Proceedings of the 4th International Conference on Advances in Computer-Human Interactions (ACHI '11)*, pp. 53–58, February 2011.

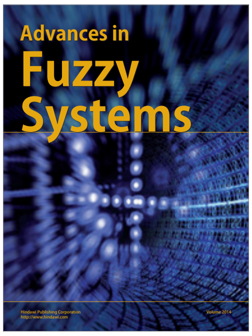

Advances in
Multimedia

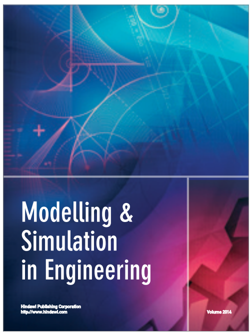
The Scientific
World Journal

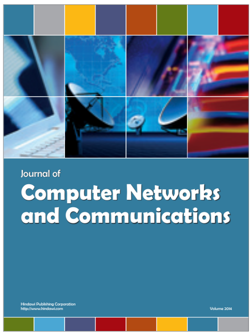

International Journal of
Distributed
Sensor Networks

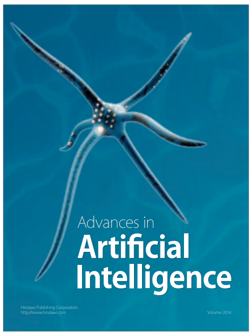

Journal of
Industrial Engineering

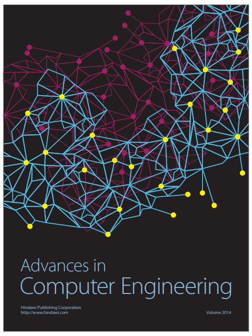

Applied
Computational
Intelligence and Soft
Computing


Advances in
Fuzzy
Systems

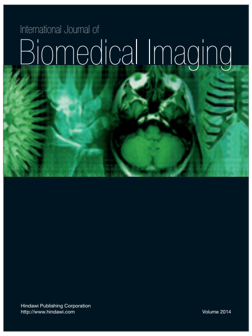

Modelling &
Simulation
in Engineering

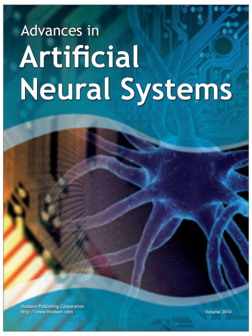

Journal of
Computer Networks
and Communications

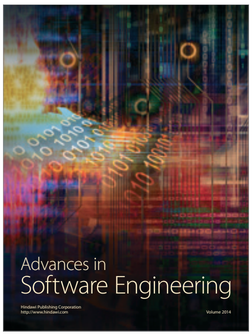

Advances in
Artificial
Intelligence


Advances in
Computer Engineering

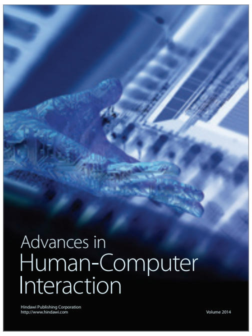
International Journal of
Computer Games
Technology

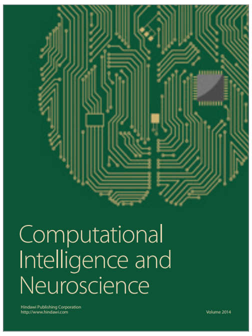

International Journal of
Biomedical Imaging


Advances in
Artificial
Neural Systems


Advances in
Software Engineering

Journal of
Robotics


Advances in
Human-Computer
Interaction


Computational
Intelligence and
Neuroscience

International Journal of
Reconfigurable
Computing

Journal of
Electrical and Computer
Engineering