A Framework for Generating Domain-Specific Heuristics for Evaluating Online Educational Websites - Further Validation

Roobaea AlRoobaea,
College of Computers & Info. Systems, Taif University, Saudi Arabia and
School of Computing Sciences, University of East Anglia, Norwich, UK
Email: r.alrobaea@uea.ac.uk

Ali H. Al-Badi,
Information Systems Dept., CEPS, SQU, Oman
Email: aalbadi@squ.edu.om

Pam J. Mayhew
School of Computing Sciences, University of East Anglia, Norwich, UK
Email: p.mayhew@uea.ac.uk

Abstract *:*

Websites are growing in use worldwide but need to be continuously evaluated and monitored to measure their efficiency, effectiveness and user satisfaction, and ultimately to improve quality. For this purpose, heuristic evaluation (such as Nielsen's heuristics) and usability testing have become the two most widely used methods for measuring a system's usability from the perspective of expert evaluators and real users of the system. It is recommended that heuristic evaluation be conducted in conjunction with usability testing because they complement each other. However, Nielsen's heuristics' are general and not readily applicable to new domains. Also, user testing is expensive and time-consuming.

For these reasons, the researchers have enhanced these evaluation methods by building a framework designed to improve the usability assessment process for websites in any chosen domain, by generating specific set of heuristics, in this case for use with educational domain. It is designed to avoid the drawbacks of having to use both general heuristics and usability testing, and it combines their advantages. Also, it helps researchers to combine feedback from both expert evaluators and potential users in a chosen domain in order to create focused heuristics. This paper aims to conduct a secondary validation stage to validate the proposed framework by conducting usability testing. Also, it is to investigate whether it is essential to conduct usability testing in conjunction with new heuristics. The results of this usability experiment, which are based on usability problems and severity of uncovered problems, are compared with Nielsen's heuristics and the newly developed heuristics, namely Educational Heuristics. This enables the assessment of their relative efficiency, effectiveness, thoroughness, validity, costs of employing and problem-area identification.

The results show that the proposed framework succeeded in building a new set of heuristics for educational websites, which managed to discover more problems than Nielsen's heuristics or usability testing. In order to remove most usability problems and to avoid wasting money, well-developed, context-specific heuristics, such as our Educational Heuristics, should be employed. These context-specific heuristics can be successfully created by the framework designed herein.

Keywords-Heuristic evaluation (HE);User testing(UT);Framework; Nielsen's heuristics (NH); Educational Heuristics (EH); Usability problem and Severity rating

I. INTRODUCTION

Dynamic websites such as educational websites are now growing in use around the world but this new learning approach must be continuously evaluated and monitored to measure their success, levels of efficiency, effectiveness and satisfaction, and ultimately to improve their quality. Web design is a key factor in determining the success of educational 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 objectives. We need an effective methodology and appropriate heuristics to perform this task. Various evaluation methods, tools and techniques can be used but the most well-known evaluation methods used to assess the user-interface and to discover usability problems are heuristic evaluation and user testing. The heuristic evaluation method depends on a set of usability principles or 'heuristics'. In 1994, Nielsen developed these guidelines (or heuristics) for measuring usability, and they are still being used. It is a well-known method used to find usability problems in a user interface [1]. However, it has been criticised as being too general and lacking in guidance for evaluators [2]. Usability testing (also known as user testing) is another

important evaluation method for ensuring system quality, in particular for websites, although it is claimed to be costly [3]. 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 [4].

Educational websites must be effective if they are to meet the student's needs. To address this, there is a need to develop or design new evaluation methods or at least to create a set of methods (i.e. processes) that help to overcome the shortcomings of both Nielsen's original heuristics and user resting in order to employ traditional evaluation methods in novel ways. The researchers here have developed a framework that is designed to avoid the drawbacks of having to use both general heuristics and usability testing, and it combines their advantages by generating specific heuristics for any domain. This framework is used to generate a new set of heuristics specifically for online educational websites and social network websites [6]. It consists of two validation stages to validate practically its efficiency in achieving its goals. The first validation stage was conducted by comparing Nielsen's heuristics (NH) with the specific educational heuristics (EH), and published in a separate paper [5].

This paper aims to conduct usability testing as a second stage to validate the proposed framework. Also, it is to investigate whether it is essential to conduct usability testing in conjunction with specific domain heuristic. The results of this usability experiment, which are based on usability problems and type of uncovered problems, are compared with Nielsen's heuristics and the newly developed heuristics (their result in [5]), namely Educational Heuristics that , in terms of the number and severity of the discovered problems. This enables the assessment of their relative efficiency, effectiveness, thoroughness, validity and problem-area identification. This paper is organized in the following way. Section 2 starts with a brief literature review, including definitions of usability problems and the concept of severity rating, and details some related work. Also, it describes the proposed framework for constructing domain heuristics and its validation stages. Section 3 shows the preparation and actual evaluation steps to conduct the user testing (second validation stage) on three case studies. Section 4 and 5 provides an data analysis and measurement and discussion of the results. Section 6 presents the conclusion and suggestions for future work.

II. LITERATURE REVIEW

Distance learning websites are becoming increasingly developed in the midst of the Internet revolution and ever-improving information technologies. In fact, the growth of the Internet has led to an explosion of educational website content, rising in accordance with demand. [7] asserts that e-learning occurs when students use the Web to proceed through the sequence of teaching, completing the learning activities, and achieving the learning results and objectives. In other words, "users in any place and time can have access to distributed learning material and processes over the Internet. It could be part of a winning strategy for particular needs, such as decongestion of overcrowded education facilities, support for students or teachers and adult education" [8]. However, some of them are difficult to use due to the inexperience of many of the designers and the lack of effective, efficient and accurate or appropriate guidelines for performing this task. Consequently, users spend more time learning how to use the website than learning the educational content. [9] states, "the quality is considered a crucial issue for education in general, and for e-learning in particular". The need for e-learning websites to be of sufficiently high quality cannot be overstated, and so it is extremely important, indeed necessary, to classify suitable criteria for addressing and assessing quality [10].

Emanating from the development of Web 2.0, there is now a need to study the usability of educational websites. The 'usability' of an interface is one of the most significant aspects affecting the quality of user experience and the success of these websites. Poor websites may have a negative impact on various aspects of educational institutions, as they do not allow users to achieve their goals efficiently, effectively, and with a high degree of satisfaction [11]. Also, users spend more time learning how to use these websites rather than on the content, causing frustration leading to abandonment of the site. ISO defines usability as "the extent to which the product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use" [11]. Muir et al. 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"[12]. Usability is part of the Human-Computer Interaction (HCI) field, which has become a major area of research; it utilizes many and various techniques for evaluating the usability of interactive systems and the phenomena surrounding them [13]. These techniques are classified into inspection, inquiry and testing methods. The most well-known methods used to assess user interfaces and

discover usability problems are heuristic evaluation and usability testing.

Heuristic evaluation was developed by [1]; it is an inspection method that guides the evaluator through a set of general usability principles or 'heuristics'. 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. [14] states, "heuristic evaluation is a widely accepted method for diagnosing potential usability problems and is popular in both academia and industry". It is superior to all other methods in carrying out evaluation in less time and by a smaller number of expert evaluators; it does not require any special tasks or equipment. Also, it can be used early in the development process, and may be used throughout the development process [1].

In the present context and in relation to heuristic evaluation, usability testing (also known as user testing) is another important evaluation method for ensuring system quality, in particular for websites, although it is claimed to be costly [3]. 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 [4]. Dumas and Redish [15] stressed that a fruitful usability testing session needs careful planning and attention to detail. 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 protocol is one technique that is widely used with user testing. There are three types of Think- Aloud: concurrent, retrospective and 'constructive interaction'. Concurrent Think-Aloud is the most common type; this involves participants verbalising their thoughts whilst performing tasks in order to evaluate an artefact. Retrospective Think-Aloud 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 [16].

The result of applying these methods is a list of usability problems [17]. These problems are classified into different groups to which a numeric scale is applied to measure the severity of each problem or issue; 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, and fixing this should be given low priority; then, this is a major usability problem, and it is important to fix this, so it should be given high priority; and finally, this issue is a usability catastrophe, and it is imperative to fix this before the product is released.

In the early years of computing, Nielsen's heuristics were widely applied in measuring the usability of Web interfaces and systems because it was the only such tool available. Many researchers then sought to compare and contrast their efficiency with other methods such as user testing. Some researchers used both Nielsen's heuristics and user testing in order to comprehend the success and popularity of the former in assessing the usability of Web-based systems such as social networks, video games, online shopping websites and e-learning applications. Lately, their findings have been almost unanimous in one respect: Nielsen's heuristics are too generic and thus not particularly useful. They argue that they did not find a sufficiently high number of usability problems; in most cases, their own extended or modified heuristics worked better [2, 18, 19]. For this reason, many researchers have enhanced Nielsen's generic heuristics by extending or modifying them in order to make them applicable to the features of other interactive systems [20, 21, 22, 23, 24]. Some researchers even found that the tested websites failed in certain respects, according to their extended or modified heuristics [19, 25]. Furthermore, a number of published studies have compared these evaluation methods. They found that user testing did not work better in discovering more problems; however, they reported that more severe problems were discovered through usability testing, as compared with heuristic evaluation. For example, Tan et al. [26] found that heuristic evaluation discovered about 60% of the usability problems, but that usability testing revealed only half this figure. Furthermore, 10% of the usability problems were uncovered by both of these methods. In addition, many researchers are claimed that user testing is more costly than heuristic evaluation, consuming more time, and needing carefully designed tasks based on real users [3, 27, 28, 29]. Consequently, there is a need for a new method to overcome the shortcomings of Nielsen's original heuristics, extended heuristics, modified

heuristics and user testing; also, it should be more efficient and cheaper and should greatly facilitate the assessment of website usability.

Sifting through the literature, the researchers found that the heuristic evaluation has many advantages that set it distinguished, so they decided to be the basic nucleus for the development of a new method by making it context-specific heuristics. The researchers here found that there is no comprehensive methodology or framework for generating context-specific heuristics. This finding and the criticality of website usability were the reasons that motivated these researchers to propose a framework for generating domain-specific heuristics for website evaluation [5,6]. This framework is applicable across numerous domains. It is characterized by being optimal, cheap and quick, and using only experts, although it should include user experience (thereby involving the advantages of user testing but avoiding its drawbacks). The development phase of the proposed framework consists of four main stages. It includes conducting a highly focused literature review, learning of the users' perspectives and requirements (through mini-usability testing and learning from errors) exploiting the expertise and experience of experts in the field (through focus group/discussion), and analysing all the gathered data. The following is an explanation of these stages:

Development Stage One (D1: Familiarisation): the aim of this stage is to review all published material in the area of usability evaluation methods with a specific focus on heuristic evaluation and to collect any information related to the target domain.

Development Stage Two (D2: User Input):this stage consists of a mini-usability test (task scenarios and questionnaire). Users are asked to perform a set of tasks on a typical domain website, and then to fill out a questionnaire. The broad aim of this is to elicit feedback on a typical system in order to appreciate the user perspectives, requirements and expectations.

Development Stage Three (D3: Expert Input (focus group/interview)): this stage entails a discussion amongst experts (in the domain and/or usability) in order to obtain a broader understanding of the specifics of the prospective domain. It is widely employed in various types of research such as product planning and usability studies. This assists in understanding the issues arising from the questionnaire results and also in garnering more information through conversations with expert evaluators (in the domain and/or usability)

Development Stage Four (D4: Draw Up Heuristics (data analysis)): the aim of this stage is to analyse all the data gathered from the previous three stages and highlight the characteristics of the target domain. The collected data consist of general areas of usability related to the selected domain. Then, the researchers establish the new heuristics for the selected domain, and a list of usability problem areas is constructed from the gathered information drawn from the three stages above. After constructing the framework, it is validated intensively through rigorous validation methods as follows:

1. Heuristic Validation (Expert Evaluation): the aim of this stage is the validation of the newly proposed heuristics by conducting a heuristic evaluation. Expert evaluators need a familiarization session before the actual evaluation. This session is important because if the application domain is unfamiliar to the evaluators, the session provides an opportunity to become acquainted with the domain and with the evaluation environment for testing the proposed heuristics on the targeted websites [17]. The collected data are then analysed empirically. The details of this stage are outlined in another paper [5].

2. Testing Validation: (User Evaluation): the aim of this stage is to complement the results obtained from the expert evaluation, by carrying out usability lab testing on the same websites. Nielsen [30] recommends conducting usability testing with heuristics evaluation because each one is complementary to the other. Then, the efficiency and effectiveness of both sets of heuristics are assessed and compared against the lab testing in terms of discovering problems and their types. This paper focuses on this validation stage to fully complete the validation stages of the proposed framework on the online educational websites.

The researcher conducted the first validation stage [5], and the result showed that the proposed framework succeeded in building specific Educational Heuristics, which managed to discover uniquely 55(69%) of the usability problems (80 problems in total), in comparison with Nielsen's heuristics, which discovered only 6(8%). 19 problems (24%) were commonly discovered (overlapping or sharing) by both sets of heuristics. The time taken using Nielsen's heuristics was less than the time taken using the EH (56 minutes vs. 72 minutes). Also, there was a slight difference in the cost of employing the sets of heuristics (NH=$3.228 vs. EH=$3.751). EH provided a more efficient result than NH through both groups in discovering usability problems (EH = 1.74 vs. NH = 1.03). Thoroughness for EH in identifying the number of real problems at Level 3 or 4 is higher than that of NH (0.1 vs.

0.04). There is a slight difference between the validity of both sets of heuristics in accurately identifying usability problems (EH = 0.12 vs. NH = 0.11). For effectiveness, EH (at 0.011) was slightly more effective in identifying usability problems than NH (0.004).

III. RESEARCH METHODOLOGY

In the User Validation Stage, a pilot experiment was conducted to make sure that there were no spelling or grammatical errors and no ambiguous words or phrases, and that all of the sentences in the instruments (task design, observer sheet and questionnaire) were sufficiently clear to be used by the evaluators. A fewer minor improvements were made and testing proper began. There is no agreement on how many users should be involved in usability testing. Dumas and Redish [15] suggested that 6 to 12 users are sufficient for testing. In this experiment, the researchers recruited 30 users to conduct the user testing, who were chosen carefully to reflect the real users of the targeted website (Skoool, AcademicEarth and BBC KS3bitesize). Then, they were divided into three groups, with 10 users in each website's group. The evaluation started by checking the suitability of the environment, such as using the same type of machine and Internet browser connection speed, and choosing a quiet room. Also, a brief introduction was given on how use the Think-Aloud protocol and on the aim of the different task designs. After that, the actual test was started; each user was given the 'task scenarios' sheet and asked to read and then perform one task at a time. There were three main tasks for each website, which included four sub-tasks in each main task. All of these tested the same goal but the formulations were different with a mixture of Think-Aloud approaches. The researchers observed the behaviour of the users as they performed the given tasks, and took notes on the different (and sometimes interesting) activities observed. Also, the Morae software package was used to record the users' behaviour. The users were asked, after they had completed all the test tasks, to answer some questions related to the usability testing, and to write down their comments and thoughts; they were also asked to explain any particular reaction that was observed during the testing in a feedback questionnaire.

IV. DATA ANALYSIS AND MEASUREMENTS

In these experiments, two usability attributes were considered, which are efficiency and effectiveness. Also, additional measures were collected, such as usability problems encountered and users' comments and feedback. The results obtained were analysed as follows:

1. Time Spent: the time spent on each task in seconds was recorded by using a stopwatch. This time can be used to measure the efficiency of the website.
2. Usability Problems Encountered and their Severity: this is the most important result, as it can help the owner, developer and designer to improve their website. Once the problems are found, they are assessed in terms of their severity, as indicated in the literature review. Also, the performance of each method in terms of unique and overlapping problems and in the five areas\attributes will be identified.
3. Users' Comments and Feedback: at the end of each session, the users were asked to comment on the usability of the website. This aims to explain some of the questions raised during the experiment.
4. Efficiency: one measure of efficiency in UEMs is the "ratio between the number of usability problems detected to the total time spent on the inspection process" [32].
   Efficiency = (No. of problems)/(Average time spent) [32].
5. Thoroughness: this 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 (major and catastrophic) [27]. It can be calculated by the following formula:
   Thoroughness = (No. of real usability problems found)/(Total no. of real usability problems) [33].
6. Validity: this is the extent to which a UEM accurately identifies usability problems. It can be calculated by the following formula:
   Validity = (No. of real usability problems found)/(No. of issues identified as a usability problem) [33].
7. Effectiveness: this can be defined as the functional ability of the user interface [35]. It is calculated by the following formula: Effectiveness = Thoroughness × Validity [33]
8. Cost of Employ: the cost estimates can be done fairly simply by following Nielsen's equation: Cost estimate = (Number of evaluation hours) x (Estimate of the loaded hourly cost of the participants) [17]. He estimated the hourly loaded cost for professional staff at $100.

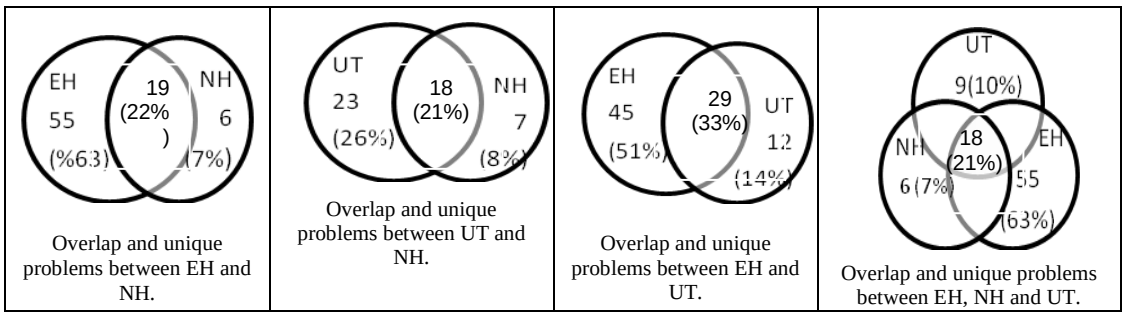
The above measures will be calculated for the user testing and compared against those calculated by the Educational Heuristics and Nielsen's heuristics in a separate paper [5]. This will allow us to determine whether our framework has generated a quality set of heuristics. To illustrate that, the results of the Educational Heuristics and Nielsen's heuristics comparison conducted in a previous paper will consequently change (in terms of total number

of usability problems found in all websites after adding the results of the user testing and this affects the results of Efficiency, Thoroughness, Validity and Effectiveness of NH and EH). The researchers conducted the debriefing session with independent evaluators to rank the severity problems for the user testing and to discuss the results of three methods (UT, EH and NH). After that, the results were separated into three lists of usability problems, in order to have all the problems for each one, together with their severity ratings. All the duplicated problems were removed and compared, in order to identify the unique and overlapping problems. A single list of usability problems was then consolidated from the three methods into one master problem list.

V. RESULTS AND DISCUSSION

Time spent: the average time taken in conducting the three experiments by using NH was 24.33 minutes, whereas for EH the average was 42.66 minutes. For UT, the average time was 139 minutes for 30 users. The group who used NH managed to evaluate the website more quickly than the other groups but discovered fewer usability problems, whereas the group that used EH spent slightly more time evaluating the website but discovered many more usability problems. The UT results are in line with the findings mentioned above in the literature review, i.e. UT is time-consuming.Explanations for the differences in time spent and number of problems located were gleaned from the evaluators' and users' feedback. They reported that NH was not particularly helpful, understandable or memorable for them. However, EH helped them to develop their skills in discovering usability problems in this application area; also, this set was more understandable and memorable during their evaluations and covered almost all broad areas. To further analyse these factors of time spent and number of problems discovered, efficiency metrics were applied.

Number and types of usability problems: the total number of problems discovered was 140, out of which 25 (28%) were identified using NH, 74 (84%) using EH and 41 (47%) using UT. When the problems from the three evaluation methods were consolidated, there were 52 duplicates; we identified a total of 88 problems. The total number of uniquely identified problems was 70. The EH identified 55 problems that were not identified by NH and UT, and there were 6 problems identified by NH that were not identified by EH and UT. Also, there were 9 problems identified by UT that were not identified by EH and NH. 18 problems out of 88 were discovered to be 'overlapping' between the three methods. Overall, a great many usability problems were discovered by EH. It was able to identify 7 catastrophic, 12 major, 20 minor and 16 cosmetic problems. This is because this method was developed in the context of the educational domain. NH did not find catastrophic problems, whereas UT uncovered 1; 1 major, 2 minor and 3 cosmetic problems were detected by NH. However, UT did not find major problems and it was able to detect 3minor and 5cosmetic problems. In this regard, NH, EH and UT each revealed different types of problem (both unique and overlapping); however, EH revealed the majority of usability problems, particularly those with high severity ratings, and EH appeared to work fruitfully for the expert evaluators, who then revealed more problems, both unique and overlapping, and discovered the most user testing problems because of the involvement of user inputs during the heuristics design stage. The usability testing revealed slightly more problems when it was combined with the Educational Heuristics, and more or less the same with Nielsen's heuristics (as depicted in Figure 1). This result was achieved because EH is appropriate for the particular characteristics of the educational domain. The result of the comparison between UT and NH confirms conducting usability testing with heuristic evaluation ; each one is complementary to the other, as recommended by [Nielsen, 1992]. On the other hand, EH as created through the proposed framework refutes this recommendation. The findings indicate that it is not essential to conduct usability testing in conjunction with heuristic evaluation (usually Nielsen's heuristics are used) in order to address the shortcomings of these heuristics; rather, to avoid wasting money, an alternative that is well-developed, context-specific and capable (such as the specifically generated Educational Heuristics) should be employed.



Overlap and unique problems between EH and NH.

Overlap and unique problems between UT and NH.

Overlap and unique problems between EH and UT.

Overlap and unique problems between EH, NH and UT.

Figure 1: Overlapping and unique problems among the methods

Usability problem areas/attributes: In sifting the literature review and consulting the expert evaluators (double

and single), the researchers found 5 key areas in the educational domain, which are User usability, Motivational factors, Content information and process orientation, Learning process and Design and media usability. EH helped the expert evaluators to identify large numbers of usability problems in all these usability areas on the three websites (74). However, NH overall worked slightly better, in discovering 25 usability problems related to three usability problem areas. However, 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. This is because NH is too general and does not cover all usability features for interactive systems [26]. Furthermore, UT worked better in discovering usability problems in three usability areas (41), but it failed to identify a sufficient n umber of usability problems in the Motivational factors and Content information and process orientation areas.

Efficiency, Thoroughness, Validity, Effectiveness and Cost of Employ: the set of formulae above were used and the results were that EH provided more efficient results than NH in discovering usability problems (EH = 1.7 vs. NH = 0.1). Also, UT provided more efficient results than NH but less than EH (UT = 0.3). That was because each group that used EH spent enough time (one hour, as estimated by Nielsen). This confirms what Nielsen stated in [29]: "the time spent on an evaluation can affect, to some extent, the number of usability problems found". The thoroughness of EH in identifying the number of real problems at Level 3 or 4 is higher than that of NH (EH = 0.75 vs. NH = 0.25). However, the thoroughness of UT is higher than NH, although less than EH (UT = 0.45). There is a slight difference between the validity of both sets of heuristics and user testing in accurately identifying usability problems, with UT being superior (EH = 0.89, NH = 0.88, UT = 0.97). For effectiveness, EH was slightly more effective in identifying usability problems than NH (0.89 and 0.22, respectively); UT scored 0.44. Finally, there was a slight difference in the cost of employ for the sets of heuristics and a greater difference with UT in this research; EH cost about $3,751, NH cost about $3,228 and UT cost about $6.400. This difference is due to the relative times consumed in the evaluations and analyses.

VI. CONCLUSION AND FUTURE WORK

Contrary to most of the efforts to construct and test enhanced sets of heuristics, our work has made explicit the process for so doing. The framework includes the views of users and usability experts to help generate new sets of heuristics for any chosen application domain. The work presented here illustrates and evaluates this process for the generation of a set of heuristics to assess the usability of educational websites. The Educational Heuristics out-performed both Nielsen heuristics and user testing, even when combined. This clearly represents a step in the right direction. Further validation of the use of our framework will indicate whether it is indeed applicable across domains.

VII. REFERENCES

1. Nielsen, J. and Molich, R. (1990). Heuristic evaluation of user interfaces, Proc. ACM CHI'90 (Seattle, WA, 1-5 April 1990), 249-256.
2. 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.
3. Oztekin, Z., J. Kong and O. Uysal (2010). UseLearn: A novel checklist and usability evaluation method for eLearning systems by criticality metric analysis. International Journal of Industrial Ergonomics, 40(4):455-469, 2010.
4. Nielsen, J. (1994), Usability engineering, Morgan Kaufmann.
5. AlRoobaea, R. Al-Badi, A., Mayhew, P. (2012). A Framework for Generating Domain-Specific Heuristics for Evaluating Online Educational Websites, 2nd International conference on Human Computer Interaction Learning Technology (ICHCILT 2013).
6. AlRoobaea, R. Al-Badi, A., Mayhew, P. (2012). Generating Domain-Specific Heuristics for Evaluating Social Network Websites, MASAUM International Conference on Information Technology (MIC-IT'13).
7. 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.
8. Ardito, C., Costabile, M., De Angeli, A. and Lanzilotti, R. (2006a). 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.
9. 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.
10. 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.

11. ISO (1998), ISO 9241-11: Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs): Part 11: Guidance on Usability.
12. Muir, A., Shield, L. and Kukulska-Hulme, A. (2003). The pyramid of usability: A framework for qualitycourse websites. In Proceedings of EDEN 12th Annual Conference of the European Distance Education Network, The Quality Dialogue: Integrating Quality Cultures in Flexible, Distance and e-Learning, Rhodes, Greece, pp. 15-18.
13. Smith-Atakan, S. (2006). Human-computer interaction. Thomson Learning Emea
14. 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
15. Dumas, J. and Redish, J. (1999). A practical guide to usability testing. Intellect Ltd.
16. 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.
17. Nielsen, J. (1994). Heuristic evaluation, Usability Inspection Methods, vol. 24, pp. 413.
18. 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, pp. 471-474. ACM.
19. 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.
20. Squires, D. and Preece, J. (1996). Usability and learning: evaluating the potential of educational software. Computers & Education, 27(1): 15-22.
21. Reeves, T. C., Benson, L., Elliott, D., Grant, M., Holschuh, D., Kim, B., Kim, H., Lauber, E. and Loh, S. (2002), Usability and instructional design heuristics for e-learning evaluation, In the proceeding of World Conference on Educational Multimedia, Hypermedia and Telecommunications, 1615-162.
22. Alsumait, A. and Al-Osaimi, A. (2009). Usability heuristics evaluation for child e-learning applications. In Proceedings of The 11th International Conference on Information Integration and Web-based Applications & Services, pp. 425-430. ACM.
23. Pinelle, D., Wong, N. and Stach, T. (2008). Heuristic evaluation for games: usability principles for video game design, In the proceeding of The twenty-sixth annual SIGCHI conference on human factors in computing systems, 1453-1462. New York, NY, USA,
24. 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
25. 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.
26. Tan, W., Liu, D. and Bishu, R. (2009). Web evaluation: Heuristic evaluation vs. user testing, International Journal of Industrial Ergonomics, vol. 39 (4), pp. 621-627.
27. 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.
28. Doubleday, A., Ryan, M., Springett, M. and Sutcliffe, A. (1997). A comparison of usability techniques for evaluating design. In Proceedings of the 2nd conference on De-signing interactive systems: processes, practices, methods, and techniques, pages 101-110. ACM.
29. Jeffries, R., Miller, J.R., Wharton, C. & Uyeda, K.M. (1991). User interface evaluation in the real world: A comparison of four techniques. Proceedings of ACMCHI'91, pp. 119-124. New York: ACM Press.
30. Nielsen, J. (1992). Finding usability problems through heuristic evaluation. In Proceedings ACM CHI'92 Conference (Monterey, CA, May 3-7), pp. 373-380. ACM.
31. Bernerus, A. and Zhang, J. (2010). A Peek at the Position of Pedagogical Aspects in Usability Evaluation of E-learning System - A Literature Review of Usability Evaluation of E-learning System conducted since 2000. Report/Department of Applied Information Technology 2010: 085.
32. Fernandez, A., Insfran, E. and Abrahão, S. (2011). Usability evaluation methods for the web: A systematic mapping study, Information and Software Technology.
33. Sears, A. (1997). Heuristic walkthroughs: Finding the problems without the noise, International Journal of Human-Computer Interaction, vol. 9 (3), pp. 213-234.
34. 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.
35. Masip, L., Granollers, T. and Oliva, M. (2011). A heuristic evaluation experiment to validate the new set of usability heuristics. In Proceedings of The 2011 Eighth International Conference on Information Technology: New Generations, pages 429-434. IEEE Computer Society.