Generating a Domain Specific Inspection Evaluation Method through an Adaptive Framework:

A Comparative Study on Social Networks

Roobaea AlRoobaea
Faculty of Computer Science and Information Technology Taif University, Saudi Arabia and School of Computing Sciences UEA, Norwich, UK

Ali H. Al-Badi
Department of Information Systems, Sultan Qaboos University, Oman

Pam J. Mayhew
School of Computing Sciences, UEA, Norwich, UK

***Abstract*—The electronic information revolution and the use of computers as an essential part of everyday life are now more widespread than ever before, as the Internet is exploited for the speedy transfer of data and business. Social networking sites (SNSs), such as LinkedIn, Ecademy and Google+ are growing in use worldwide, and they present popular business channels on the Internet. However, they need to be continuously evaluated and monitored to measure their levels of efficiency, effectiveness and user satisfaction, ultimately to improve quality. Nearly all previous 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, a new evaluation method is developed using traditional evaluations (HE and UT) in novel ways.**

The lack of an adaptive 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 adaptive framework that is readily capable of adaptation to any domain, and then evaluates it by generating an evaluation method for assessing and improving the usability of products in a particular domain. The evaluation method is called Domain Specific Inspection (DSI), and it is empirically, analytically and statistically tested by applying it on three websites in the social networks 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 our newly identified set of comprehensive usability problem areas as well as relevant usability evaluation method (UEM) metrics, with minimum input in terms of the cost and time usually spent on employing traditional usability evaluation methods (UEMs).

Keywords—Heuristic Evaluation (HE); User Testing (UT); Domain Specific Inspection (DSI); adaptive framework; social networks domain.

I. INTRODUCTION

Online social gatherings are known by a variety of names, such as online community and social network websites (SNSs). SNSs are increasingly recognized as one of the most popular mediums of online communication, and they are increasingly attracting the attention of academic and industry researchers intrigued by their affordability and reach. They have attracted millions of users around the world; many of them have integrated these sites to be part of their daily activities. Nowadays, SNSs play a vital role in many fields, such as e- government and business. They have gained in popularity not only because of their many interactive and innovative features, but also because their purpose is clearly established and the audience is targeted effectively [Pessagno, 2010]. Software development organizations are paying increased levels of attention to the usability of such social networking websites (SNSs); however, the majority of SNSs still have low levels of usability [Fu et al., 2008]. The motivation for this research is that SNSs are increasingly interesting as a topic of research in Information Systems, and so assessing and hence improving the usability of these SNSs is becoming crucial [Fox and Naidu, 2009]. Also, SNSs are becoming increasingly popular, andso it is important that the usability of individual sites be tested, assessed and improved in an objective manner as possible.

It is clear that Heuristic Evaluation (HE) and User Testing (UT) are the most important usability evaluation methods for ensuring system quality and usability [Chattratichart and Lindgaard, 2008; Chattratichart and Brodie, 2004]. Currently, the growth of a new breed of dynamic websites, complex computer systems, mobile devices and their applications have made usability evaluation methods even more important; however, usability differs from one website to another depending on website characteristics. It is clear that users have become the most important factor impacting on the success of a website; if a website 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) [Nielsen, 2001]. Nayebi et al. (2012) asserted, "Companies are endeavoring to understand both user and product, by investigating the interactions between them".

The traditional usability measures of effectiveness, efficiency and satisfaction are not adequate for the new contexts of use [Zaharias and Poylymenakou, 2009]. 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 [Hertzum and Jacobsen, 2001; Chattratichart and Lindgaard, 2008]. UT has been claimed to be costly, time consuming, prone to missing consistency problems and subject to environmental factors

[Oztekin et al., 2010]. 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) [Alias et al., 2013; Gutwin and Greenberg, 2000]; 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 [Coursaris and Kim, 2011].

The adaptive methodological framework in this paper was originally constructed and then evaluated practically in the educational domain; in this, it delivered interesting results by discovering more real usability problems in specific usability areas than HE or UT [Roobaea et al., 2013a]; [Roobaea et al., 2013b]. The main objective of this paper is to address the challenges that were raised and to continue testing the validity of the adaptive framework by applying it on the social networks domain, through three case studies. Furthermore, it is to conduct a comprehensive comparison between UT, HE and our domain-specific Inspection (DSI) method, which is developed through the adaptive framework, in terms of number of real usability problems found and their severity in each of a number of usability problem areas, as well as in terms of certain UEM metrics and other measurements. The paper is organized in the following way. Section 2 starts with a brief literature review relating to this study; it includes a definition of usability problems, and describes the concept of severity rating. Section 3 details the construction of the adaptive framework. Section 4 is on 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 on three cases, and then provides an analysis and discussion of the results. Section 7 presents a discussion of the findings. Section 8 presents the conclusion and future work.

A. 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.

II. Literature Review

SNSs have quickly become one of the most popular means of online communication in the last few years, and their users are dramatically growing in number by the day. SNSs can be defined as 'web-based services' that allow individuals to construct a public or private profile within a bounded system, and to explore connections with others within the system. They can be used to seek out new friendships or to group together to chat with friends, share activities or interests and extend one's own social network [Ellison et al., 2007; Fox and Naidu, 2009]. Most of the existing social networks on the Internet offer a range of services to users, such as instant messaging, private messages and e-mail, video and file sharing, blogging and playing online games, but they are also used by businesses, advertisers and employers, and those who wish to follow companies in order to receive information, updates and RSS feeds [Ellison et al., 2007; Estes et al., 2009]. It is now apparent that SNSs have had a significant impact on how individuals and social groups communicate and exchange information. These networks involve a great many users at any one time, and they are divided into broad categories according to purpose; there are networks for making new friends, for study and for work, in addition to networks based on interest or activity. The most well-known SNSs are Facebook, MySpace and Twitter, but others are emerging as this is an evolving field. Consequently, this is a productive environment for informational conflict between these websites, who seek to increase their number of users by attracting new users, attracting users from competitors' websites and maintaining current users [Tufekci, 2008]. Essentially, the success of SNSs depends to a large extent on the degree of users' contributions and activities, and so they need to be highly usable; if websites are not usable, users will leave and find others that better cater to their needs.

Emanating from the development of Web 2.0, there is now a real need to study the usability of SNSs [Ali et al., 2013]. The reviewed literature shows that the techniques for measuring the quality of user experience have been classified under the headings of ergonomics and ease-of- use, but more lately under the heading of usability [Oztekin et al., 2010]. 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 product and its user experience. A website is a product, and the quality of a product takes a significant amount of time and effort to develop. A high-quality website 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 an inordinate length of time learning how to use it; these are the fundamentals of the usability of a website. However, poor website 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 [ISO, 1998]. Nielsen (1994b) stated, "usability is associated with learnability, efficiency, memorability, errors and satisfaction" [Nielsen, 1994b]. 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 attributes covering four dimensions that influence the acceptance of a product, which are effectiveness, learnability, flexibility and attitude [Shackel and Richardson, 1991]. Nielsen (1994b) introduced five major attributes of usability based on a System Acceptability model [Nielsen, 1994b], 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

[Molich and Nielsen, 1990], 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 [Magoulas et al., 2003]. It can be used early in the development process, and may be used throughout the development process [Nielsen and Molich, 1990]. 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 [Holzinger, 2005; Nielsen and Loranger, 2006; Chattratichart and Lindgaard, 2008; Hertzum and Jacobsen, 2001].

TABLE I. HEURISTIC EVALUATION

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

There are two kinds of expert evaluators in HE. 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%). 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 [Nielsen, 1994a] 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 [Chen and Macredie, 2005]. 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, as HE does not otherwise address this important issue. Finally, the results of the evaluations are collected into a series of 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 [Nielsen, 1994a].

In the present context and in relation to HE, usability testing (UT; 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 [Nielsen, 1994b]. Dumas and Redish (1991) 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 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 [Van den Haak et al., 2004].

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 [Wilson, 2007]. Dumas and Redish (1999) suggested that the tasks could be selected from four different perspectives. 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 [Dumas & Redish, 1999]. 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 [Alshamari & Mayhew, 2008].

The result of applying HE and UT is a list of usability problems [Nielsen, 1994a]. 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 [Nielsen, 2000]. Furthermore, [Chattratichart and Brodie, 2002; Chattratichart and Lindgaard, 2008] 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 [Thompson and Kemp, 2009; Alrobai et al., 2012]. 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 [Liljegren and

Osvalder, 2004; Doubleday et al., 1997; Jeffries et al., 1991]. 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 [Ling and Salvendy, 2005; Hasan, 2009; Hart et al., 2008]. Nevertheless, 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 [Nielsen, 1992; Law and Hvannberg, 2002].

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 considers context of use and that includes 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 such a framework, i.e. for generating a context-specific method for evaluating 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 stages employed in the adaptive framework; also, it describes the process used to test it.

III. Construction of the Adaptive Framework

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

Development Stage One (D1: Familiarization): This stage starts from the desire to develop a method that is context-specific, 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 Stage Two (D2: User Input): This stage consists of mini-user testing (task scenarios, TA and a questionnaire). Users are asked to perform a set of tasks on a typical domain website, to 'think aloud' whilst so doing and then 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 stage in the framework directly benefits from including the advantages of user testing.

Development Stage 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 Stage Four (D4: Draw Up DSI Method: data analysis): The aim of this stage 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.

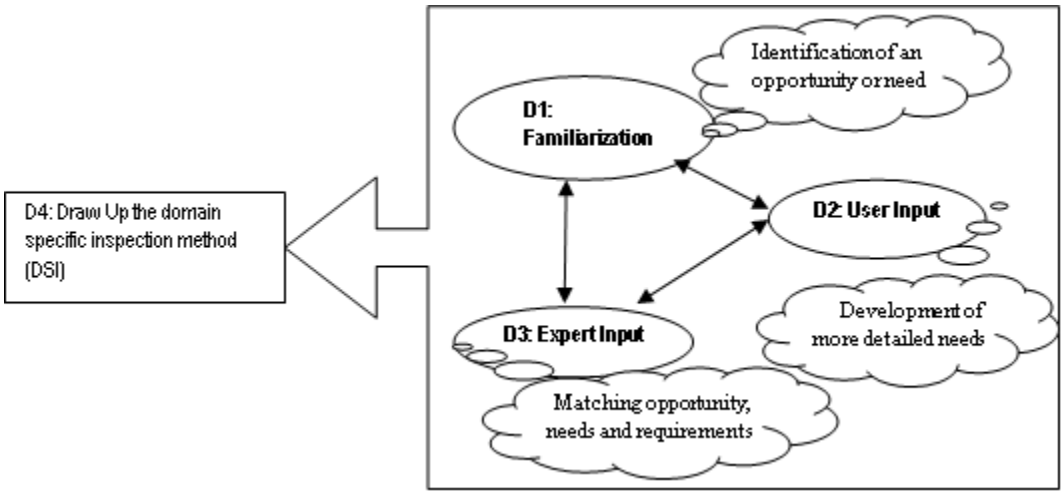

Fig.1. Development stages of the adaptive framework

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).

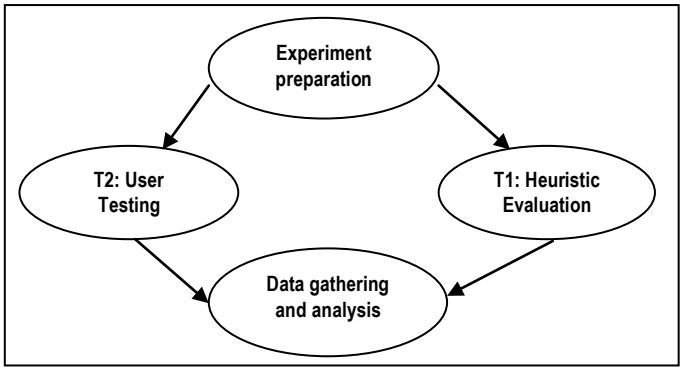

Fig.2. Testing stages of the adaptive framework

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

1) Experiment Preparation stage (for DSI, HE and UT): Before the actual evaluation formally starts, the following initial preparative steps 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. This initial experiment preparation stage is concluded with a pilot experiment to make sure that everything is in place and ready for the actual evaluation.

2) Heuristic Validation stage (Expert Evaluation (HE)): The aim of this stage 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 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 stage 4.

3) Testing Validation stage (User Evaluation (UT)): 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, 1992] 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 stage 4.

4) Data Analysis stage: This stage 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 steps; 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 step, 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 SNSs have recently been the subject of much study by researchers interested in areas such as privacy, identity, community dynamics and the behaviour of adolescents [Ellison et al., 2007]. However, it has not yet been fully explored, nor have any context-specific evaluation 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 SNS (i.e. one that is aesthetically attractive and is easy to use) can positively affect the number of people who become members. If these

are considered, an SNS will gain members more quickly because it will allow users to carry out social tasks more easily [Pessagno, 2010].

IV. 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:

A. 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.

B. Development; Evaluation of the Practicality of the Framework

In the first of the four stages within the development phase, the researchers conducted a literature review on the materials relating to usability and UEMs as well as on the requirements of the social networks 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 SNS users. In stage three, a focus group discussion session was conducted with six experts in usability and the social networks 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 0.9, 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 social networks as well as business networking websites, taking into an account what is called 'user experience'. The method was created and classified according to the usability problem areas detailed in Table 2 below.

C. Validation Stage 1; Preparation

a) Selection of the targeted websites

The first step in an initial preparation stage (of the validation phase) was 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; five aspects were determined and verified for each website, and these are: 1) Good interface design, 2) Rich functionality, 3) Good representatives of SNSs, 4) Not familiar to the users, and 5) No change will occur before and during the actual evaluation. In order to achieve a high level of quality in this research, the researchers chose three well-known websites in this domain, which are LinkedIn, Google+ and Ecademy. All of these have all the aspects mentioned above.

a) Experts and Users Recruitment

The selection of usability experts and users was the second important step in the initial preparation phase in this experiment. The researchers decided to recruit six expert evaluators, divided into two groups of three, who were carefully balanced in terms of experience. In each group, there are two double expert evaluators (usability specialists in SNSs) and one single expert evaluator (usability specialists in general). 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.

TABLE II. FINAL VERSION OF DOMAIN SPECIFIC INSPECTION (DSI)

Usability problem areas/attributes	Domain Specific Inspection (DSI)
Layout and formatting (LF)	Design consistency
	Simple user interface
Content quality (CQ)	Correct, relevant, reliable, error-free & up to date
	Site upload time & less memory utilization
	Representation with familiar terminology & understandable content
	Appropriate & approachable content
Security and privacy (SP)	Awareness of security mechanisms/settings & protection
	Transparency of transactions
Business support (BS)	Advertise or sales pitches mechanism
	Trust & credibility of information sources and company advertising
	Easy to follow & share
	Forum/blog facilities and connectivity with different groups/businesses
	Syndication of Web content (such as RSS tools)
	Frequent posting & updating
User usability, sociability and management activities (USM)	Manageable personal profile & user-driven content
	Easy functionality, participation & user privileges, such as revoke & join friends/connection
	Supports user's skills & freedom, such as customize/modify user's content/messaging and notification.
	Offers informative feedback - action & reaction
	Appropriate multimedia with complete user control
	Help & support
Accessibility and compatibility (AC)	Accessibility and compatibility of hardware devices
	Accessible path-contact details & site map
	Easy access through universal design
Navigation system and search quality (NS)	Correct & reliable navigation/directions
	Easy identification of links and menus
	Search support & functionality

Appropriate users will deliver results that are more reliable; they will also be intrinsically motivated to conduct the experiment [Dumas & Redish, 1999]. 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 [Nielsen & Loranger, 2006]. At this point, 30 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 10 users for each website. The majority of the users are students and employees, and they were mixed across the three users groups in terms of gender, age, education level and computer skills.

b) Task Sequencing

The third step was planning the sequence of the groups' evaluations. Each group employed two methods, namely DSI and HE, to evaluate the three different websites. The evaluations were carried out in a prescribed sequence, i.e. one group used DSI on Google+ and then HE on LinkedIn, and finally DSI on Ecademy, while the second group used HE on Google+ and Ecademy and then DSI on LinkedIn 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 set of heuristics, i.e. each evaluation was conducted with a fresh frame of mind.

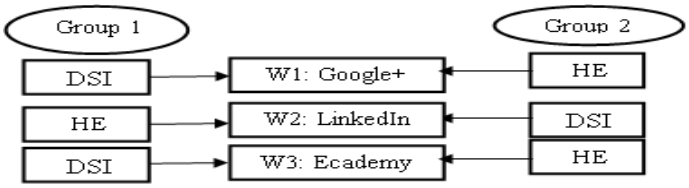

Fig.3. Usage of method sets by Group

c) Building the Experiment Documents

In the preparation phase, the fourth step entailed building a set of task-preparation documents for HE, DSI and UT, such as introduction sheets, HE and DSI checklists, tasks sheets, problem-ranking sheets, observer sheets, demographics, satisfaction and Likert questionnaires, sheets for collated problems 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 the checklists that had been developed by the researchers to

facilitate the evaluation process for DSI and HE. Users used the task sheets that were designed according to the main functions that users would normally expect to perform on the three 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 task for all three task groups, 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) to complete the five-point scale (1 for strongly disagree and 5 for strongly agree) to rate their satisfaction on the evaluation method they had used (DSI or HE). 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 [Brooke, 1996]. 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 [Nielsen and Loranger, 2006]. Evaluators and users were asked to fill in an open-ended 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.

d) Piloting the Experiment

The final step in the initial preparation stage was a pilot experiment. It was conducted by two independent evaluators and fifteen users. All materials were checked to make sure that there were no spellings or grammatical errors and no ambiguous words or phrases. 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 the equipment that the users regularly use and set it up for them before the test, for example, using the same type of machine and browser.

D. Validation Stage 2: Heuristic Evaluation

The heuristic validation stage 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, the SUS questionnaire and the Likert questionnaire and to give feedback on their own evaluation results.

E. Validation Stage 3: User Testing

The UT validation stage started with a training (familiarization) session for the 30 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.

F. Validation Stage 4: 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 and evaluators' attitude of HE and DSI by using System Usability Scale (SUS) and Likert Scale.

4) Reliability of HE and DSI: This can be measured from employing the 'evaluators' effect formula' (Any-Two-Agreement). It is used on each evaluator in order to measure their performance on an individual basis [Hertzum and Jacobsen, 2001].

5) Any-Two-Agreement = Average of |Pi∩Pj| / |Pi∪Pj| over all / 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.

6) 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" [Fernandez et al., 2011]. 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 [Liljegren, 2006]. Validity is the extent to which a UEM accurately identifies usability problems [Sears, 1997]. Effectiveness is defined as the ability of a UEM to identify usability problems related to the user interface [Khajouei et al., 2011]. The reliability of user testing can be measured by the mean number of evaluators to the number of real problems identified [Chattratichart and Lindgaard, 2008]. 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 [Nielsen, 1994]. 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 t-test, One-way ANOVA, Pearson's correlation, Mann- Whitney and Wilcoxon were chosen (at the 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.

V. Analysis and Discussions

This section describes the results obtained from using the three methods adopted in this experiment. It starts by detailing the results of the HE and DSI methods separately, including quantitative and qualitative analyses. This is followed by detailing the results 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.

A. Analysis for HE and DSI Results

1) Time spent: It is clear from Tables 3 and 4 below that the average time taken for doing the three experiments using HE was 56 minutes, whereas for DSI the average was 72 minutes. This difference in time spent between them is not significant (F = 0.139, p = 0.714) using the t-test. The group who used HE managed to evaluate the websites more quickly than the other group but discovered fewer usability problems, whereas, the group that used DSI spent slightly more time evaluating the websites, but discovered many more real usability problems. There was a statistically significant positive relationship between time spent and problems discovered through using Pearson's correlation test, where the 'Sig' value is 0.041 at the 0.05 level. This result reveals that the users who spent more time were able to discover more usability problems.

TABLE III. Average time taken and number of problems found during the evaluation by Group 1

Website	**Google+**	**LinkedIn**	**Ecademy**
Evaluator 'G1'	**Time**	**Time**	**Time**
1	90	70	80
2	60	50	60
3	70	60	75
Method	DSI	HE	DSI
# of problems	55	13	33
Mean time taken	73	60	72

TABLE IV. Average time taken and number of problems found during the evaluation by Group 2

Website	**Google+**	**LinkedIn**	**Ecademy**
Evaluator 'G1'	**Time**	**Time**	**Time**
1	60	80	60
2	50	70	50
3	40	65	60
Method	HE	DSI	HE
# of problems	22	47	12
Mean time taken	50	72	57

Explanations for the differences in time spent and number of problems located were 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 evaluations and covered most broad areas. 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 = 0.6 vs. HE = 0.4) as Table 5 shows. The t-test reveals significant difference in terms of efficiency between HE and DSI (t = -3.070, df =

11.391, p = 0.01).

TABLE V. MEAN SCORE OF EFFICIENCY FOR THE TWO METHODS

Method	Google+	LinkedIn	Ecademy	Mean
	Efficiency	Efficiency	Efficiency	
HE	0.5	0.29	0.3	0.4
DSI	1.1	0.8	0.8	0.6

2) Number of usability problems: Table 5 shows that HE was able to uncover 26% of the total number of real usability problems. However, DSI was able to uncover 74% of the total number of real usability problems in the websites.

TABLE 5: SUMMARY OF USABILITY PROBLEMS (NUMBERS AND PERCENTAGES) UNCOVERED BY EACH WEBSITE, EACH GROUP, EACH EVALUATOR AND EACH METHOD

Website	Group	Expert and type	Method	# of problems found by each evaluator	Total # of problems with repetition	Total # of problems without repetition	Total # of problems in each site with repetition	% of problems found by each evaluator	% # of problems found by each group
Google+	G1	Ev. 1^	DSI	16	66	55	77	21%	71%
		Ev. 2+	DSI	33				43%	
		Ev. 3+	DSI	17				63%	
	G 2	Ev. 1+	HE	6	22	22		8%	29%
		Ev. 2^	HE	5				7%	
		Ev. 3+	HE	11				14%	
LinkedIn	G1	Ev. 1^	HE	2	16	13	60	3%	22%
		Ev. 2+	HE	8				13%	
		Ev. 3+	HE	6				10%	
	G 2	Ev. 1+	DSI	24	59	47		40%	78%
		Ev. 2^	DSI	8				13%	
		Ev. 3+	DSI	27				45%	
Ecademy	G 1	Ev. 1^	DSI	6	57	33	45	14%	73%
		Ev. 2+	DSI	28				67%	
		Ev. 3+	DSI	23				55%	
	G 2	Ev. 1+	HE	5	12	12		11%	27%
		Ev. 2^	HE	3				7%	
		Ev. 3+	HE	4				9%	
Total number of usability problems discovered by each method						**Methods**		**Total number**	**%**
						HE		47	26%
						DSI		135	74%

(+) Double Expert (^) Single Expert (Ev.) Evaluator

TABLE VI. RESULTS OF T-TEST BETWEEN GROUPS AND METHODS IN EACH WEBSITE

Website	Group	Method	*t*-value	*df*-value	*p*-value
Google+	Group 1	HE	-5.524	5.448	0.045
	Group 2	DSI			
LinkedIn	Group 1	HE	-5.429	5.455	0.040
	Group 2	DSI			
Ecademy	Group 1	HE	-5.922	5.973	0.001
	Group 2	DSI			

In terms of the performance of each method in discovering unique and overlapping problems, Table 5 illustrated that the total number of real problems discovered was 182 in all three websites, out of which 47 were identified using HE and 135 using DSI. When the problems from the three evaluation groups were consolidated, there were 24 duplicates; we thus identified a total of 158 problems in all websites. The total for uniquely identified real problems in all websites was 128 problems.

The heuristic evaluation using DSI identified 96 real problems (61% of the 158 problems) that were not identified by HE, and there were 32 real problems (20% out of 158) identified by HE that were not identified by DSI. 30 real problems (19%) out of 158 were discovered by both methods (as depicted in Figure 4).

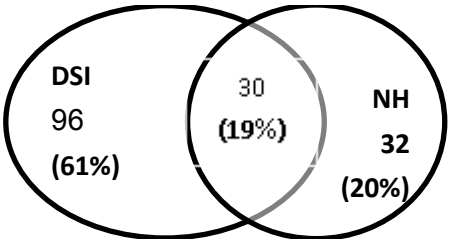

Fig.4. Overlap between both methods

In order to further compare the two methods, the severity ratings of the problems discovered (cosmetic, minor, major and catastrophic) were assessed, as Table 7 shows. Overall, a great many real usability problems were discovered across the rating scale. The most important results were obtained from using DSI, while HE found fewer (or no) usability problems. The Wilcoxon test revealed that there is a significant difference between the two methods in terms of problem severity (Cosmetic $z = -1.997$, $p = 0.04$; Minor $z = -2.207$, $p = 0.027$; Major $z = -2.003$, $p = 0.045$; Catastrophic $z = -2.100$, $p = 0.03$). This quantitative assessment between the two methods also entailed a comparison in terms of the usability problem areas in which the problems were found. These classifications/areas assisted in identifying how each method performed in each usability problem area or sub- heuristic. The six expert evaluators discussed, agreed and decided on the categories to which the problems should belong in both methods, as Tables 8 and 9 illustrate. The overall results from both tables show that the two groups (and the three websites) revealed more usability problems by using DSI than HE, particularly in User usability, sociability and management activities, business support (from assessing the DSI heuristics). The results show some level of agreement between the methods, regardless of the number of problems discovered in each usability area; both of them found the most problems in the area entitled , Navigation system and layout and formatting (DSI) or User control and freedom (HE). These were followed in DSI by Content quality, Navigation system and sear quality, Layout and formatting and Security and privacy; these areas were not discovered efficiently or sufficiently by the equivalent areas in HE. This suggests that HE is rather general and unlikely to encompass all the usability attributes of user experience and design.

TABLE VII. TOTAL NUMBER OF USABILITY PROBLEMS BY SEVERITY RATING FOR BOTH METHODS

Website	**Severity of Problems**	**Method**			
		DSI		HE	
Google+	Cosmetic	**Group 1**	16	**Group 2**	6
	Minor		28		13
	Major		11		3
	Catastrophic		0		0
	Severity (average)	1.9		1.9	
LinkedIn	Cosmetic	**Group 2**	11	**Group 1**	0
	Minor		19		8
	Major		11		5
	Catastrophic		6		0
	Severity (average)	2.3		2.3	
Ecademy	Cosmetic	**Group 1**	16	**Group 2**	4
	Minor		11		8
	Major		6		0
	Catastrophic		0		0
	Severity (average)	1.7		1.7	
Overall Severity (average)		2		2	
No. of discovered problems		135		47	

TABLE VIII. USABILITY PROBLEMS FOUND BY EACH HEURISTIC IN HE

Nielsen's Heuristics	**Google+**	**LinkedIn**	**Ecademy**
Visibility of system status	2	1	5
Match between the system and the real world	4	1	1
User control and freedom	5	3	1
Consistency and standards	3	2	0
Error prevention	0	1	0
Recognition rather than recall	2	3	1
Flexibility and efficiency of use	2	0	0
Aesthetic and minimalist design	1	0	0
Helps users recognize, diagnose and recover from errors	1	1	1
Help and documentation	2	1	3
Total problems	22	13	12

TABLE IX. USABILITY PROBLEMS FOUND BY CATEGORY THROUGH DSI

Usability problem area	Google+	LinkedIn	Ecademy
Layout and formatting	9	2	4
Content quality	12	11	5
Security and privacy	2	4	2
Business support	0	1	3
User usability, sociability and management activities	21	20	14
Accessibility and compatibility	1	1	3
Navigation system and search quality	10	8	2
Total problems	55	47	33

One striking result is that the number of problems identified by each evaluator who used HE was always fewer than the number of problems identified by any evaluator using DSI for the same website. An explanation of this was found in the evaluators' answers in the questionnaire. They said that the HE set was difficult to use and 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; it did indeed help them to remember all the functions that needed to be tested, it is specific and was designed to cover all the aspects needed for social networking websites.

3) UEM Performance Metrics:

After employing the above formulae and as depicted in Table 10, the Mann-Whitney test was used to investigate the statistical differences between the two methods in terms of the UEMs and reliability. The thoroughness of DSI in identifying the number of real problems was higher than for HE (0.3 vs. 0.1); also, Mann-Whitney revealed a significant difference between them ($z = -2.235$, $p = 0.025$). Further,

the validity of DSI was higher in accurately identifying real usability problems than HE (0.2 vs. 0.04); there was significant difference between them ($z = -2.600$, $p = 0.009$). The effectiveness of DSI was higher than that for HE (0.1 vs. 0.01); there was significant difference between them ($z = -2.230$, $p = 0.025$). Finally, the reliability values for DSI were slightly higher than for HE (0.76 vs. 0.64), and the results reveal that the difference between the two methods is significant ($z = -3.202$, $p = 0.001$). It can now be concluded that there is general agreement amongst the evaluators on the usability problems ($z = -3.202$, $p = 0.001$). Finally, the average results in terms of the cost of employing the two methods show that there is a slight difference in this research (Table 11); DSI = $863.33 vs. HE = $706.66.

TABLE X. MEAN SCORE OF UEM FOR TWO METHODS

Method	Google+		LinkedIn		Ecademy		Mean overall	
	HE	DSI	HE	DSI	HE	DSI	HE	DSI
Thoroughness	0.3	0.47	0.1	0.23	0	0.14	0.1	0.3
Validity	0.09	0.15	0.04	0.13	0	0.16	0.04	0.2
Effectiveness	0.03	0.07	0.004	0.03	0	0.03	0.01	0.1
Reliability	0.5	0.9	0.64	0.6	0.8	0.8	0.64	0.76

TABLE XI. COST OF EMPLOYING BOTH METHODS IN THIS RESEARCH

Mean cost	Ecademy	LinkedIn	Google+	Evaluation Method
$706.66	**$710** Time spent by 3 evaluators (2.8 hours) + 1 hour collecting data from the evaluation sessions + 3.3 hours analysing data.	**$730** Time spent by 3 evaluators (3 hours) + 1 hour collecting data from the evaluation sessions + 3.3 hours analysing data.	**$680** Time spent by 3 evaluators (2.5 hours) + 1 hour collecting data from the evaluation sessions + 3.3 hours analysing data.	Heuristic evaluation (HE)
$863.33	**$860** Time spent by 3 evaluators (3.5 hours) + 1.3 hours collecting data from the evaluation sessions + 3.8 hours analysing data.	**$860** Time spent by 3 evaluators (3.5 hours) + 1.3 hours collecting data from the evaluation sessions + 3.8 hours analysing data.	**$870** Time spent by 3 evaluators (3.6 hours) + 1.3 hours collecting data from the evaluation sessions + 3.8 hours analysing data.	Domain Specific Inspection (DSI)

4) Post- test questionnaire

- Satisfaction Score: The researchers used the System Usability Scale (SUS), and the results reveal that HE delivered a lower overall score, at 51, whereas DSI delivered slightly higher score, at 76. The evaluators gave this result because the process of the evaluation was smoother when using DSI and it was generated to cover all social network aspects.
- **Opinions and Attitudinal Questions** (Likert **scale)**

The Likert scores were collated for each statement in order to obtain overall results concerning the opinions of the expert evaluators with respect to 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. Cronbach's Alpha test was used to measure the reliability of responses and the result was 0.89. The Likert scores reveal that the evaluators were satisfied overall with DSI, and the results reveal significant differences between DSI and HE (using Mann-Whitney), as Table 12 shows.

TABLE XII. RESULTS OF MANN-WHITNEY FOR BOTH METHODS

Methods	Q1	Q2	Q3	Q4	Q5	Q6	Q7
Mann-Whitney U	1.500	6.000	5.000	6.500	4.500	3.000	0.000

B. Quantitative Analysis for Usability Testing Result

1) Time spent

Table 13 shows the time spent by each user on performing the experiment. The Google+ groups spent the longest time, more than the LinkedIn and Ecademy groups, with 429, 377 and 372 minutes, respectively. This 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 Google+ website, as some tasks were abandoned because the users had doubts about how to accomplish them. Also, in the LinkedIn website, the group spent time thinking about how to perform some tasks, such as the 'find group' task and the 'upload CV' task. The average time spent by each user in all three groups was more than 3.72 minutes. The efficiency formula used for UT for all the experiments, in terms of number of usability problems discovered over time spent, delivered a mean score 0.47 (Google+ = 0.64, LinkedIn = 0.46, Ecademy = 0.30). One-way ANOVA was used to determine whether there was a significant difference in terms of time spent, and the result reveals that indeed there was ($F = 15.033$, $p < 0.001$). Moreover, there was a statistical difference in terms of efficiency ($F = 24.694$, $p < 0.001$).

TABLE XIII. TIME TAKEN ON CONDUCTING THE EVALUATION

Usability measure	Google+	LinkedIn	Ecademy
Total time spent by all users (I n minutes)	429	377	372
Average time per user per task (in minutes)	4.29	3.77	3.72
Average time per user over ten tasks	42.9	37.7	37.2

2) User Satisfaction

It can be seen clearly that Ecademy delivered the highest overall score, at 6.5, whereas LinkedIn delivered the second highest score, at 4.9, and Google+ delivered the lowest score among the three websites, at 4.2. 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.

3) Number of usability problems discovered

Table 14 explains the total usability problems found by user testing and their severity rating. All the redundant problems are removed.

The usability problems detected in Google+ were 34, higher than in the LinkedIn and Ecademy websites (26 vs. 19). The One-way ANOVA test was used, and it delivered statistical differences amongst the number of problems ($F = 15.033$, $p < 0.001$). Pearson's correlation was used and the result reveals a positive relationship between time spent and problems discovered, with a 'Sig' value of 0.02. This result reveals that the users who spent more time were able to discover more usability problems.

4) UEM Performance Metrics

By applying the UEM and reliability formulae, Table 15 explains that the thoroughness of UT in identifying real usability problems was 0.23. The validity of UT in finding the known usability problems was 0.04. The effectiveness of UT in identifying usability problems related to the user interface was 0.03. The One-way ANOVA test was used to find significant differences between the websites (as a dependent factor). The results reveal that there are no significant differences ($p > 0.05$). The results for the cost of employing UT on each website were a little different with an average of $1,404, as shown Table 16.

TABLE XIV. NUMBER OF USABILITY PROBLEMS DISCOVERED

Problem type	Google+	LinkedIn	Ecademy	Total problems in all websites
	Total usability problems	Total usability problems	Total usability problems	
Catastrophic	4	2	0	6
Major	9	5	3	17
Minor	11	8	6	25
Cosmetic	10	11	11	32
No. of problems	34	26	19	79

TABLE XV. THE MEAN RESULT OF UEM METRICS

Metric / Websites	Google+	LinkedIn	Ecademy	Mean Total
Thoroughness	0.23	0.21	0.24	0.23
Validity	0.11	0.14	0.15	0.04
Effectiveness	0.02	0.03	0.032	0.03
Reliability	0.4	0.28	0.23	0.3

TABLE XVI. COST OF EMPLOYING UT IN THIS RESEARCH

Mean cost	Ecademy	LinkedIn	Google+	Evaluation Method
$1,404	**$1,370** Time spent by 10 users (6.2 hours) + 5 hours collecting data from the evaluation sessions + 2.5 hours analysing data	**$1,378** Time spent by 10 users (6.28 hours) + 5 hours collecting data from the evaluation sessions + 2.5 hours analysing data.	**$1,465** Time spent by 10 users (7.15 hours) + 5 hours collecting data from the evaluation sessions + 2.5 hours analysing data.	User Testing (UT)

VI. COMPARATIVE ANALYSIS TO EVALUATE THE ADAPTIVE FRAMEWORK

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

A. Types of problems found by UT in relation to DSI and HE

Two independent expert evaluators were involved in discussing, agreeing and deciding where the UT problems should be in HE, and to which category they should belong in DSI, as Tables 17 and 18 illustrate. The overall results from both tables show that all the UT problems were successfully classified into DSI. However, 30 problems out of 34 in Google+, and 12 problems out of 19 in Ecademy were successfully classified into HE. This proves that HE is rather general, and is unlikely to encompass all user problems, such as usability problems in the 'User usability, sociability and management activities', 'Business support', and 'Security and privacy' areas. Thus, this proves that DSI was able to discover user problems, and the unique problems that were discovered by UT and did not discovered by HE and DSI; were classed as missed problems for DSI and HE. The tasks given to the users during the usability testing seem to have 'walked them through' the activities, which could have increased the opportunity to discover problems. Furthermore, the findings confirm that 'Visibility of system status', 'Match between the system and the real world', 'Aesthetic and minimalist design' in HE, as well as the seven areas in DSI are a common weakness in dynamic websites (particular for SNSs). All three websites found nearly equal numbers of usability problems related to navigation and visibility. In conclusion, UT worked better than HE because 11 problems were not classified in it. However, all the users' problems were classified in the DSI.

TABLE XVII. USABILITY PROBLEMS FOUND COMPARED TO THE HE

Nielsen's Heuristics	Google+	LinkedIn	Ecademy
Visibility of system status	4	2	4
Match between the system and the real world	5	3	2
User control and freedom	3	2	0
Consistency and standards	1	1	2
Error prevention	2	3	0
Recognition rather than recall	2	1	1
Flexibility and efficiency of use	0	2	0
Aesthetic and minimalist design	6	1	2
Helps users recognize, diagnose and recover from errors	4	2	1
Help and documentation	3	1	0
Total problems	**30**	**19**	**12**

TABLE XVIII. USABILITY PROBLEMS FOUND COMPARED TO THE DSI

Usability problem area	Google+	LinkedIn	Ecademy
Layout and formatting	3	4	3
Content quality	7	6	2
Security and privacy	3	1	0
Business support	5	3	0
User usability, sociability and management activities	8	5	6
Accessibility and compatibility	2	0	0
Navigation system and search quality	6	7	8
Total problems	**34**	**26**	**19**

B. Performance of the Three Methods

Generally, Tables 19, 20 and 21 show how UT, HE and DSI revealed different types and numbers of usability problems. One-way ANOVA reveals that there is significant difference between three methods in terms of discovering usability problems on the whole ($F = 13.32$, $p < 0.001$). UT, HE and DSI revealed 47%, 31% and 75% of the usability problems found in Google+, respectively. One-way ANOVA-Tukey HSD was used and the results show that there is a strongly significant mean difference amongst the methods in finding usability problems in Google+ between HE and UT, where $p < 0.03$ and the mean difference = -14.667, as well as between DSI and HE, where $p < 0.003$ and mean difference = -16.767. In LinkedIn, UT, HE and DSI revealed 46%, 23% and 84% 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 LinkedIn, particular between HE and DSI ($p < 0.046$ and mean difference = -14.333) and between HE and UT ($p < 0.009$ and mean difference = -15.367). Finally, UT, HE and DSI revealed 50%, 32% and 87% of the found usability problems in Ecademy, respectively. One- way ANOVA-Tukey HSD was used and the results show that there is significant difference amongst the methods in finding usability problems in Ecademy between HE and DSI, where $p = 0.012$ and mean difference = -15.000. The performance of HE in discovering usability problems during the experiment ranged from 23% to 31%. UT discovered usability problems ranging from 40% to 47%, while DSI discovered usability problems ranging from 69% to 84%. 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. Thus, it can be seen that DSI was the best in discovering real problems; this was followed by UT, and then finally HE.

TABLE XIX. FINDINGS IN GOOGLE+

Method / Problem type	UT	HE	DSI	Total problems in the site from three methods (no duplicates)
Catastrophic	4 (100%)	0 (0%)	0 (0%)	4
Major	9 (82%)	3 (27%)	11 (100%)	11
Minor	11 (37%)	13 (43%)	28 (93%)	30
Cosmetic	10 (37%)	6 (22%)	16 (59%)	27
No. of problems	**34 (47%)**	**22 (31%)**	**55 (75%)**	**72**

TABLE XX. FINDINGS IN LINKEDIN

Method / Problem type	UT	HE	DSI	Total problems in the site from three methods (no duplicates)
Catastrophic	2 (33%)	0 (0%)	6 (100%)	6
Major	5 (39%)	5 (39%)	11 (85%)	13
Minor	8 (32%)	8 (32%)	19 (76%)	25
Cosmetic	11 (92%)	0 (0%)	11(92%)	12
No. of problems	**26 (46%)**	**13 (23%)**	**47 (84%)**	**56**

TABLE XXI. FINDING IN ECADEMY

Method / Problem type	UT	HE	DSI	Total problems in the site from three methods (no duplicates)
Catastrophic	0 (0%)	0 (0%)	0 (0%)	0
Major	3 (50%)	0 (0%)	6 (100%)	6
Minor	6 (50%)	8 (67%)	11 (92%)	12
Cosmetic	11 (37%)	4 (13%)	16 (53%)	30
No. of problems	**19 (40%)**	**12 (25%)**	**33 (69%)**	**48**

C. Overlapping and Unique Problems

Many researchers recommend conducting UT together with HE because they have found thateach method discovers unique problems [Nielsen, 1992], 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 22 shows the performance of the three methods on a unique performance basis for the three websites, illustrating the number of problems revealed by the UT but not identified by the HE and DSI and vice versa. DSI was able to discover 6 catastrophic, 24 major, 41 minor and 25 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 4 major, 19 minor and 9 cosmetic problems. UT was able to discover 6 catastrophic, 17 major, 25 minor and 32 cosmetic problems that were not revealed by the other 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 it also appeared to work fruitfully for the expert evaluators, who then revealed more real problems, both unique and overlapping.

For example, DSI found 41% uniquely of the total number of real usability problems (n = 73 out of 176). HE found 14% uniquely of the total number of real usability problems (n = 24 out of 176), and UT identified 32% uniquely of the total number of real usability problems (n = 56 out of 176). 23 (13%) real problems out of 176 were found to be 'overlapping' by the three methods. The clear superiority of DSI was due to involving user inputs in designing the method (as it is included in one stage of the adaptive framework), and due DSI being appropriate for the particular characteristics of the SNS domain.

TABLE XXII. SEVERITY PROBLEMS OF EACH METHOD'S PERFORMANCE, UNIQUELY AND WORKING IN PAIRS

Problem Types	HE (unique)	DSI (unique)	UT (unique)	HE & UT (overlapping)	DSI & UT (overlapping)	DSI & HE (overlapping)	Total number of problems in three websites (unique)
Catastrophic	0	6	6	0	8	0	10
Major	4	24	17	3	11	4	30
Minor	19	41	25	4	18	17	67
Cosmetic	9	25	32	5	21	9	69
Total	32	96	79	12	58	30	176

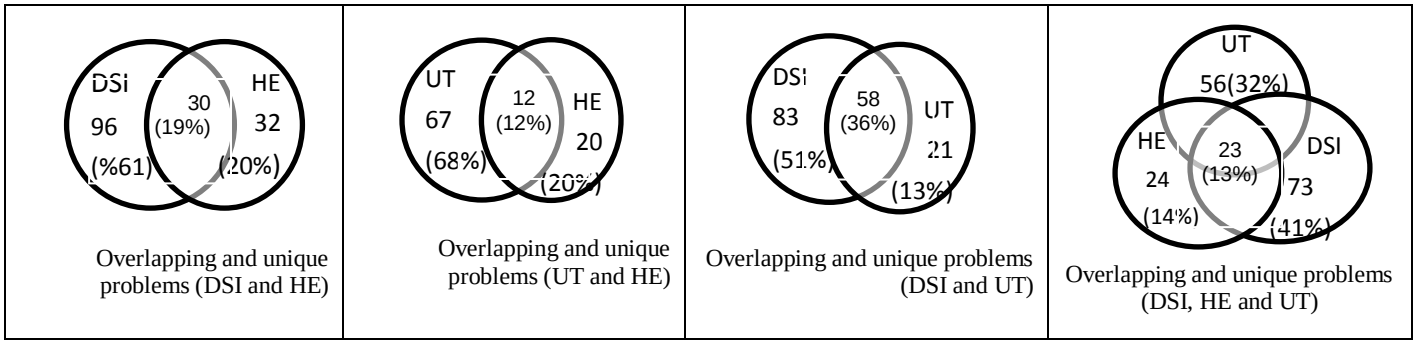

Fig.5. Overlapping and unique problems among the methods

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. In summary, the result of comparison between UT and HE confirms conducting UT with HE in order to overcome the shortcomings of each, because each one is complementary to the other. On the other hand, DSI (as created by the adaptive framework) refutes this recommendation.

D. Usability Problem Areas

It can be seen in Table 23 that DSI helped to identify large numbers of real usability problems in all usability areas on the three websites (135). However, HE overall worked slightly better in discovering 47 real usability problems related to just four usability problem areas, although it failed to expose any usability problems in three main usability problems areas, which are 'Security and privacy', 'Business support', and 'Accessibility and compatibility'. Furthermore, UT worked better in discovering usability problems (76) in three usability areas, but it failed to identify a sufficient number of usability problems in the 'Accessibility and compatibility' area.

TABLE XXIII. NUMBER OF USABILITY PROBLEM AREAS IDENTIFIED BY THREE METHODS

Usability Problem Areas	UT	DSI	HE
Layout and formatting	10	15	9
Content quality	12	28	4
Security and privacy	4	8	-
Business support	8	4	-
User usability, sociability and management activities	19	55	19
Accessibility and compatibility	2	5	-
Navigation system and search quality	21	20	15
Total Problems	76	135	47

E. Comparison between the Three Methods in UEM performance

It can be seen from Table 24 that DSI are more efficient, thorough and effective in terms of identifying the total number of real problems against total time spent, and in its relative ability to identify usability problems related to the user interface than the other methods. UT is the second best method, and HE is the last method. However, HE is the cheapest in terms of employment, and DSI is slightly more expensive than HE; both are cheaper than UT. One- way ANOVA reveals that there is significant difference among the methods used in terms of the UEM metric results, as shown in Table 25.

TABLE XXIV. COMPARING THE METRICS BETWEEN THE THREE METHODS

Metrics / Methods	Efficiency	Thoroughness	Validity	Effectiveness	Reliability	Cost
HE	0.4	0.1	0.04	0.01	0.6	**$706,66**
DSI	0.6	0.3	0.2	0.1	0.8	**$863,33**
UT	0.5	0.023	0.04	0.03	0.3	**$1,404**

TABLE XXV. ONE-WAY ANOVA RESULTS FOR THE THREE METHODS

Metrics	F	Sig. (*p*-value)
Efficiency	19.809	P< 0.001
Thoroughness	8.902	0.001
Validity	3.210	0.049
Effectiveness	3.367	0.48
Reliability	3.344	0.44

F. Advantages and Disadvantages of the Three Methods

We now assess the relative advantages and disadvantages of the three methods in evaluating user interfaces (see Table 26). 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.

TABLE XXVI. SUMMARY OF THE STUDY'S FINDINGS

Method	Advantages	Disadvantages
Usability Testing (UT)	* Helps define and achieve users' goals * Identifies the users' real problems * Identifies recurrent and general real problems	* Misses some severe real problems * High cost * takes more time * Conducting under lab condition
Heuristics Evaluation (HE)	* Identifies little 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 real users' problems and helps define and achieve users' goals	* A little higher in cost than HE and cheaper than UT. * Slightly higher in time than HE

VII. DISCUSSION AND FINDINGS

This section explores the results of this experiment and highlights the main findings. It thendraws out the lessons learned from the research. The main objective of this experiment was to evaluate the adaptive framework through its ability to generate a new method, specifically an inspection method designed for the social networks domain, by comparing its results with usability testing (UT) and Heuristic Evaluation (HE). It has been clearly shown that the hypotheses were accepted, and that Domain Specific Inspection (DSI) was able to find all the real problems that were discovered by UT and HE and more, 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 quality of the selected websites, which were represented in seven usability areas. 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 undermine their confidence 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 websites. 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 the one generated here for SNSs (or in another research on the

educational domain [Roobaea et al., 2013c], should be employed. Furthermore, the adaptive framework provided optimal results regarding the identification of comprehensive 'usability problem areas' on the SNSs, 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 problems given by [Cockton and Woolrych, 2002], we can consider the problems found by any one method and not found by the others as missed problems. From this standpoint, DSI missed discovering 80 real usability problems. However, HE and UT missed 129 and 97 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 websites. The selection of the method or methods will depend on the types of problem best identified by each of them.

VIII. Conclusions 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 and improve the usability of social network 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, which is specific for evaluating social network websites (Table 2). The third contribution is the identification of usability problem areas in the social network domain (seven areas in Table 2).

Acknowledgements

We thank the expert evaluators and users in the School of Computing Sciences at the University of East Anglia (UEA) and the Aviva company for their participation in the comparative study and the mini-usability testing.

References

[1] Ali H. Al-Badi, Michelle, O. Okam, Roobaea Al Roobaea and Pam J. Mayhew (2013), "Improving Usability of Social Networking Systems: A Case Study of LinkedIn," Journal of Internet Social Networking & Virtual Communities, Vol. 2013 (2013), Article ID 889433, DOI: 10.5171/2013.889433.

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

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

[8] 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.

[9] 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.

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

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

[12] 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.

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

[14] Ellison, N. et al. (2007). Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication, 13(1): 210-230.

[15] Estes, J., Schade, A. and Nielsen, J. (2009), 109 User Experience Guidelines for Improving Notifications, Messages, and Alerts Sent Through Social Networks and RSS, Accessed on 5/8/2012, Available at: [http://www.nngroup.com/reports/streams/].

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

[17] Fox, D. and Naidu, S. (2009). Usability evaluation of three social networking sites. Usability News, 11(1): 1-11.

[18] Fu, F., Liu, L., and Wang, L. (2008). Empirical analysis of online social networks in the age of web 2.0. Physica A: Statistical Mechanics and its Applications, 387(2-3):675-684.

[19] 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.

[20] 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.

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

[22] 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.

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

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

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

[26] 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.

[27] 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.

[28] Latchman, H., Salzmann, C., Gillet, D. and Bouzekri, H. (1999), Information technology enhanced learning in distance and conventional education, Education, IEEE Transactions on, vol. 42 (4), pp. 247-254.

[29] 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.

[30] 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.

[31] 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.

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

[33] Mack, R. and Nielsen, J. (1994). Usability inspection methods. edited book, John Wilely & Sons, Inc., ISBN 0-471-01877-5.

[34] 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.

[35] Mankoff, J., Dey, A., Hsieh, G., Kientz, J., Lederer, S. and Ames, M. (2003). Heuristic evaluation of ambient displays. In Proceedings of the SIGCHI conference on Human factors in computing systems, pages 169-176. ACM.

[36] 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.

[37] McCarthy, J. and Wright, P. (2004). Technology as experience. Interactions, 11(5):42-43.

[38] 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.

[39] Nielsen, J. (1992). Finding usability problems through heuristic evaluation. In Proceedings ACM CHI'92 Conference (Monterey, CA, May 3-7), pages 373-380. ACM.

[40] Nielsen, J. (1994), Heuristic evaluation, Usability Inspection Methods, vol. 24, pp. 413.

[41] Nielsen, J. (1994a), Usability engineering, Morgan Kaufmann.

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

[43] Nielsen, J. and Molich, R. (1990), Heuristic evaluation of user interfaces, Proc. ACM CHI'90 (Seattle, WA, 1-5 April 1990), 249-256.

[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] Pessagno, R. (2010), Design and usability of social networking web sites, BSc dissertation at California Polytechnic State University - San Luis Obispo.

[46] Pinelle, D., Wong, N. and Stach, T. (2008). Heuristic evaluation for games: usability principles for video game design. In Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems, pages 1453-1462. ACM.

[47] Preece, J. and Maloney-Krichmar, D. (2003). Online communities: focusing on sociability and usability. Handbook of Human-computer Interaction, pages 596-620.

[48] Rabiee, F., (2004), Focus-group interview and data analysis, Proceedings of the Nutrition Society, vol. 63 (4), pp.655.

[49] Roobaea AlRoobaea, Ali H. Al-Badi, Pam J. Mayhew. (2013a). A Framework for Generating Domain-Specific Heuristics for Evaluating Online Educational Websites, 2nd International conference on Human Computer Interaction & Learning Technology (ICHCILT 2013), MARCH 05-06, 2013, Abu Dhabi, United Arab Emirates (UAE).

[50] Roobaea AlRoobaea, Ali H. Al-Badi, Pam J. Mayhew.(2013b) .A Framework for Generating Domain-Specific Heuristics for Evaluating Online Educational Websites- Further Validation, 2nd International conference on Human Computer Interaction & Learning Technology (ICHCILT 2013), MARCH 05-06, 2013, Abu Dhabi, United Arab Emirates (UAE).

[51] Roto, V., Rantavuo, H. and Väänänen-Vainio-Mattila, K. (2009), Evaluating user experience of early product concepts, In the proceeding of DPPI, 2009, 199-208.

[52] Sears, A. (1997), Heuristic walkthroughs: Finding the problems without the noise, International Journal of Human-Computer Interaction, vol. 9 (3), pp. 213-234. Shackel, B. and Richardson, S. J. (1991), Human factors for informatics usability, Cambridge University Press.

[53] Smith-Atakan, S. (2006), Human-computer interaction. Thomson Learning Emea.

[54] Sutcliffe, A. and Gault, B. (2004), Heuristic evaluation of virtual reality applications. Interacting with Computers, 16(4): 831-849.

[55] 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.

[56] 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

[57] Special Interest Group on Human-Computer Interaction, pages 29-36. ACM.

[58] Tsui, K. M., Abu-Zahra, K., Casipe, R., M Sadoques, J. and Drury, J. L. (2009), A Process for Developing Specialized Heuristics: Case Study in Assistive Robotics, University of Massachusetts Lowell, Tech. Rep, vol. 11, pp. 2009.

[59] Tufekci, Z. (2008). Can you see me now? Audience and disclosure regulation in online social network sites. Bulletin of Science, Technology & Society, 28(1), 20-36.

[60] 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.

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

[62] 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.

[63] Roobaea AlRoobaea,Ali H. Al-Badi, Pam J. Mayhew,. (2013c). Generating a Domain Specific Inspection Evaluation Method through an Adaptive Framework: A Comparative Study on Educational Websites, International Journal of Human Computer Interaction (IJHCI) Volume 4, Issue 2.