Generating a Domain Specific Inspection Evaluation Method through an Adaptive Framework: A Comparative Study on Educational Websites

Roobaea S. AlRoobaea *r.alrobaea@uea.ac.uk*
Faculty of Computer Science and Information Technology Taif University, Saudi Arabia and School of Computing Sciences, UEA, Norwich, UK

Ali H. Al-Badi *aalbadi@squ.edu.om*
Department of Information Systems, Sultan Qaboos University, Oman

Pam J. Mayhew *p.mayhew@uea.ac.uk*
School of Computing Sciences, UEA, Norwich, UK

Abstract

The growth of the Internet and related technologies has enabled the development of a new breed of dynamic websites and applications that are growing rapidly in use and that have had a great impact on many businesses. These websites need to be continuously evaluated and monitored to measure their efficiency and effectiveness, to assess user satisfaction, and ultimately to improve their quality. Nearly all the studies have used Heuristic Evaluation (HE) and User Testing (UT) methodologies, which have become the accepted methods for the usability evaluation of User Interface Design (UID); however, the former is general, and unlikely to encompass all usability attributes for all website domains. The latter is expensive, time consuming and misses consistency problems. To address this need, new evaluation method is developed using traditional evaluations (HE and UT) in novel ways.

The lack of a methodological framework that can be used to generate a domain-specific evaluation method, which can then be used to improve the usability assessment process for a product in any chosen domain, represents a missing area in usability testing. This paper proposes an adapting framework and evaluates it by generating an evaluation method for assessing and improving the usability of a product, called Domain Specific Inspection (DSI), and then analysing it empirically by applying it on the educational domain. Our experiments show that the adaptive framework is able to build a formative and summative evaluation method that provides optimal results with regard to the identification of comprehensive usability problem areas and relevant usability evaluation method (UEM) metrics, with minimum input in terms of the cost and time usually spent on employing UEMs.

Keywords: Heuristic Evaluation (HE), User Testing (UT), Domain Specific Inspection (DSI), Adaptive Framework, Educational Domain.

1. INTRODUCTION

The growth of the Internet and new technologies has created new dynamic websites that are growing rapidly in use and are having a significant impact on many businesses. These dynamic websites are becoming increasingly developed in the midst of the Internet revolution and ever-improving information technologies. For example, e-learning websites are now essential for all universities that have a physical workplace. They have websites, which have become an integrated part of their business, particular in their e-learning systems. Nowadays, the Internet revolution has led to a large number of universities being solely online, without needing a physical workplace. To keep pace with this development, some companies and organizations seek to build free online learning websites that are oriented to world-class education for all educational levels, such Intel® Education and the BBC. This development in lifelong learning has made learners' intention to continue using e-learning an increasingly critical issue. Consequently, quality is considered crucial to education in general, and to e-learning in particular. Web design is a key factor in determining the success of e-learning websites, and users should be the priority in the designers' eyes because usability problems in educational websites can have serious ramifications, over and above the users failing to meet their needs.

It is clear that Heuristic Evaluation (HE) and User Testing (UT) are the most important traditional usability evaluation methods for ensuring system quality and usability [10;13]. Currently, complex computer systems, mobile devices and their applications have made usability evaluation methods even more important; however, usability differs from one product to another depending on product characteristics. It is clear that users have become the most important factor impacting on the success of a product; if a product is produced and is then deemed not useful by the end-users, it is a failed product (nobody can use it and the company cannot make money) [42]. Nayebi et al., (2012) asserted, "companies are endeavouring to understand both user and product, by investigating the interactions between them" [44].

Traditional usability measures of effectiveness, efficiency and satisfaction are not adequate for the new contexts of use [52]. HE has been claimed to be too general and too vague for evaluating new products and domains with different goals; HE can produce a large number of false positives, and it is unlikely to encompass all the usability attributes of user experience and design in modern interactive systems [21; 11]. UT has been claimed to be costly, time consuming, prone to missing consistency problems and subject to environmental factors [44]. Several studies have also emphasised the importance of developing UEMs as a matter of priority, in order to increase their effectiveness. To address these challenges, many frameworks and models have been published to update usability evaluation methods (UEMs) [3; 20]; however, these frameworks and models are not applicable to all domains because they were developed to deal with certain aspects of usability in certain areas [16].

The main objective of this paper is to address these challenges and to construct a methodological framework and then to test its validity by applying it on the educational domain, through three case studies. Furthermore, it is to conduct a comprehensive comparison between UT, HE and our domain specific investigation (DSI) method, which is from the adaptive framework, in terms of discovering the number of real usability problems and their severity in each of the usability problem areas, UEM metrics, and other measurements. The paper is organized in the following way. Section 2 starts with a brief literature to this study and includes a definition of usability problems, a severity rating and related work. Section 3 describes the construction of the adaptive framework. Section 4 is details the research methodology. Section 5 details the set of measurements and analysis metrics. Section 6 validates the adaptive framework by applying the new method (DSI), HE and UT in practice to three cases and provides an analysis and discussion of the results. Section 7 presents a discussion of the findings. Section 8 presents the conclusion and future work.

1.1 Research Hypotheses

This research hypothesizes that:

1. There are significant differences between the results of HE and DSI, where the latter method outperforms the former in terms of achieving higher ratings from evaluators on the issues relating to the number of usability problems, the usability problem areas, the UEM performance metrics, and the evaluators' confidence, concluding that it is not essential to conduct HE in conjunction with DSI.

2. There are significant differences between results of UT and DSI, where the latter method outperforms the former in terms of achieving higher ratings on the issues relating to the number of usability problems, the usability problems areas, the UEM performance metrics, concluding that it is not essential to conduct UT in conjunction with DSI.

2. LITERATURE REVIEW AND RELATED RESEARCH

A website is a product, and the quality of a product takes a significant amount of time and effort to develop. A high-quality product is one that provides all the main functions in a clear format, and that offers good accessibility and a simple layout to avoid users spending more time learning how to use it; these are the fundamentals of the 'usability' of a product. Poor product usability may have a negative impact on various aspects of the organization, and may not allow users to achieve their goals efficiently, effectively and with a sufficient degree of satisfaction [26]. The growth of the Internet has led to an explosion of educational website content, rising in accordance with demand. E-learning occurs when students in any place and time access the Internet to proceed through the sequence of teaching, completing the learning activities and achieving learning results and objectives. This could be part of a winning strategy for particular needs, such as decongestion of overcrowded education facilities, and support for students or teachers and adult education [1; 7]. However, some of these websites are difficult to use due to the inexperience of many of the designers and the lack of effective, efficient and accurate appropriate guidelines for performing this task. Consequently, users spend more time learning how to use the website than learning the educational content, causing frustration leading to abandonment of the site. Alkhattabi et al. 2010 state, "quality is considered a crucial issue for education in general, and for e-learning in particular" [4]. Thus there is a need for e-learning websites to be of sufficiently high quality. In this, it is extremely important to classify suitable criteria for addressing and assessing their quality [48].

The reviewed literature shows that the techniques for measuring the quality of user experience have been classified under the heading of ergonomics and ease-of-use, but more lately under the heading of usability [44]. This aims to ensure that the user interface is of sufficiently high quality. Usability is one of the most significant aspects affecting the quality of a website and its user experience. Nielsen (1994b) stated, "usability is associated with learnability, efficiency, memorability, errors and satisfaction" [38]. Muir et al. (2003) defined pedagogic usability as a branch of usability that "affects educational website design and development, particularly in the context of supported open and distance learning" [35]. Usability is not a single 'one-dimensional' property of a user interface. There are many usability attributes that should be taken into account and measured. Shackel and Richardson (1991) proposed four-dimensional attributes that influence the acceptance of a product, which are effectiveness, learnability, flexibility and attitude [46]. Nielsen (1994b) introduced five major attributes of usability based on a System Acceptability model [38], and they are as follows; 1) Easy to learn: a system should be easy to learn for the first time; 2) Efficient to use: the relationship between accuracy and time spent to perform a task; 3) Easy to remember: a user should be able to use the system after a period without spending time learning it again; 4) Few errors: the system should prevent users from making errors (this also addresses how easy it is to recover from errors); and 5) Subjectively pleasing: this addresses the user's feeling towards the system.

Usability evaluation methods (UEMs) are a set of techniques that are used to measure usability attributes. They can be divided into three categories: inspection, testing and inquiry. Heuristic Evaluation (HE) is one category of the inspection methods. It was developed by [34], and is guided by a set of general usability principles or 'heuristics' as shown Table 1. It can be defined as a process that requires a specific number of experts to use the heuristics in order to find usability problems in an interface in a short time and with little effort [33]. It can be used early in the development process, and may be used throughout the development process [40]. However, it is a subjective assessment and depends on the evaluator's experience, and can produce a large number of false positives that are not usability problems at all or can miss some real problems [23; 39; 11; 21].

Heuristic Evaluation
Visibility of system status
Match between system and the real world
User control and freedom
Consistency and standards
Error prevention
Recognition rather than recall
Flexibility and efficiency of use
Aesthetic and minimalist design
Helps users recognize, diagnose, and recover from errors
Help and documentation

TABLE 1: Heuristic Evaluation.

There are two kinds of expert evaluators. One is a 'single' evaluator, who can be defined as a person with general usability experience. The second is a 'double' evaluator who can be defined as a person with a usability background in a specific application area. Molich and Nielsen (1990) recommended from previous work on heuristic evaluation that between three and five single expert evaluators are necessary to find a reasonably high proportion of the usability problems (between 74% and 87%) [34]. For the double expert evaluators, it is sufficient to use between two and three evaluators to find most problems (between 81% and 90%). There is no specific procedure for performing HE. However, Nielsen [37] suggested a model procedure with four steps. Firstly, conducting a pre-evaluation coordination session (a.k.a training session) is very important. Before the expert evaluators evaluate the targeted website, they should take few minutes browsing the site to familiarize themselves with it. Also, they should take note of the actual time taken for familiarisation. If the domain is not familiar to the evaluators, the training session provides a good opportunity to present the domain. Also, it is recommended that in the training session, the evaluators evaluate a website using the heuristics in order to make sure that the principles are appropriate [14]. Secondly, conducting the actual evaluation, in which each evaluator is expected to take around 1 to 1.5 hours listing all the usability problems. However, the actual time taken for evaluation should always be noted. Next, there should be a debriefing session, which would be conducted primarily in a brainstorming mode and would focus on discussion of possible redesigns to address the major usability problems and general problematic aspects of the design. A debriefing is also a good opportunity for discussing the positive aspects of the design, since heuristic evaluation does not otherwise address this important issue. Finally, the results of the evaluations are collected into actual evaluation tables, and then combined into a single table after removing any redundant data. After the problems are combined, the evaluators should agree on the severity of each individual problem [37].

In the present context and in relation to HE, usability testing (also known as user testing), is another important evaluation method for ensuring system quality, in particular for websites. It needs participants to perform a set of tasks, usually in a laboratory. These tasks are performed without information or clues as to how to complete them, and with no help provided to the user during the test session. Also, the completion of these tasks is monitored and assessed by an

observer who records the usability problems encountered by the users. All the observed data, such as error numbers, time spent, success rate and user satisfaction, need to be recorded for analysis [38]. Dumas and Redish (1991) stressed that a fruitful usability testing session needs careful planning and attention to detail [17]. Accordingly, there is a general procedure for conducting user testing, thus: 1) Planning a usability test; 2) Selecting a representative sample and recruiting participants; 3) Preparing the test materials and actual test environment; 4) Conducting the usability test; 5) Debriefing the participants; 6) Analysing the data of the usability test; and 7) Reporting the results and making recommendations to improve the design and effectiveness of the system or product. The Think-Aloud technique (TA) is used with UT. There are three TA types, which are concurrent, retrospective and constructive interaction. The concurrent TA type is the most common; this involves participants verbalising their thoughts whilst performing tasks in order to evaluate an artefact. Retrospective TA is less frequently used; in this method, participants perform their tasks silently, and afterwards comment on their work on the basis of a recording of their performance. Constructive interaction is more commonly known as Co-Discovery Learning, where two participants work together in performing their tasks, verbalising their thoughts through interacting [51].

One important factor in usability testing is setting the tasks. Many researchers are aware that task design is an important factor in the design of adequate usability tests. The tasks designed for Web usability testing should be focused on the main functions of the system. The tasks should cover the following aspects: 1) Product page; 2) Category page; 3) Display of records; 4) Searching features; 5) Interactivity and participation features; and 6) Sorting and refining features [50]. Dumas and Redish (1999) suggested that the tasks could be selected from four different perspectives [17]. These are: 1) Tasks that are expected to detect usability problems; 2) Tasks that are based on the developer's experience; 3) Tasks that are designed for specific criteria; and 4) Tasks that are normally performed on the system. They also recommended that the tasks be short and clear, in the users' language, and based on the system's goals [17]. Alshamari and Mayhew (2008) found that task design can play a vital role in usability testing results, where it was shown that changing the design of the task can cause differences in the results [6].

The result of applying HE and UT is a list of usability problems [37]. These problems are classified into different groups to which a numeric scale is used to measure the severity of each problem. Firstly, this issue is not a usability problem at all. Secondly, this is a cosmetic problem that does not need to be fixed unless extra time is available on the project. Next, this issue is a minor usability problem; fixing this should be given low priority. Then, this is a major usability problem; it is important to fix this, so it should be given high priority. Finally, this issue is a usability catastrophe; it is imperative to fix this before the product can be released.

In the early years of computing, HE was widely applied in measuring the usability of Web interfaces and systems because it was the only such tool available. These heuristics have been revised for universal and commercial websites as HOMERUN heuristics [41]. Furthermore, [11;12] have proposed UEMs called HE-Plus and HE++, which are extensions to HE by adding what is called a "usability problem profile". However, some researchers have found that their tested websites failed in certain respects according to these extended or modified heuristics [49; 5]. On the other hand, many researchers then sought to compare and contrast the efficiency of HE with other methods such as UT. They found that HE discovered approximately three times more problems than UT. However, they reported that more severe problems were discovered through UT, compared with HE [30; 18; 27]. Lately, researchers' findings have been almost unanimous in one respect: HE is not readily applicable to many new domains with different goals and are too vague for evaluating new products such as web products because they were designed originally to evaluate screen-based products; they were also developed several years before the web was involved in user interface design [24; 25; 31]. Thus, each method seems to overcome the other method's limitations, and researchers now recommend conducting UT together with HE because each one is complementary to the other, and then combining the two methods to offer a better picture of a targeted website's level of usability [36; 32]. To address these challenges, many frameworks and models have been published to update usability

evaluation methods (UEMs) [3]; [20]; however, these frameworks and models are not applicable to all domains because they were developed to deal with certain aspects of usability in certain areas [16].

It can be seen from the above that there is need to an effective and appropriate methodology for evaluating the emerging domains/technology to measure their levels of efficiency, effectiveness and satisfaction, and ultimately to improve their quality. Also, there is need for a method that is context of use and that considers expert and user perspectives. This finding and the criticality of website usability has encouraged researchers to formulate such a framework. This framework should be applicable across numerous domains. In other words, it should be readily capable of adapting in any domain and for any technology. This paper constructs this framework for generating a context- specific method for the chosen domain that can be applied without needing to conduct user testing. However, developing and testing a method is not quick and it should involve some key stages. The next section describes the steps employed in the adaptive framework, also, it describes the process used to test it.

3. CONSTRUCTION OF THE ADAPTIVE FRAMEWORK

The adaptive framework was developed according to an established methodology in HCI research. It consists of two distinct phases: 1) Development phases that consist of four main steps for gathering together suitable ingredients to develop a context-specific method (DSI) for website evaluation; and 2) Validation phases for testing the developed DSI method practically (these are outlined in Figures 1 and 2). Below is an explanation of the four development steps:

Development Step One (D1: Familiarization): This stage starts from the desire to develop a method that is context of use, productive, useful, usable, reliable and valid, and that can be used to evaluate an interface design in the chosen domain. It entails reviewing all the published material in the area of UEMs but with a specific focus on knowledge of the chosen domain. Also, it seeks to identify an approach that would support developers and designers in thinking about their design from the intended end-users' perspective.

Development Step Two (D2: User Input): This stage consists of mini-user testing (task scenarios, think aloud protocol and questionnaire). Users are asked to perform a set of tasks on a typical domain website and then asked to fill out a questionnaire. The broad aim of this is to elicit feedback on a typical system from real users in order to appreciate the user perspective, to identify requirements and expectations and to learn from their errors. Understanding user needs has long been a key part of user design, and so this step directly benefits from including the advantages of user testing.

Development Step Three (D3: Expert Input): This stage aims to consider what resources are available for addressing the need. These resources, such as issues arising from the mini-user testing results and the literature review, require a discussion amongst experts (in the domain and/or usability) in order to obtain a broader understanding of the specifics of the prospective domain. Also, it entails garnering more information through conversations with expert evaluators to identify the areas/classification schemes of the usability problems related to the selected domain from the overall results. These areas provide designers and developers with insight into how interfaces can be designed to be effective, efficient and satisfying; they also support more uniform problem description and they can guide expert evaluators in finding real usability problems, thereby facilitating the evaluation process by judging each area and page in the target system.

Development Step Four (D4: Draw Up DSI Method: data analysis): The aim of this step is to analyse all the data gathered from the previous three. Then, the DSI method will be established (as guidelines or principles) in order to address each area of the selected domain.

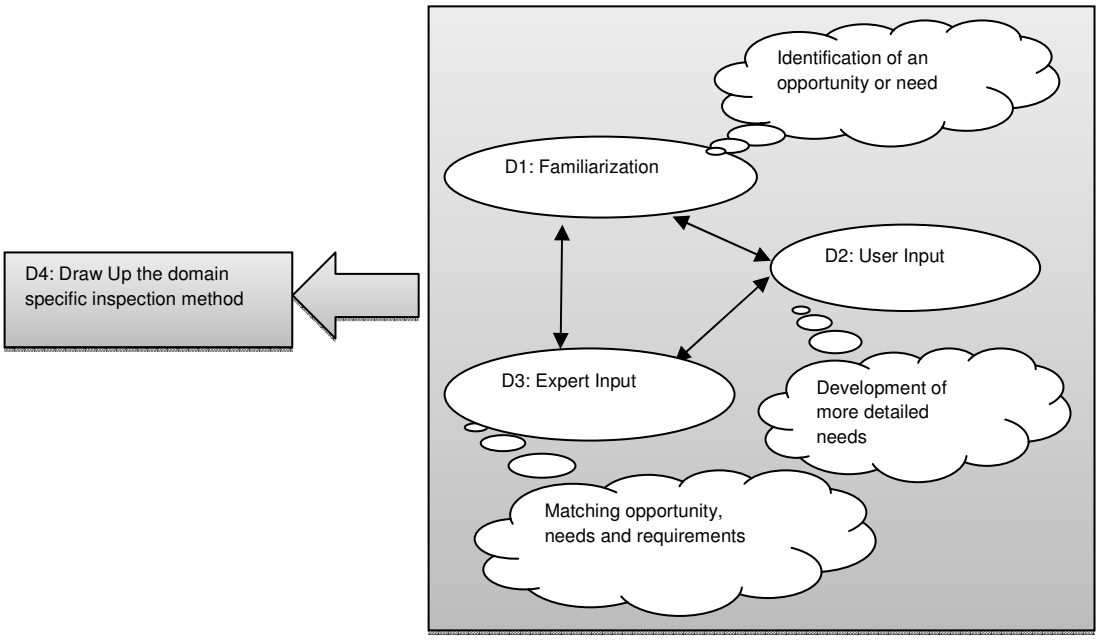

FIGURE 1: Development Stages of the Adaptive Framework.

After constructing the DSI framework, the researchers test it intensively through rigorous validation methods to verify the extent to which it achieves the identified goals, needs and requirements that the method was originally developed to address (this validation is outlined in Figure 2).

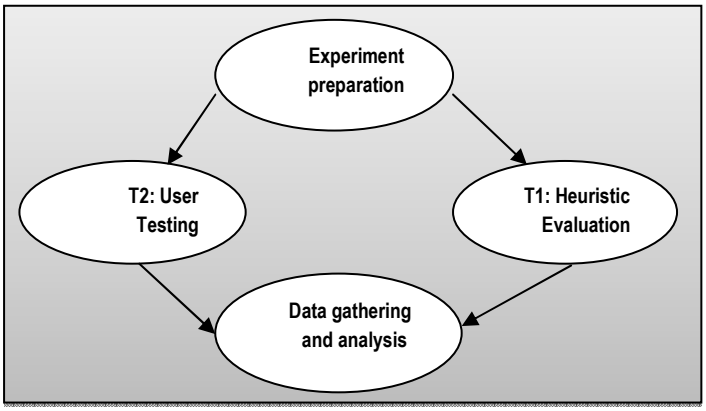

FIGURE 2: Testing stages of the adaptive framework

The validation phase of the adaptive framework again consists of four separate steps, as explained below:

1. Experiment Preparation (for DSI, HE and UT): Before the actual evaluation formally starts, the following initial preparations are needed: 1) Select a number of systems/websites that are typical of the chosen domain; 2) Recruit expert evaluators and users; 3) Plan the sequence of conducting the evaluations by each group in such a way as to avoid any bias; and 4) Prepare the experiment documents. The initial experiment preparation phase is concluded with a pilot experiment to make sure that everything is in place and ready for the actual evaluation.

2. Heuristic Validation (Expert Evaluation (HE)): The aim of this step is the validation of the newly developed method by conducting a heuristic evaluation (HE). Expert evaluators need a familiarization session before the actual evaluation. The actual expert evaluation is then conducted using the newly developed DSI method alongside HE. The aim of this process is to collect data ready for analysis (analytically and statistically), as explained in step 4.

3. Testing Validation: (User Evaluation (UT)): The aim of this step is to complement the results obtained from the expert evaluation, by carrying out usability lab testing on the same websites. [36] recommends conducting usability testing (UT) with HE because each one is complementary to the other. Then, the performance of HE is compared with the lab testing to identify which problems have been identified by UT and not identified by HE and/or DSI, and vice versa. The aim of this process is to collect data ready for analysis (empirically and statistically) in step 4.

4. Data Analysis: This step aims to analyse all the results and to answer all the questions raised from the above steps in a statistical manner. It is conducted in two parts; one focused on HE and the other on UT. The researchers extract the problems discovered by the experts from the checklists of both DSI and HE. Then, they conduct a debriefing session with the same expert evaluators to agree on the discovered problems and their severity, and to remove any duplicate problems, false positives or subjective problems. Then, the problems approved upon are merged into a master problem list, and any problems upon which the evaluators disagree are removed. Ultimately, the researchers conduct a comparison on the results of both methods (DSI and HE) in terms of the number of problems discovered (unique and overlapping), their severity ratings, which problems are discovered by HE and not discovered by DSI and vice versa, the areas of the discovered problems, the UEM performance metrics, evaluator reliability and performance, and the relative costs entailed in employing the two methods.

In the second part, the researchers conduct a debriefing session with independent evaluators to rank the severity of the problems derived from the user testing and to remove any duplicate problems. Following this, they establish the list of usability problems for UT. Subsequently, a single unique master list of usability problems is consolidated from the three methods. A comparison of the results of the three methods is then conducted in terms of the number of problems discovered (unique and overlapping), their severity ratings, and the areas of the discovered problems; this is to identify which problems were discovered by HE and DSI and not discovered by UT, and vice versa. Also, the UEM performance metrics of each method are measured, together with other measures, such as their relative costs and reliability. Moreover, this final step seeks to prove or refute the efficacy of conducting UT and HE with DSI.

Having proposed the framework above, it was decided to evaluate its practicality by applying it to a real-life experiment. From the literature review, it was found that the evaluation of the free educational websites domain is a subject area that has not yet been fully explored, nor have any context-specific methods been generated for this domain (to overcome the shortcomings of HE and UT); this is an important area of research because these websites are now essential to many users and companies. A well-designed educational website (i.e. one that is aesthetically attractive and is easy to use) can positively affect the number of people who become members.

4. RESEARCH METHODOLOGY

The experimental approach was selected to address the research hypotheses outlined above. Essentially, this section describes the methodology employed in this comparative study. Before conducting this experiment, a set of procedures were followed by the researchers, as follows:

4.1 Design

This experiment employs the between-subject and within-subject designs. The independent variables are the three methods (HE, DSI and UT). The dependent variables are the UEM

performance metrics, which are calculated from the usability problems reported by the evaluators/users, and from the reliability and efficiency measurements.

4.2 Evaluation of the Practicality of the Framework

In the first stage, the researchers conducted a literature review on the materials relating to usability and UEMs as well as on the requirements of the educational websites domain. In stage two, a mini-user testing session was conducted through a brief questionnaire that entailed four tasks, which were sent to ten users who are regular educational website users. In stage three, a focus group discussion session was conducted with eight experts in usability and the educational domain (i.e. single and double experts). Cohen's kappa coefficient was used on the same group twice to enable a calculation of the reliability quotient for identifying usability problem areas. In stage four, the researchers analysed the results of the three stages and incorporated the findings. The intra-observer test-retest using Cohen's kappa yielded a reliability value of 0.8, representing satisfactory agreement between the two rounds. After that, the usability problems areas were identified to facilitate the process of evaluation and analysis, and to help designers and programmers to identify the areas in their website that need improvement. Then, the DSI method was established, closely focused on educational websites, taking into an account what is called "learner-centred design". The method was created and classified according to the usability problem areas detailed in Table 2 below.

Usability problem area/attributes	Domain Specific Inspection (DSI)
User usability	Supports modification and progress of evaluation
	Supports user tasks and avoids difficult concepts
	Feedback and support services
	Easy to remember
Motivational factors	Supports leaner curiosity
	Learning content design and Attractive screen design
	Motivation to learn
Content information and process orientation	Relevant, correct and adequate information
	Reliability and Validity
	Privacy and Security
Learning process	Assessment
	Interactivity
	Evokes mental images for the learner
	Resources
	Learning management
	Learnability
Design and media usability	Multimedia representations
	Accessibility and compatibility of hardware devices
	Functionality
	Navigation and Visual clarity

TABLE 2: Final Version of Domain Specific Inspection (DSI).

4.3 Selection of the Targeted Websites

The first step in an initial preparation phase is selecting the websites. The researchers sought to ensure that the selected websites would support the research goals and objectives. The selection process was criteria-based; six aspects were determined and verified for each website, and these are: 1) Good interface design, 2) Rich functionality, 3) Good representatives of the free educational websites, 4) Not familiar to the users, 5) No change will occur before and during the actual evaluation, and 6) Completely free educational websites. In order to achieve a high level of quality in this research, the researchers chose three well-known websites in this domain. The first website was 'skoool'. It is an Intel driven initiative that delivers highly innovative and interactive learning resources via cutting-edge technologies and devices [47]. The second website was

AcademicEarth. It is an organization founded with the aim of giving everyone on earth access to a world-class education [2]. The third website was BBC KS3bitesize. This website helps school students from 11 to 14 with their coursework, homework and test preparation [8]. All of these have all the aspects mentioned above.

4.4 Recruitment of Experts and Users

The selection of usability experts and users is the second important step in the initial preparation phase in this experiment. The researchers decided to recruit eight expert evaluators, divided into two groups of four, who were carefully balanced in terms of experience. In each group, there are two double expert evaluators (usability specialists in educational websites) and two single expert evaluators (usability specialists in general). Each evaluator was to conduct his/her evaluation separately in order to ensure independent and unbiased evaluations [37]. Also, each group employed two methods, namely DSI and HE, to evaluate the three different websites. The evaluation was carried out in a prescribed sequence, i.e. one group used HE, DSI and HE on Skoool, AcademicEarth and BBC KS3bitesize, while the second group used DSI, HE and DSI as shown in Figure 3 below. The researchers adopted this technique to avoid any bias in the results and also to avoid the risk of any expert reproducing his/her results in the second session through over-familiarity with one method, i.e. each evaluation was conducted with a fresh frame of mind.

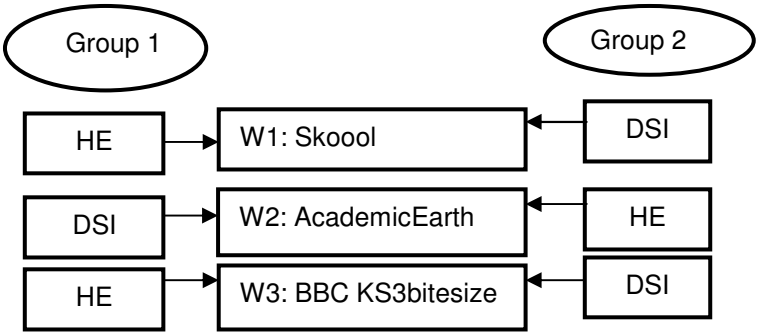

FIGURE 3: Usage of Methods by Group.

Selecting and recruiting users must be done carefully; the participants must reflect the real users of the targeted website because inappropriate users will lead to incorrect results, thereby invalidating the test. Appropriate users will deliver results that are more reliable; they will also be motivated to conduct the experiment [17]. There is no agreement on how many users should be involved in usability testing. Dumas and Redish (1999) suggested that 6 to 12 users are sufficient for testing, whereas other studies have recommended that 7, 15 and 20 users are the optimal numbers for evaluating small or large websites; particularly 20 users if benchmarking is needed [39]. At this point, 60 users were engaged; they were chosen carefully to reflect the real users of the targeted websites and were divided into three groups for each website, i.e. a total of 20 users for each website. The majority of the users are students, and they were mixed across the three users groups in terms of gender, age, education level and computer skills.

4.5 Building the Experiment Documents

In experiment preparation phase, the third step was building the set of preparation documents for HE, DSI and UT, such as an introduction sheet, heuristics checklists, tasks sheet, ranking problems sheet, observer sheet, demographics, satisfaction and Likert questionnaires, problems sheet and a master problem list. The introduction sheet contains the goals and objectives of the evaluation and the roles of users and experts. Before starting actual evaluation, users and experts completed a demographics questionnaire to obtain more information about them. Expert evaluators used checklists that had been developed by the researchers to facilitate the evaluation process for DSI and HE. Users used the task sheet that was designed according to the main functions that users would normally expect to perform on the three educational websites. A combination of task designs and TA approaches (as mentioned in the literature review) were used in this experiment. There were four sub-tasks in each three task group, they were kept to a

reasonable time limit and they were interesting and engaging enough to hold the users' attention. As briefly mentioned above, usability testing requires an observer, and the researcher adopted this role in all the sessions, noting all the comments made by the users. The researcher used a stopwatch to record the time spent by each user on each task, and an observation sheet to write down the behaviour of each user and the number of problems encountered. The ranking sheet aims to help the expert evaluators and independent evaluators (for user testing) to rank the severity of usability problems by using Nielsen's scale as mentioned above. After the evaluators had finished their evaluations and had ranked HE and DSI problems, they were asked to complete a satisfaction questionnaire using the System Usability Scale (SUS) on both methods. It is made up of ten items in the form of scale questions ranging from 0 to 100 to measure the satisfaction of expert evaluators [9]. Also, when the users finished their tasks, they were asked to rate their level of satisfaction in a questionnaire on a scale of one to seven, where one refers to 'highly unsatisfactory' and seven indicates 'highly satisfactory'. This scale has been suggested to truthfully measure the levels of satisfaction that are felt by users on a website interface following a test [39]. Evaluators and users were asked to fill in an open-end questionnaire by writing down their comments and feedback on the methods used and explaining any reaction that was observed during the test. Subsequently, the Likert scale was used by the evaluators for measuring either positive or negative responses to a statement in both the DSI and HE methods. Moreover, the researchers extracted the problems of three methods from the problems sheet and removed all false positive ('not real') problems, evaluators' 'subjective' problems and duplicated problems during the debriefing session. The problems agreed upon were merged into a unique master problem list and any problems upon which the evaluators disagreed were removed.

4.6 Piloting the Experiment

The forth step in an initial preparation phase is a pilot experiment. It was conducted by two independent evaluators and fifteen users. All materials were checked to make sure that there were no spelling or grammatical errors and no ambiguous words or phrases, and that all of sentences in the instruments (heuristics, check-lists, task scenarios, questionnaires and procedures) were sufficiently clear to be used by the evaluators. A fewer minor improvements were made, such as before establishing the final version of DSI, the independent evaluators suggested removing game heuristics from Design and Media Usability area because approximately 95% of games websites are not educational websites as confirmed by [22].

Furthermore, to assess the time needed for testing, the fifteen users were divided into three groups (five users in each). Each group performed its tasks. The users' behaviour was monitored, and all the usability measures were assessed as they would be in real testing. All of these steps resulted in useful corrections and adjustments for the real test. Also, the test environment was a quiet room. We attempted to identify what equipment the users regularly use and set it up for them before the test, for example, using the same type of machine and browser.

4.7 Actual Evaluation

The Heuristics Validation phase started with a training (familiarization) session for the eight expert evaluators. They were given a UEM training pack that contained exactly the same information for both groups. The researchers emphasized to each evaluator group that they should apply a lower threshold before reporting a problem in order to avoid misses in identifying real problems in the system. Then, the actual expert evaluation was conducted and the evaluators evaluated all websites consecutively, rating all the problems they found in a limited time (which was 90 minutes). After that, they were asked to submit their evaluation report and to complete the five-point scale in the SUS questionnaire (1 for strongly disagree and 5 for strongly agree) to rate their satisfaction on the evaluation method they had used (DSI or HE), and to give feedback on their own evaluation results.

The Testing Validation phase started with a training (familiarization) session for the 60 users; it involved a quick introduction on the task designs, the TA approach and the purpose of the study. The next step entailed explaining the environment and equipment, followed by a quick demonstration on how to 'think aloud' while performing the given tasks. Prior to the tests, the

users were asked to read and sign the consent letter, and to fill out a demographic data form that included details such as level of computer skill. All the above steps took approximately ten minutes for each test session. The actual test started from this point, i.e. when the user was given the task scenario sheet and asked to read and then perform one task at a time. Once they had finished the session, they were asked to rate their satisfaction score relating to the tested website, to write down their comments and thoughts, and to explain any reaction that had been observed during the test, all in a feedback questionnaire. This was followed by a brief discussion session.

5. DATA ANALYSIS AND MEASUREMENTS

To determine whether our adaptive framework has generated an evaluation method of sufficiently high quality, the results of the comparison process between the three methods needed a meta-analysis to be performed, as follows:

1. Compare the average time spent by each group when using each method during the evaluation sessions.

2. Compare the results of the usability problems and their severity in order to assess the performance of each method in terms of identifying unique and overlapping problems and of identifying real usability problems in the usability problem areas.

3. Comparing the satisfaction scores of HE and DSI by using System Usability Scale (SUS).

4. Reliability of HE and DSI: This can be measured from employing the 'evaluators' effect formula' (Any-Two-Agreement). It is used on each single evaluators in order to measures the performance of the evaluators individually [21].

   Any-Two-Agreement = Average of $|Pi \cap Pj| / |Pi \cup Pj|$ over all $^1/_2$ n (n-1) pairs of evaluators, where Pi is the set of problem discovered by evaluator i and the other evaluator j, and n refers to the number of evaluators.

5. Evaluators' Performance: This can be measured by the performance of single and double expert evaluators in discovering usability problems by using HE and DSI in each group and website.

To make further comparisons between the performance of HE, DSI and UT in identifying usability problems, a set of UEM and other metrics were used for examining their performance; none of these metrics on their own addresses errors arising from false positive, subjective and missed problems. They are efficiency, thoroughness, validity, effectiveness, reliability and cost. Efficiency in UEMs is the "ratio between the numbers of usability problems detected to the total time spent on the inspection process" [19]. Thoroughness is perhaps the most attractive measure; it is defined as a measure indicating the proportion of real problems found when using a UEM to the total number of known real problems [29]. Validity is the extent to which a UEM accurately identifies usability problems [45]. Effectiveness is defined as the ability of a UEM to identify usability problems related to the user interface [28]. The reliability of user testing can be measured by the mean number of evaluators to the number of real problems identified [11]. The cost can be calculated by identifying the cost estimates. It can be done fairly simply by following Nielsen's equation who estimated the hourly loaded cost for professional staff at $100 [37]. All of them are computed as follows:

1. Efficiency = (No. of problems)/(Average time spent)

2. Thoroughness = (No. of real usability problems found)/(Total no. of real usability problems present)

3. Validity = (No. of real usability problems found)/(No. of issues identified as a usability problem)

4. Effectiveness = Thoroughness × Validity

5. Reliability of UT = (Mean no. of evaluators)/(No. of real problems identified)

6. Cost = (No. of evaluation hours) × (Estimate of the loaded hourly cost of participants)

To test the research hypotheses and choose the correct statistical test in SPSS, the normality of the data should be examined. The results of both Skewness and Kurtosis are equal to 0, and the Sig-value of the Kolmogorov-Smirnov test are greater than 0.05. As a result, the data are normally distributed, and t-Test, One-way ANOVA and Pearson correlation were chosen at 5% significance level as our methods for statistical analysis, as the dependent variables in our data are independent of each other, improving the validity of using analysis of variance. The Mann-Whitney test was used to analyse the Likert score (considered as an ordinal scale).

6. ANALYSIS AND DISCUSSIONS

This section describes the results obtained from using the three method adopted in this study. It starts by detailing the result of the HE and DSI methods separately, including quantitative and qualitative analyses. This is followed by detailing the result of the UT method alone, including quantitative and qualitative analyses. Ultimately, all the results derived from the three methods were compared in terms of the numbers of problems and types, as well as the other usability metrics as mentioned above.

6.1 Analysis for HE and DSI Results

6.1.1 Time spent: It can be seen from Tables 3 and 4 that the average time taken for doing the three experiments using DSI was 24.25 minutes with a standard deviation of 6.7, whereas the DSI average was 42.58 minutes with a standard deviation of 7.1. This difference in time spent is not significant ($F = 0.199$, $p = 0.660$) using the t-test. The group who used DSI managed to evaluate the website more quickly than the other group but discovered fewer usability problems. The group that used DSI spent almost double the time evaluating the website but discovered almost three times as many real usability problems. There was a statistically significant positive relationship between time spent and problems discovered by using the Pearson correlation test, where the Sig value is 0.020 at the 0.05 level. This result reveals that the users who spent more time were able to discover more usability problems.

Website	Skoool	AcademicEarth	BBCKS3bitesize
Evaluator	**Time**	**Time**	**Time**
1	25	45	24
2	30	50	40
3	25	55	22
4	20	29	23
Heuristics	HE	DSI	HE
# of problems	10	29	2
Mean time taken	25	45	27

TABLE 3: Average Time Taken and Number of Problems for Group 1.

Website	Skoool	AcademicEarth	BBCKS3bitesize
Evaluator	**Time**	**Time**	**Time**
1	42	30	50
2	40	17	38
3	38	15	35
4	45	20	44
Heuristics	DSI	HE	DSI
# of problems	33	13	12
Mean time taken	41	21	42

TABLE 4: Average Time Taken and Number of Problems for Group 2.

An explanation for the differences in time spent and number of problems located is gleaned from the evaluators' feedback. They said that HE was not particularly helpful, understandable or memorable for them. However, DSI helped them to develop their skills in discovering usability problems in this application area; also, this set was more understandable and memorable during their evaluation. To further analyse these factors of time spent and number of problems discovered, efficiency metrics were applied. DSI proved to be more efficient than HE in discovering usability problems (DSI = 1 vs. HE = 0.7), as Table 5 shows.

Method	Skoool	Academic Earth	BBC KS3bitesize	**Mean**
	Efficiency	**Efficiency**	**Efficiency**	
HE	0.77	1.1	0.2	0.7
DSI	1.55	1.2	0.6	1

TABLE 5: Mean Score of Efficiency for Two Methods.

6.1.2 Number of Usability Problems

Table 6 shows that HE was able to uncover 25% of the total number of real usability problems. However, DSI was able to uncover 75% of the total number of real usability problems in the websites (no false and subjective problems).

Website	Group	Expert and type	Method	# of problems found by each evaluator	# of problems with repetition	# of problems without repetition	# of problems with repetition between groups	% of problems found by each evaluator	% # of problems found by each group
Skoool	G 1	Ev. 1+	HE	8	19	10	43	19%	23%
		Ev. 2^	HE	5				12%	
		Ev. 3^	HE	1				2%	
		Ev. 4+	HE	5				12%	
	G 2	Ev. 1+	DSI	21	64	33		49%	77%
		Ev. 2^	DSI	15				35%	
		Ev. 3+	DSI	13				30%	
		Ev. 4^	DSI	15				35%	
Academic Earth	G 1	Ev. 1+	DSI	10	42	29	42	24%	69%
		Ev. 2^	DSI	11				26%	
		Ev. 3^	DSI	9				21%	
		Ev. 4+	DSI	12				29%	
	G 2	Ev. 1+	HE	6	23	13		14%	31%
		Ev. 2^	HE	6				14%	

		Ev. 3+	HE	7				17%	
		Ev. 4^	HE	4				10%	
BBC KS3bitesize	G 1	Ev. 1+	HE	1	5	2	14	7%	14%
		Ev. 2^	HE	1				7%	
		Ev. 3^	HE	1				7%	
		Ev. 4+	HE	2				14%	
	G 2	Ev. 1+	DSI	6	24	12		43%	86%
		Ev. 2^	DSI	6				43%	
		Ev. 3+	DSI	7				50%	
		Ev. 4^	DSI	5				36%	

Total number of usability problems discovered by each set of heuristics	**Heuristics**	**Total number**	**Approx. %**
	HE	25	25%
	DSI	74	75%

+ Double Expert ^ Single Expert Ev. = Evaluator

TABLE 6: Summary of numbers and percentages for usability problems uncovered on each website, by each group, each evaluator and each set of heuristics.

Table 6 also shows that DSI discovered more real problems in all three evaluations. For example, it discovered 33 problems in the Skoool website, which is equivalent 77% of the total problems on the website. However, HE discovered only 10 problems for the same website, which equates to 23% of the total problems on the website. Also, the results for the BBC KS3bitesize website were considerable. DSI identified 12 problems, which represent 86%; however, HE found only 2 problems, which equates to 14% of the total problems on this website. One striking result is that the number of problems identified by each evaluator who used HE was always less than the number of problems identified by any evaluator using DSI for the same website. An explanation of this was evident in the evaluator answers in the questionnaire. They said that the HE set was difficult to use, did not remind them of aspects they might have forgotten about, and they did not believe that this set encouraged them to be thorough in their evaluation. On the other hand, they said that the DSI set was easy to use; indeed, it helped them remember all the functions that needed to be tested, as it is specific and designed to cover all the aspects needed for educational websites. The t-test revealed that the difference in discovering usability problems by each method in each website is significant (see Table 7).

Website	**Group**	**Method**	***t*-value**	***df*-value**	***p*-value**
Skoool	Group 1	HE	-5.000	5.801	0.003
	Group 2	DSI			
Academic Earth	Group 1	DSI	-5.270	5.996	0.002
	Group 2	HE			
BBC KS3bitesize	Group 1	HE	-9.922	4.973	P < 0.001
	Group 2	DSI			

TABLE 7: Results of t-test between Groups and Methods in Each Website.

In terms of the performance of each set of heuristics in discovering unique and overlapping problems, Table 6 illustrates the total number of real problems discovered, which was 99 on the three websites, out of which 25 were identified using HE and 74 using DSI. All the duplicated problems were removed and compared by two independent evaluators, in order to identify the unique and overlapping problems. When problems from the two evaluation groups were consolidated, there were 19 duplicates; we thus identified a total of 80 real problems in all websites. The total for uniquely identified problems in all websites was 61; DSI identified 55 real problems (69% of the 80 problems) that were not identified by HE, and there were 6 real problems (8% out of 80) identified by HE that were not identified by DSI. 19 real problems (24%) out of 80 problems were discovered by both methods (as depicted in Figure 5).

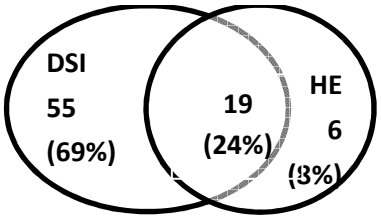

FIGURE 5: Overlapping Problems between Both Methods.

Table 8 shows the severity ratings for the real problems discovered (cosmetic, minor, major and catastrophic). Overall, a great many real usability problems were discovered. The most significant results were obtained from using DSI, while HE found fewer (or no) usability problems. The t-test reveals that there is significant difference between the two methods in terms of severity problems ($t = -1.877$, $df = 4$, $p = 0.026$).

Website	Severity of problems	Type of method			
		HE		DSI	
Skoool	Cosmetic	Group 1	3	Group 2	12
	Minor		3		11
	Major		2		6
	Catastrophic		2		4
	Severity (average)	2.3		2.1	
AcademicEarth	Cosmetic	Group 2	2	Group 1	11
	Minor		7		11
	Major		3		4
	Catastrophic		1		3
	Severity (average)	2.2		2	
BBC KS3bitesize	Cosmetic	Group 1	2	Group 2	2
	Minor		0		6
	Major		0		3
	Catastrophic		0		1
	Severity (average)	1		2.25	
Overall severity (average)		1.8		2.1	
No. of discovered real problems		25		74	

TABLE 8: Total Number of Usability Problems with Severity Ratings and Averages.

The quantitative assessment of the heuristic evaluation results was a comparison between the two sets of methods, in particular, in terms of the areas of usability problems found in this experiment. These areas assisted in identifying how each method performed in each usability problem area. The eight expert evaluators discussed, agreed and decided on the categories to which the problems should belong in both sets, as Tables 9 and 10 illustrate. The overall results from both tables show that the two groups (and the three websites) revealed more usability problems by using DSI in all areas than HE, particularly in Learning process, Motivational factors, Design and media usability, User usability and Content information and process orientation, respectively. Three out of the ten HE heuristics performed more efficiently than four others and the three remaining failed to expose a sufficient number of usability problems. This suggests that the HE heuristics are rather too general, and are unlikely to encompass all the usability attributes of user experience and design in interactive learning systems.

Heuristic Evaluation	Skoool	AcademicEarth	BBC KS3bitesize
Visibility of system status	1	2	0
Match between the system and the real world	0	4	0
User control and freedom	1	3	0
Consistency and standards	1	0	0
Error prevention	0	1	0
Recognition rather than recall	3	1	0
Flexibility and efficiency of use	1	0	0
Aesthetic and minimalist design	2	1	0
Helps users recognize, diagnose and recover from errors	0	0	0
Help and documentation	1	1	2
Total problems	10	13	2

TABLE 9: Usability Problems Found by Category through HE.

Usability problem area	Skoool	Academic Earth	BBC KS3bitesize
User usability	4	5	2
Motivational factors	5	6	1
Content information and process orientation	4	3	0
Learning process	11	7	7
Design and media usability	9	8	3

TABLE 10: Usability Problems Found by Category through DSI.

6.1.3 UEM Performance Metrics

After employing the above formulae in terms of the UEM metrics, and as depicted in Table 11 below, the t-test was used to investigate the statistical differences between the DIS and HE. The measure of thoroughness for the DSI set in identifying the number of real problems was higher than for HE (0. 5 vs. 0.2). Also, the t-test revealed significant difference between them ($t = -2.227$, $df = 19.208$, $p = 0.037$). Also, the measure of validity for DSI was higher (in accurately identifying real usability problems) than for HE (0.7 vs. 0.4). There was a significant difference between them ($t = -2.966$, $df = 20.705$, $p = 0.007$). The effectiveness of DSI was also higher that for HE (0.3 vs. 0.1). Again, there was a significant difference between them ($t = -2.212$, $df = 21.717$, $p = 0.038$). Furthermore, the reliability values for DSI were slightly higher than for HE (0.5 vs. 0.4). It can now be concluded that there is agreement amongst the evaluators over the usability problems, and it is of a high level. Also, the t-test revealed a significant difference ($t = 3.181$, $df = 11.326$, $p < 0.000$). Finally, the average results for the cost of employing the two methods show that there is a slight difference (Table 12); (DSI = $1.240 vs. HE = $1.017).

Method	Skoool		Academic Earth		BBC KS3bitesize		**Overall mean**	
	HE	DSI	HE	DSI	HE	DSI	HE	DSI
Thoroughness	0.3	0.5	0.3	0.4	0	0.3	0.2	0.5
Validity	0.3	0.7	0.5	0.8	0	0.5	0.4	0.7
Effectiveness	0.1	0.4	0.2	0.3	0	0.2	0.1	0.3
Reliability	0.4	0.4	0.4	0.6	0.3	0.5	0.4	0.5

TABLE 11: Mean Score for UEM for the Two Methods.

Method	Skoool	Academic Earth	BBC KS3bitesize	Mean cost
Heuristic Evaluation (HE)	**$1,020** This includes the time spent by 4 evaluators (1.6 hours) + hours collecting data from the evaluation sessions + 6 hours analysing data.	**$990** This includes the time spent by 4 evaluators (1.3 hours) + hours collecting data from the evaluation sessions + 6 hours analysing data.	**$1,040** This includes the time spent by 4 evaluators (1.8 hours) + 2.6 hours collecting data from the evaluation sessions +6 hours analysing data.	**$1017**
Domain Specific Inspection (DSI)	**$1,130** This includes the time spent by 4 evaluators (2.7 hours) + 3 hours collecting data from the evaluation sessions + 6.6 hours analysing data.	**$1,250** This includes the time spent by 4 evaluators (2.9 hours) + 3 hours collecting data from the evaluation sessions + 6.6 hours analysing data.	**$1,240** This includes the time spent by 4 evaluators (2.8 hours) + 3 hours collecting data from the evaluation sessions + 6.6 hours analysing data.	**$1206**

TABLE 12: Cost of Employing Both Methods.

6.1.4 Post- test Questionnaire

- **Satisfaction score:** The researchers used the System Usability Scale (SUS) as previously mentioned. HE delivered a lower overall score, at 46, whereas DSI delivered a much higher score, at 71.

- **Opinion and attitudinal questions (Likert scale)**
  The expert evaluators completed the aforementioned Likert scale, and the scores were calculated for each statement of Likert questionnaire in order to obtain the overall results concerning the expert evaluators' opinion with the DSI and HE. A Likert score of 1-2 was regarded as a negative response, 4-5 a positive response, and 3 a neutral one. The Cronbach's Alpha test used to measure the reliability of responding and the result was 0.84. The Likert scores revealed that evaluators satisfied overall with the DSI and the results were significant differences between DSI and HE by using Mann-Whitney test as Table 13 shows.

Method	Q1	Q2	Q3	Q4	Q5	Q6	Q7
Mann-Whitney U	3.000	8.000	11.000	5.000	14.500	12.000	.000
P-value (2-tailed)	.002	.009	.020	.004	.042	.023	$p <$ 0.001

TABLE 13: Results of Mann-Whitney for Both Methods.

6.2 Quantitative Analysis for Usability Testing Result

6.2.1 Time Spent

Table 14 shows the time spent by each user on performing the experiment. The Skoool groups spent the longest time, more than the BBC KS3bitesize and Academic Earth groups, with 112, 96 and 88 minutes, respectively. This again was probably due to problems in navigation, structure and function in the three websites, which caused the users to spend more time in accomplishing their tasks. This was particularly so in the Skoool website, as some tasks were abandoned because the users had doubts about how to accomplish them. Also, in the BBC KS3bitesize

website, the group spent time thinking about how to perform some tasks, such as the 'registration' task and the 'post a question' task. The average time spent by each user in all three groups was more than 1.1 minutes. The efficiency formula used for UT for all the experiments, in terms of number of usability problem discovered over time spent, delivered a mean score of 0.4 (Skoool = 0.1, Academic Earth = 0.1, BBC KS3bitesize = 0.2). The One-way ANOVA test was used to determine any significant difference in terms of time spent; the results reveal significant a difference ($F = 6.616$, $p = 0.003$). However, there was no statistical difference in terms of efficiency ($F = 0.109$, $p = 0.458$).

Usability measure	Skoool	Academic Earth	BBC KS3bitesize
Total time spent by all users (in minutes)	112	88	96
Average time per user per task (in minutes)	1.4	1.1	1.2
Average time per user over four tasks	5.6	4.4	4.8

TABLE 14: Time Taken on Conducting the Evaluation.

6.2.2 User Satisfaction

It can be seen clearly that BBC KS3bitesize delivered the highest overall score, at 7, whereas Skoool delivered the second highest score, at 5, and Academic Earth delivered the lowest score among the three websites, at 3. This indicates that there were certain factors that influenced the users, which then affected the satisfaction rating for the tested website, as evidenced by the critical user comments on the design features of each website. These factors are the various activities, such as the test and revise functions that each website provided (or the games); also, the users were encouraged by simple and attractive designs. These results are similar to what evaluators stated.

6.2.3 Number of Usability Problems Discovered

Table 15 details the total numbers of usability problems found by user testing and their severity rating. All the redundant problems were removed. The usability problems detected in BBC KS3bitesize numbered 16, higher than in the Skoool and Academic Earth websites (13 vs. 12). The One-way ANOVA test was used, revealing no statistical difference amongst the numbers of problems found ($p > 0.05$). Pearson correlation was used, and the results reveal a positive relationship between time spent and problems discovered (the Sig value is 0.013). This result reveals that the users who spent more time were able to discover more usability problems.

Problem type	Skoool	Academic Earth	BBC KS3bitesize	Total no. of problems without duplication
	Total no. of usability problems	Total no. of usability problems	Total no. of usability problems	
Catastrophic	1	0	0	1
Major	3	3	2	8
Minor	2	2	5	9
Cosmetic	7	7	9	23
No. of problems	13	12	16	41

TABLE 15: Numbers of Usability Problems Discovered.

6.2.4 UEM Performance Metrics

By applying the UEM and reliability formulae, Table 16 details the thoroughness of UT in identifying real usability problems, with a mean of 0.3. The validity of UT in finding the known

usability problems was 0.2, and the effectiveness of UT in identifying usability problems related to the user interface was 0.1. The One-way ANOVA test was used to identify any significant difference between them in each website as a dependent factor. The results reveal that there are significant differences in regard to thoroughness ($F = 9.873$, $p < 0.001$), validity ($F = 8.435$, $p = 0.001$), effectiveness ($F = 7.754$, $p = 0.001$) and reliability ($F = 9.612$, $p < 0.001$). The results for the costs of employing UT on each website were little different with an average $1,667, as shown in Table 176.

Metric / Websites	Skoool	Academic Earth	BBC KS3bitesize	Mean Total
Thoroughness	0.4	0.3	0.1	0.3
Validity	0.3	0.3	0.1	0.2
Effectiveness	0.1	0.1	0.01	0.1
Reliability	0.8	0.9	0.7	0.8

TABLE 16: The Mean Results for the UEM Metrics.

Evaluation method	Skoool	Academic Earth	BBC KS3bitesize	Mean cost
User Testing (UT)	**$1,690** This includes the time spent by 20 users (1.9 hours) + 10 hours collecting data from the evaluation sessions + 5 hours analysing data.	**$1,650** This includes the time spent by 20 users (1.5 hours) + 10 hours collecting data from the evaluation sessions + 5 hours analysing data.	**$1,660** This includes the time spent by 20 users (1.6 hours) +10 hours collecting data from the evaluation sessions + 5 hours analysing data	**$1667**

TABLE 17: Cost of Employing UT in This Research.

6.3 Comparative Analysis to Evaluate the Adaptive Framework

This section represents comprehensive and comparative analysis between the three methods.

6.3.1 Types of Problems Found by UT in Relation to DSI and HE

Two independent expert evaluators were involved in discussing, agreeing on and deciding where the UT problems should be in HE, and to which category they should belong in DSI, as Tables 18 and 19 illustrate. The overall results from both tables show that all the UT problems were successfully classified into DSI, whereas just 11 problems out of 16 in the BBC KS3bitesize were classified into HE. This proves that HE is rather general, and is unlikely to encompass all user problems, such as usability problems in the Learning process area. Also, the tasks given to the users during the usability testing seem to have 'walked them through' the quality of the 'learning process', which could have increased the opportunity to discover problems. Furthermore, the findings confirm that User control and freedom, Motivational factors and Content information and Process orientation are a common weakness in dynamic websites (particular for educational websites). All the three websites found nearly equal numbers of usability problems related to navigation and visibility. In conclusion, UT worked better than HE because seven problems were not classified in it. However, all the users' problems were classified in the DSI.

Heuristic Evaluation	Skoool	Academic Earth	BBC KS3bitesize
Visibility of system status	6	1	4
Match between the system and the real world	2	0	3
User control and freedom	0	0	0
Consistency and standards	0	1	0
Error prevention	0	4	0
Recognition rather than recall	0	0	0
Flexibility and efficiency of use	1	0	0
Aesthetic and minimalist design	2	5	3
Helps users recognize, diagnose and recover from errors	0	0	0
Help and documentation	1	0	1
Total problems	**12**	**11**	**11**

TABLE 18: Usability Problems Found Compared with HE.

Usability problem area	Skoool	Academic Earth	BBC KS3bitesize
User usability	2	3	2
Motivational factors	1	0	1
Content information and Process orientation	1	1	1
Learning process	6	7	9
Design and media usability	3	1	3
Total problems	**13**	**12**	**16**

TABLE 19: Usability Problems Found Compared with the Domain Specific Inspection (DSI).

6.3.2 Performance of the Three Methods

Generally, Tables 20, 21 and 22 show how UT, HE and DSI revealed different types and numbers of usability problems. One-way ANOVA reveals that there is significant difference between the three methods in terms of discovering usability problems on the whole ($F = 13.447$, $p < 0.001$). UT, HE and DSI revealed 80%, 10% and 60% of the usability problems found in the BBC KS3bitesize website, respectively. One-way ANOVA-Tukey HSD was used and the results show that there is a strongly significant difference amongst the methods in finding usability problems on the BBC KS3bitesize website between HE and UT, where $p = 0.003$. In the Skoool website, UT, HE and DSI revealed 38%, 29% and 97% of the found usability problems, respectively. One-way ANOVA-Tukey HSD was used and the results show that there is a strongly significant difference amongst the methods in finding usability problems in Skoool (as a dependent factor), particular between HE and DSI and between DSI and UT, where $p < 0.001$. Finally, UT, HE and DSI revealed 34%, 37% and 83% of the found usability problems in Academic Earth, respectively. One-way ANOVA-Tukey HSD was used and the results show that there is a significant difference amongst the methods in finding usability problems in Academic Earth between HE and UT, where $p = 0.044$. The performance of HE in discovering usability problems during the experiment ranged from 10% to 37%. UT discovered usability problems ranging from 38% to 80%, while DSI discovered usability problems ranging from 60% to 97%. Also, UT and HE performed better in discovering major, minor and cosmetic real usability problems, but DSI was the best in discovering more catastrophic, major, minor and cosmetic real usability problems. Furthermore, 9 unique problems were discovered in all experiments on the three websites through UT (6 in BBC KS3bitesize and 3 in Academic Earth), whereas the remaining UT problems were discovered by DSI (although one was discovered by HE). Thus, it can be seen that DSI was the best in discovering real problems; this was followed by UT, and then finally HE.

Method / Problem type	UT	HE	DSI	Total problems (no duplicates)
Catastrophic	0 (0%)	0 (0%)	1 (100%)	1
Major	2 (66%)	0 (0%)	3 (100%)	3
Minor	5 (100%)	0 (0%)	5 (100%)	5
Cosmetic	9 (100%)	2 (22%)	2 (22%)	11
No. of problems	**16 (80%)**	**2 (10%)**	**12(60%)**	**20**

TABLE 20: Findings in BBC KS3bitesize.

Method / Problem type	UT	HE	DSI	Total problems (no duplicates)
Catastrophic	1 (25%)	2 (50%)	4 (100%)	4
Major	3 (30%)	2 (20%)	6 (89%)	7
Minor	2 (29%)	3 (43%)	11 (85%)	11
Cosmetic	7 (54%)	3 (23%)	12 (92%)	12
No. of problems	**13 (38%)**	**10 (29%)**	**33 (97%)**	**34**

TABLE 21: Findings in Skoool.

Method / Problem type	UT	HE	DSI	Total problems (no duplicates)
Catastrophic	0 (0%)	1 (33%)	3 (100%)	3
Major	3 (50%)	3 (50 %)	4 (66%)	6
Minor	2 (17%)	7 (58 %)	11 (92%)	12
Cosmetic	7 (50%)	2 (14%)	11 (79%)	14
No. of problems	**12 (34%)**	**13 (37%)**	**29 (83%)**	**35**

TABLE 22: Findings in Academic Earth.

6.3.3 Overlapping and Unique Problems

Many researchers recommend conducting UT together with HE because they have found that each method discovers unique problems [36], so when they are conducted together, they can reveal and present all the problems in the targeted website. Again, this experiment may confirm or deny this recommendation, depending on the following results. Table 23 shows the performance of the three methods on a unique performance basis for the three websites. DSI was able to discover 7 catastrophic, 12 major, 20 minor and 16 cosmetic problems that were not revealed by the other methods. HE was not able to identify any catastrophic problems alone; however, it was able to identify 1 major, 2 minor and 3 cosmetic problems. UT was not able to discover any major problems; however, it discovered 1 catastrophic, 3 minor and 5 cosmetic problems.

Figure 6 also shows the overlapping real usability problems discovered by the three methods. In fact, each method revealed different types of problem (both unique and overlapping). However, DSI revealed the majority of real usability problems, indicating those with high severity ratings, and DSI also appeared to work fruitfully for the expert evaluators, who then revealed more real problems, both unique and overlapping. For example, DSI found 62% unique problems out of the total number of real usability problems (n = 55 out of 89). HE found 7% unique problems out of the total number of real usability problems (n = 6 out of 89), and UT identified 10% unique problems out of the total number of real usability problems (n = 9 out of 89). 19 (21%) real problems out of 89 were discovered as 'overlapping' by the three methods. The clear superiority of DSI was due to involving user inputs in drawing up the method (as in one stage of the adaptive framework), and due to DSI having characteristics that are appropriate to the educational domain.

Problem types	HE (unique)	DSI (unique)	UT (unique)	HE & UT (overlap)	DSI & UT (overlap)	DSI & HE (overlap)	Total number of problems in three websites (unique)
Catastrophic	0	7	1	0	1	3	8
Major	1	12	0	2	6	3	13
Minor	2	20	3	7	7	6	33
Cosmetic	3	16	5	9	27	7	35
Total	6	55	9	18	29	19	89

TABLE 23: Each Method's Performance with Severity Rating.

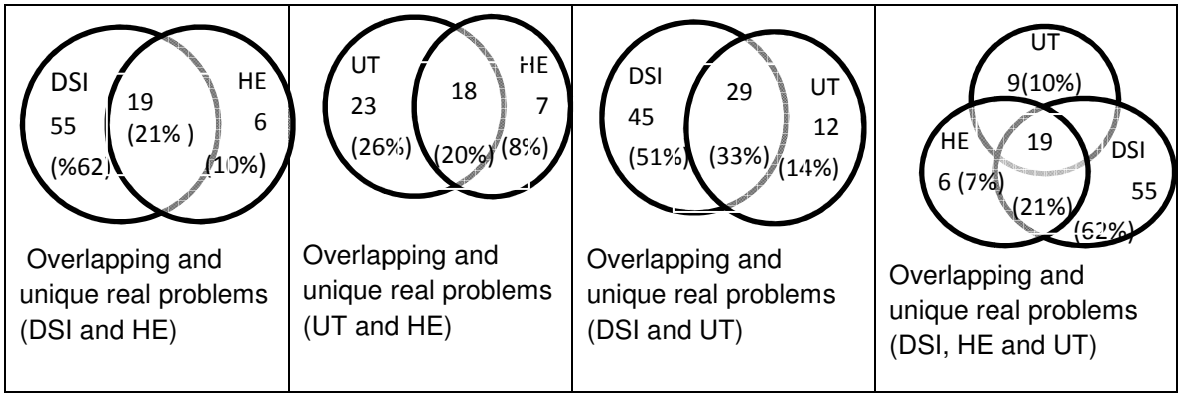

FIGURE 6: Each Method's Performance, Uniquely and Working in Pairs.

It can also be seen that combining the results of DSI with either HE or UT offers better performance in terms of catastrophic, major, minor and cosmetic problems, whereas combining HE with either DSI or UT offers quite good results in terms of cosmetic problems. Combining UT with either DSI or HE offers better results in terms of minor and cosmetic problems. To sum up, the result of the comparison between UT and HE confirms conduct UT with HE in order to overcome the shortcomings of each, because each one is indeed complementary to the other. On the other hand, DSI (as created from the adaptive framework) refutes this recommendation.

6.3.4 Usability Problem Areas: As evident in Table 24, DSI identified large numbers of real usability problems in all usability areas on the three websites (74). However, HE overall worked slightly better in discovering 25 real usability problems related to three usability problem areas but it failed in exposing any usability problems in two main usability problems areas, which are Motivational factors and Learning process, and it failed to identify a sufficient number of usability problems in the Content information and process orientation area. Furthermore, UT worked better in discovering usability problems (41) in three usability areas but it failed to identify a sufficient number of usability problems in the Content information and process orientation area.

Usability problem areas	UT	DSI	HE
User usability	✓ (14 problems)	✓ (20 problems)	✓ (15 problems)
Motivational factors	✳ (1 problem)	✓ (8 problems)	-
Content information and process orientation	✳ (4 problems)	✓ (7 problems)	✳(2 problems)
Learning process	✓ (6 problems)	✓ (20 problems)	-
Design and media usability	✓ (16 problems)	✓ (19 problems)	✓(8 problems)
Total number of problems	**41**	**74**	**25**

TABLE 24: Number of Usability Problem Areas Identified by the Three Methods.

6.3.5 Comparison between the Three Methods in UEM Performance Metrics

It can be seen from Table 25 that DSI was more efficient, thorough and effective in terms of identifying the total number of real problems to total time spent, and its ability to identify usability problems related to the user interface than the other two methods. UT is the second best method but it is more reliable than DSI or HE. HE delivered the worst result in terms of identifying real problems; however, it is the cheapest to use. Moreover, DSI is slightly more expensive than HE, and is cheaper that UT. One-way ANOVA reveals that there is significant difference between the methods used in terms of the UEM metrics results, as shown in Table 26.

Metric / Method	Efficiency	Thoroughness	Validity	Effectiveness	Reliability	Cost
HE	0.7	0.2	0.4	0.1	0.4	**$1017**
DSI	1	0.5	0.7	0.3	0.5	**$1206**
UT	0.4	0.3	0.2	0.1	0.7	**$1667**

TABLE 25: Comparing the Metrics between the Three Methods.

Metric	F	Sig. (*p*-value)
Efficiency	7.613	0.001
Thoroughness	3.950	0.023
Validity	3.525	0.034
Effectiveness	4.369	0.016
Reliability	38.571	$p < 0.000$

TABLE 26: One-way ANOVA Results between the Three Methods.

6.3.6 Advantages and Disadvantages of the Three Methods

This study has addressed the relative effectiveness of three methods for evaluating user interfaces, and it now offers some insights into each (see Table 27). Overall, DSI, as applied here, produced the best results; it found the most real problems, including more of the most serious ones, than did HE and UT, and at only a slightly higher cost. HE missed a large number of the most severe problems, but it was quite good in identifying cosmetic and minor problems. UT is the most expensive method and it missed some severe problems; however, it helps in discovering general problems and it assists, as does DSI, in defining the users' goals.

Method	Advantages	Disadvantages
Usability Testing	• Helps define and achieve the users' goals • Identifies the users' real problems • Identifies recurrent and general real problems	• Misses some severe real problems • High cost • Takes more time • Conducted under lab conditions
Heuristic Evaluation	• Identifies a few real problems • Low cost	• Misses some severe problems • Too general • Not readily applicable to many new domains
Domain Specific Inspection (DSI)	• Identifies many more real problems • Identifies more serious, major, minor and cosmetic real problems • Improves the evaluator's performance • Identifies the users' real problems and helps define and achieve the users' goals	• A little higher in cost than HE and cheaper than UT • Slightly higher in time than HE

TABLE 27: Summary of the Study's Findings.

7. DISCUSSION AND FINDINGS

This section explores the results of this experiment and highlights the main findings. It then draws out the lessons learned from the research. The main objective of this experiment was to evaluate the adaptive framework through its ability to generate new method, and specifically the specific inspection method for Educational domain\ (DSI) by comparing its results with usability testing (UT) and Heuristic Evaluation (HE). It has been clearly shown that the hypotheses were accepted and DSI was able to find all the real problems that were discovered by the UT and HE, but with greater efficiency, thoroughness and effectiveness. Also, DSI was better at discovering catastrophic, major, minor and cosmetic real problems. It seemed to guide the evaluators' thoughts in judging the usability of the website through clear guidelines that included all aspects of the educational quality of the website, which were represented in Content information and process orientation, Management of learning process, Quality of design and media, and Motivational factors to learn. As a result, it is unsurprising that the DSI method revealed a number of problems not discovered by the other two methods. HE method did not perform as well as either DSI or UT, based on the number of usability problems discovered during this experiment. The experts that used HE seemed to have their confidence undermined whilst performing the evaluation, for example, when they performed the evaluation, they found no readily applicable heuristic within HE for performing some of the main functions in these educational websites, such as Educational process and management. Consequently, HE performed poorly in discovering problems. The UT method performed modestly against DSI, and well against HE, based on the number of problems identified. Thus, the findings indicate that it is not essential to conduct UT in conjunction with HE, in order to address the shortcomings of these methods; rather, to avoid wasting money, an alternative that is well-developed, context-specific and capable, such as what has been generated here for educational domain, should be employed. Furthermore, the adaptive framework provided optimal results regarding the identification of comprehensive 'usability problem areas' on the educational websites, with minimal input in terms of cost and time spent in comparison with the employment of usability evaluation methods. The framework was used here to generate DSI, which helped to guide the evaluation process as well as reducing the time that it would have taken to identify these usability issues through current evaluation methods. In terms of the definition of missed problem given by [15], we can consider that the problems that were found by any one method and not found by the others as missed problems. From this standpoint, DSI missed discovering 15 real usability problems. However, HE and UT missed 64 and 61 real usability problems, respectively.

The above findings facilitate decision-making with regard to which of these methods to employ, either on its own or in combination with another, in order to identify usability problems on

educational websites. The selection of the method or methods will depend on the types of problem best identified by each of them.

8. CONCLUSION AND FUTURE WORK

Contrary to most of the efforts to construct and test enhanced usability methods, our work here has made explicit the process for so doing. The adaptive framework includes the views of users and usability experts to help generate a context-specific method for evaluating any chosen domain. The work presented here illustrates and evaluates this process for the generation of the DSI method to assess the usability of educational websites. DSI outperformed both HE and UT, even when taken together. This clearly represents a step in the right direction. Further validation of the use of our adaptive framework will indicate whether it is indeed applicable across domains. In order to consolidate and confirm the findings, future research could include testing the adaptive framework by developing DSI for different fields such as e-commerce and healthcare systems.

In conclusion, this research contributes to the advancement of knowledge in the field. Its first contribution is the building of an adaptive framework for generating a context-specific method for the evaluation of whichever system in any domain (Figure 1). The second contribution is the introduction of DSI that is specific for evaluating educational systems (Table 2). The third contribution is the identification of usability problem areas in the educational domain (five areas in Table 2).

9. ACKNOWLEDGEMENTS

We thank the expert evaluators in the School of Education and Lifelong Learning and School of Computing Sciences at the University of East Anglia (UEA), MSc and PhD students in both schools, the Saudi Norwich School (UK), and the expert evaluators in Aviva company in the UK for their participation in the comparative study and in the mini-usability testing experiments.

10. REFERENCES

[1] Abuzaid, R. (2010). Bridging the Gap between the E-Learning Environment and E-Resources: A case study in Saudi Arabia. *Procedia-Social and Behavioral Sciences*, 2(2): 1270-1275.

[2] AcademicEarth, (2012), *AcademicEarth*, Accessed on 3/4/2012, Available at: [http://academicearth.org/]

[3] Alias, N., Siraj, S., DeWitt, D., Attaran, M. & Nordin, A. B. (2013), Evaluation on the Usability of Physics Module in a Secondary School in Malaysia: Students' Retrospective. *The Malaysian Online Journal of Educational Technology*, 44.

[4] Alkhattabi, M., Neagu, D. and Cullen, A. (2010). Information Quality Framework for E-Learning Systems. *Knowledge Management & E-Learning: An International Journal (KM&EL)*, 2(4): 340-362.

[5] Alrobai, A. AlRoobaea, R. Al-Badi, A., Mayhew, P. (2012). Investigating the usability of e-catalogue systems: modified heuristics vs. user testing, *Journal of Technology Research*.

[6] Alshamari, M. and Mayhew, P. (2008). Task design: Its impact on usability testing. In Internet and Web Applications and Services, 2008, ICIW'08. Third International Conference on, pages 583-589. IEEE.

[7] Ardito, C., Costabile, M., De Angeli, A. and Lanzilotti, R. (2006). Systematic evaluation of e-learning systems: an experimental validation. In Proceedings of *The 4th Nordic Conference on Human-Computer Interaction: changing roles*, pp. 195-202. ACM.

[8] BBC KS3bitesize, (2012), *BBCKS3bitesize*, Accessed on 3/4/2012, Available at: [http://www.bbc.co.uk/schools/ks3bitesize/]

[9] Brooke, J. (1996). SUS - A quick and dirty usability scale. *Usability Evaluation in Industry*, pages 189-194.

[10] Chattratichart, J. & Lindgaard, G., (2007). Usability testing: what have we overlooked?. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 1415-1424). ACM.

[11] Chattratichart, J. and Lindgaard, G. (2008), A comparative evaluation of heuristic-based usability inspection methods, In the proceeding of *CHI'08 extended abstracts on Human factors in computing systems*, 2213-2220

[12] Chattratichart, J. and Brodie, J. (2002), Extending the heuristic evaluation method through contextualisation. Proc. *HFES2002*, HFES (2002), 641-645.

[13] Chattratichart, J. & Brodie, J. (2004), Applying user testing data to UEM performance metrics. In *CHI'04 extended abstracts on Human factors in computing systems* (pp. 1119-1122). ACM.

[14] Chen, S. Y. and Macredie, R. D. (2005), The assessment of usability of electronic shopping: A heuristic evaluation*, International Journal of Information Management*, vol. 25 (6), pp. 516-532.

[15] Cockton, G. and Woolrych, A. (2002). Sale must end: should discount methods be cleared off HCI's shelves? *Interactions*, 9(5): 13-18. ACM.

[16] Coursaris, C. K. & Kim, D. J. (2011), A meta-analytical review of empirical mobile usability studies. *Journal of usability studies*, 6(3), 117-171.

[17] Dumas, J. and Redish, J. (1999). *A practical guide to usability testing*. Intellect Ltd.

[18] Doubleday, A., Ryan, M., Springett, M., & Sutcliffe, A. (1997). A comparison of usability techniques for evaluating design. In *Proceedings of the 2nd conference on Designing interactive systems: processes, practices, methods, and techniques* (pp. 101-110). ACM.

[19] Fernandez, A., Insfran, E. and Abrahão, S. (2011), Usability evaluation methods for the web: A systematic mapping study*, Information and Software Technology*.

[20] Gutwin, C. & Greenberg, S. (2000), The mechanics of collaboration: Developing low cost usability evaluation methods for shared workspaces. In *Enabling Technologies: Infrastructure for Collaborative Enterprises, 2000. (WET ICE 2000). Proceedings. IEEE 9th International Workshops on* (pp. 98-103). IEEE.

[21] Hertzum, M. and Jacobsen, N. (2001). The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction*, 13(4): 421-443.

[22] Holsapple, C. W. and Wu, J. (2008), Building effective online game websites with knowledge-based trust, *Information Systems Frontiers*, vol. 10 (1), pp. 47-60.

[23] Holzinger, A. (2005), Usability engineering methods for software developers*, Communications of the ACM*, vol. 48 (1), pp. 71-74.

[24] Hart, J., Ridley, C., Taher, F., Sas, C. and Dix, A. (2008), Exploring the Facebook experience: a new approach to usability. In Proceedings of the *5th Nordic Conference on Human-Computer Interaction: Building Bridges*, pages 471-474. ACM.

[25] Hasan, L. (2009), Usability evaluation framework for e-commerce websites in developing countries.

[26] ISO (1998), *ISO 9241-11: Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs): Part 11: Guidance on Usability*.

[27] Jeffries, R., Miller, J. R., Wharton, C., & Uyeda, K. (1991, March). User interface evaluation in the real world: a comparison of four techniques. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 119-124). ACM.

[28] Khajouei, R., Hasman, A. and Jaspers, M. (2011), Determination of the effectiveness of two methods for usability evaluation using a CPOE medication ordering system*, International Journal of Medical Informatics*, vol. 80 (5), pp. 341-350.

[29] Liljegren, E. (2006), Usability in a medical technology context assessment of methods for usability evaluation of medical equipment*, International Journal of Industrial Ergonomics*, vol. 36 (4), pp. 345-352.

[30] Liljegren, E., & Osvalder, A. L. (2004). Cognitive engineering methods as usability evaluation tools for medical equipment. *International Journal of Industrial Ergonomics*, *34*(1), 49-62.

[31] Ling, C. and Salvendy, G. (2005), Extension of heuristic evaluation method: a review and reappraisal*, Ergonomia IJE & HF*, vol. 27 (3), pp. 179-197.

[32] Law, L. and Hvannberg, E. (2002). Complementarily and convergence of heuristic evaluation and usability test: a case study of universal brokerage platform. In Proceedings of *The second Nordic conference on human-computer interaction*, pages 71-80, ACM.

[33] Magoulas, G. D., Chen, S. Y. and Papanikolaou, K. A. (2003), Integrating layered and heuristic evaluation for adaptive learning environments, In the proceeding of *UM2001*, 5-14.

[34] Molich, R. and Nielsen, J. (1990), Improving a human-computer dialogue, *Communications of the ACM*, vol. 33 (3), pp. 338-348.

[35] Muir, A., Shield, L. and Kukulska-Hulme, A. (2003), The pyramid of usability: A framework for quality course websites, In the proceeding of *EDEN 12th Annual Conference of the European Distance Education Network, The Quality Dialogue: Integrating Quality Cultures in Flexible, Distance and eLearning*, 188-194. Rhodes, Greece.

[36] Nielsen, J. (1992), Finding usability problems through heuristic evaluation, In the proceeding of *ACM CHI'92 Conference* 373-380. Monterey, CA, USA, May 3-7.

[37] Nielsen, J. (1994a), Heuristic evaluation, *Usability inspection methods*, vol. 24, pp. 413.

[38] Nielsen, J. (1994b), *Usability engineering*, Morgan Kaufmann.

[39] Nielsen, J. and Loranger, H. (2006), *Prioritizing web usability,* New Riders Press, Thousand Oaks, CA, USA.

[40] Nielsen, J. and Molich, R. (1990), Heuristic evaluation of user interfaces, In the proceeding of *SIGCHI conference on Human factors in computing systems: Empowering people*, 249-256.

[41] Nielsen, J. (2000), HOMERUN Heuristics for Commercial Websites, in www.useit.com.

[42] Nielsen, J. (2001), "Did poor usability kill e-commerce", in www.useit.com.

[43] Nayebi, F., Desharnais, J. M. & Abran, A. (2012), The state of the art of mobile application usability evaluation. In *Electrical & Computer Engineering (CCECE), 2012 25th IEEE Canadian Conference on* (pp. 1-4). IEEE.

[44] Oztekin, A., Kong, Z. J. and Uysal, O. (2010), UseLearn: A novel checklist and usability evaluation method for eLearning systems by criticality metric analysis, *International Journal of Industrial Ergonomics*, vol. 40 (4), pp. 455-469.

[45] Sears, A. (1997), Heuristic walkthroughs: Finding the problems without the noise, *International Journal of Human-Computer Interaction*, vol. 9 (3), pp. 213-234.

[46] Shackel, B. and Richardson, S. J. (1991), *Human factors for informatics usability*, Cambridge University Press.

[47] Skoool, (2012), *Skoool*, Accessed on 3/4/2012, Available at: [http://lgfl.skoool.co.uk//]

[48] Stracke, C. and Hildebrandt, B. (2007). Quality Development and Quality Standards in e-Learning: Adoption, Implementation and Adaptation. In Proceedings of *World Conference on Educational Multimedia, Hypermedia and Telecommunication 2007*, pp. 4158-4165.

[49] Thompson, A. and Kemp, E. (2009), Web 2.0: extending the framework for heuristic evaluation. In Proceedings of *The 10th International Conference NZ Chapter of the ACM's Special Interest Group on Human-Computer Interaction*, pp. 29-36. ACM.

[50] Wilson, C. (2007). Taking usability practitioners to task. *Interactions*, 14(1): 48-49.

[51] Van den Haak, M., de Jong, M. and Schellens, P. (2004), Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: a methodological comparison. *Interacting with computers*, 16(6): 1153-1170.

[52] Zaharias, P. & Poylymenakou, A. (2009), Developing a usability evaluation method for e-learning applications: Beyond functional usability. *Intl. Journal of Human-Computer Interaction*, *25*(1), 75-98.