Generating a Domain-Specific Inspection Method through an Adaptive Framework

Roobaea Salim Alrobaea

Supervisor:

Dr Pam Mayhew

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy at the University of East Anglia, School of Computing Sciences, Norwich, 2016

Abstract

Many recent innovations and inventions have contributed to rapid technological development, which in turn have produced a wide variety of products that have had a major impact on many businesses in several different domains. These products have their own contextual attributes that have made their usability evaluation, by using traditional usability evaluation methods (UEMs), all the more critical. Almost all previous usability studies have used the Heuristic Evaluation (HE) and User Testing (UT) methods; however, the majority of such studies have described these methods as being not directly applicable to the product being tested, not directly related to the context of the tested product, and not able to identify specific areas and types of usability problems. Furthermore, the lack of a methodological framework that can be used systematically to generate a domain-specific inspection method, which can then be used to assess the usability for a product in any chosen domain and to improve the usability assessment process, represents a missing area in usability testing.

Thus, the goal of this research is to generate a domain-specific inspection evaluation method that does not involve users in an actual testing session, i.e. one that is applied by only experts. To reach this goal, firstly, a systematic adaptive framework is presented, called Domain Specific Inspection (DSI), which is characterized as being pertinent to the context and specific target of a chosen domain. This framework is designed to generate a method that avoids the drawbacks of having to use both HE and UT, although it combines their advantages. In addition, this framework assists researchers as it combines feedback from both expert evaluators and potential users in the chosen domain in order to create a focused method. Secondly, this research seeks to validate the adaptive framework practically by generating a DSI method for assessing the usability of selected products. In this regard, websites are chosen as the targeted product, and two experiments are conducted; the first examines the utility of the generated DSI method on the educational domain. The second examines another generated DSI method on the social network domain. In both experiments, the DSI methods are tested intensively through rigorous validation methods and a number of usability metrics to verify the extent to which it achieves the identified goals, needs and requirements that the methods were originally developed to address, and to identify which problems are identified by UT but not identified by HE and/or DSI, and vice versa. Also, an investigation into whether it is essential to conduct the DSI method in conjunction with UT or HE will be undertaken. Furthermore, the roles and numbers of evaluators (together with their types) and users will be examined.

The results show that the adaptive framework is able to generate a DSI method that can be used to generate ideas from the different perspectives of multidisciplinary teams in order to create engaging user experiences and to facilitate interactive design. This method enables the discovery of a larger number of serious problems than UT and HE. In addition, it provides optimal results with regard to the identification of comprehensive usability problem areas, and it is more efficient and effective than UT and HE, with minimum input in terms of cost and time. Furthermore, it is able to improve the evaluator performance; thus, the results of the single evaluators, who used the DSI method, provided results that approached or outperformed the effectiveness of the double evaluators, who used HE. Consequently, few evaluators are needed to find a majority of the usability problems if DSI is used.

Acknowledgments

First and foremost, all thanks are due to God alone, not to any of the objects that are being worshipped instead of Him, nor to any of His creations. Through His mercy, I have been able to complete this thesis, and I beg His forgiveness for all the errors and omissions.

This research project would not have been possible without the support of many people. Firstly, I wish to express my sincere gratitude to my supervisor Dr. Pam Mayhew, who made this dream become a reality. She has immense knowledge, so she was abundantly helpful and offered invaluable assistance, support, encouragement, patience and intellectual guidance, and she provided me with an excellent atmosphere for conducting my PhD research. Thanks should also be extended to Prof. John Glauert for his constructive advice, thoughts and suggestions.

Secondly, I give a special thanks to my parents, sisters and brothers, who have supported me spiritually throughout my life, and who have always wished me success in my PhD research. In addition, many thanks go to my beloved wife Ala Alghamdi and my son Fisal; many thanks to them for their support, help and patience, without them I could not have completed this PhD research.

In addition, it is my pleasure to take this opportunity to thank the Saudi Government, Royal Embassy of Saudi Arabia Cultural Bureau and particularly Taif University for giving me this opportunity to develop my scientific knowledge. In addition, I would like to thank all the participants and evaluators in the experiments for their cooperation and willingness to spend time with me. Furthermore, I would like to express my sincere thanks to Dr. Ali Al-Badi for his consistent help, encouragement and supportive messages. I also would like to thank all journals and conferences for their feedbacks on published papers, and the website owners for granting me permission to use their websites in this PhD research.

Finally, I would like to express my appreciation to all my colleagues at the School of Computing in the University of East Anglia for their assistance.

Dedication

To my dearest parents

To my dearest family

To my dearest friends

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Background

Recent technological developments have opened the door to stiff competition amongst companies, forcing them to develop and produce new products on an annual basis. Nayak (1991, p.1) stated, "today, the pace at which companies introduce new technology has become a principal determinant of competitive success or failure". Competition has facilitated the production of a wide variety of products, and each product has specific design characteristics and context for use. Consequently, there is a need to develop a specific and applicable usability evaluation methods (UEMs) for evaluating each product based on its context of use and on its particular characteristics (Inostroza et al., 2012).

The most distinctive example of this rapid development in technology is the Internet revolution. The growth of the Internet and of ever-improving information technologies has enabled the development of a new breed of dynamic websites and applications that are growing rapidly in use and that have had a great impact on many businesses. These websites and applications should be developed in an interactive manner. Jiang (2009, p.101) defines Web usability as, "an application of usability in domains where Web browsing can be considered as a general metaphor for constructing the user interface". In fact, the primary concern of interaction design is to develop interactive products or technologies that are usable. On this point, Rogers et al. (2011) defined the meaning of interaction design as, "designing interactive products to support the way people communicate and interact in their everyday and working lives". This means that the products should be easy to learn, effective to use, and offer a pleasurable user experience. In this regard, a website is a product, and the quality of a product takes a substantial amount of time and effort to develop because if a product is produced and is then deemed not useful by the end-users, then it is classified as a failed product. Nielsen (2001) described this failure as, "nobody can use it and the company cannot make money". He studied the impact of poor usability on e-commerce websites, and said, "e-commerce sites lose almost half of their potential sales because users cannot use the

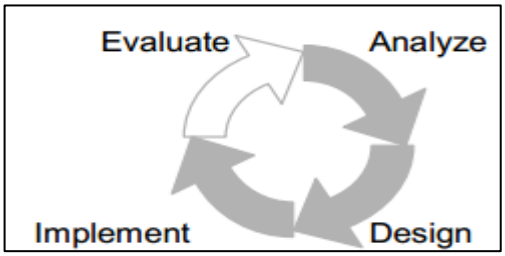
site". Also, he suggested that these sites have a great potential for having their usability improved, and that if that happens, the average site could increase its current sales by 79% (Nielsen, 2001). Later, Delone (2003) confirmed this by indicating that poor usability, usefulness or responsiveness can discourage the customer usage of an e-commerce system. Also, a poor commercial website can have serious consequences for the host company in a competitive environment (Oztekin et al., 2010);(Osterbauer et al., 1999);(Wild and Macredie, 2000).

In addition, website consultants and marketing specialists have understood that the number of hits, the customer return rate, and customer satisfaction are all directly affected by the usability of a website (Rogers et al., 2011). To avoid this scenario, Nayebi et al. (2012, p.1) assert, "companies are endeavouring to understand both user and product, by investigating the interactions between them". Consequently, a high-quality product is one that provides all the main functions in a clear format, and that offers good accessibility and a simple layout in order to avoid forcing users to spend more time on learning how to use it; these are the fundamentals of the 'usability' of a product. Poor product usability may have a negative impact on various aspects of the organization, and may not allow users to achieve their goals efficiently, effectively and with a sufficient degree of satisfaction (ISO, 1998a).

Web design is a key factor in determining the success of websites, and users should be the priority in the designers' eyes because usability problems in a website can have serious ramifications, over and above the users failing to meet their needs (Chen and Macredie, 2005). The term 'design' has been used in various different aspects, such as Web design, user centred design and product design; however, interaction design has been accepted as the umbrella term for all these aspects. Designing interactive products and evaluating them are common procedures and represent an important stage in the development lifecycle of a product, as shown in Figure 1.1. Capra (2006, p.1) argued that evaluation in this context entails, "identifying usability problems present in an interface that the designer should fix in the next design iteration". In this regard, there is an overlap between interaction design and Human Computer Interaction (HCI). The former is concerned with the theory and practice of designing user experiences for all manner of products, whereas the latter is concerned with the design, assessment and application of interactive computing systems for the benefit of end-users and the study of major phenomena surrounding them. The reason for the overlap

is that designers need to understand how users think, examine how they react to events, consider how they communicate and interact with each other, and comprehend how their emotions are affected in order to create effective user experiences and a highly usable interactive design (Rogers et al., 2011). Thus, user experience has become a central concept of interaction design and HCI. Garrett (2010, p.10) defines user experience thus, "every product that is used by someone has a user experience: newspapers, ketchup bottles, reclining armchairs, cardigan sweaters". This means that any reaction, feeling and impression about a product in terms of how good it is to use can be called user experience.

Figure 1. 1: Product development lifecycle (Capra, 2006)



In conclusion, usability is considered a critical quality aspect for websites, particularly for interactive ones. In this regard, quality assessment, and in particular usability evaluation is an important phase in the development of a website, which is often overlooked by modern Web application developers. Assessing the usability of a website by using the traditional usability methods (in a novel way or through developing a new method) has become necessary nowadays as the Web has developed gradually into a platform of complex applications that have increasing levels of interactivity, and into a front end of business databases and corporate information systems. In fact, the Web is now so complex that one person alone would not be able to have adequate knowledge of all of the different aspects of users in all of their diverse areas, and to use them in the process of interaction design; therefore, bringing together people with different types of knowledge and training is helpful in generating new evaluation methods, and in producing more creative designs (Rogers et al., 2011). From this standpoint, a methodological framework is proposed that is readily capable of adaptation to any domain, which can thus help anyone, including designers, programmers and engineers, to design a context-specific inspection method (DSI) in order to assess the quality of a product in a chosen domain. Also, it can be used in the initial development stages

to generate ideas from the different perspectives of multidisciplinary teams in order to create engaging user experiences and to facilitate interaction design.

1.2 Research process overview

The experimental approach was adopted as the most effective approach of achieving the research objectives outlined. Two experiments were conducted, each employing three usability evaluation methods to complement and compare, including: heuristic evaluation (HE), user testing (UT) with thinking out loud (TA), and domain specific inspection (DSI). Research entails a number of particular activities to strength and validate the findings, results and interpretations, for instance, a problem statement, a research question, definitions, a literature review, a sample of subjects, tests, a description of the methodology used (or other measuring instruments to collect data from the subjects), a description of the procedures to be undertaken, and a description of the intended data analyses (Fraenkel and Wallen, 2012). These activities have been adopted by the researcher here in order to develop the research plan, and the following sections and chapters explain the above in detail.

1.3 Definitions

This research is concerned with usability evaluation methods (UEMs) and their processes, with constructing a methodological adaptive framework for generating domain-specific inspection (DSI) method, and with validating it in practice; it is also concerned with the numbers of evaluators and users, and sets of usability measures. A framework is one source of inspiration and knowledge that can be used to conduct research, such as models and theories (Carroll, 2003). The meaning of ‘adaptive framework’ in this research indicates a flexible framework in which one is able to use its components in any domain by adapting those components in order to generate a new method for evaluating a specific product within a targeted domain.

The evaluation of a product, as it has been established, is at the heart of interaction design. This process aims to guarantee that the product is usable. The most common approaches to achieving this are through two kinds of UEMs, specifically designed to measure the usability of websites (as the chosen product in this study); they are inspection methods and testing methods. Heuristic evaluation (HE) is an inspection method that is guided by a set of general usability principles or ‘heuristics’ to identify usability problems. User testing (UT) is an

evaluation method that involves testing users' performance and assessing their satisfaction with the system in question via set of tasks in a laboratory (Rogers et al., 2011) However, both of these methods have advantages and drawbacks, and so it has been recommended that they be regarded as complementary.

A systematic adaptive framework is urgently needed in order to generate an alternative method, rather than continuing to use the current methods, to assess the usability of a product; also, this framework is needed to be applicable across different domains. This would make the evaluation process much easier for beginner and experienced programmers/designers to track particular steps in order to assess what they want. The numbers of users and evaluators indicates how many participants are needed by the HE, UT and DSI methods to perform an evaluation. This is a topic that has not been agreed upon yet, i.e. what are the most efficacious numbers of users and evaluators to achieve the desired results? Usability measures indicate the types of data that need to be collected in order to measure each of the three methods throughout the comparative evaluation process; they include efficiency, validity, number of usability problems, their relative severity, and so on.

1.4 Problem statement

A great deal of HCI research during the 1980s was conducted on how users interact with simple interfaces, such as dialog boxes and error messages; this was the focus of many researchers. In the beginning of the 1980s, the field of usability engineering methods (UEMs) appeared, along with the growth in graphical user interfaces. By the middle of 1990s, a major shift in HCI had occurred, following the wide acceptance of the Internet; thus there was a need to research new types of interfaces and technologies, such as web pages. Around 2004-2005, HCI research shifted more towards user-generated content that was shared, such as photos, videos, blogs and wikis. In 2006, research focused on collaboration, connections, emotions and communication. At that time, research did not focus on workplace efficiency; rather, it focused on whether a user liked an interface and wanted to use it. Every time there was a shift in the focus of research, there was a need to adapt or develop new research methods. In this evolving field, the traditional UEMs for ensuring system quality and usability, such as HE and UT, are in more demand than ever before; however, in the midst of the computer revolution, technological innovations, complex computer systems, mobile devices and their applications, usability now differs from one product to another depending

on product characteristics (Lazar et al., 2010). In a debate at an HCI conference, Greenberg and Buxton (2008) pointed out that the real world is complex and that innovation technologies are ever-changing; this means that it is now more difficult to conduct an evaluation with our classic UEMs. In the same debate, Cockton (2007) supported the above by saying, "the problem is not whether one should do evaluations, but that there is a lack of methods that are useful to various design stages".

Furthermore, the growth of the Internet has led to an explosion in dynamic website content, rising in accordance with demand, particularly after Web 2.0; for example, the e-learning and social networks domains. A variety of technologies has been developed for educational objectives, such as multi-media learning tools, mobile applications and digital content (Abuzaid, 2010) (Ardito et al., 2006). Nowadays, websites are essential for all universities that have a physical workplace. They all now have websites, and these have become an integrated part of their business, particular in their e-learning systems, such as the portal of the University of East Anglia (UEA). Developments within the Internet revolution (and related technologies) have led to the establishment of a large number of universities that exist solely online, i.e. without needing a physical workplace, such as the Open University in the UK. To keep pace with such developments, some companies and organizations are seeking to build free online learning websites that are oriented to world-class education for all educational levels, such Intel® Education and the BBC. This development in lifelong learning has made learners' intention to continue using e-learning an increasingly critical issue. However, some of these websites are difficult to use due to the inexperience of many of the designers and the lack of effective, efficient, accurate and appropriate guidelines for performing this task. Consequently, users spend more time learning how to use the website than learning the educational content, causing frustration, and leading to the abandonment of the site. Zaharias and Poylymenakou (2009) found that most e-learning programs were exhibit higher dropout rates when compared with traditional instructor-led courses. They found the poor usability of e-learning applications to be a major reason explaining the high dropout rates. Alkhattabi et al. (2010, p. 341) state, "quality is considered a crucial issue for education in general, and for e-learning in particular". Thus, there is a need for e-learning websites to be of sufficiently high quality. In terms of social network websites (SNSs), 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. It is now apparent that SNSs have had a major impact on how individuals and social groups communicate and exchange information. The impact of SNSs did not stop at this point; rather, they were used more imaginatively than anyone expected when people used them to change their governments, as happened during the revolutions of the Arab Spring. They are increasingly attracting the attention of academic and industry researchers intrigued by their affordability and reach. 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 (Fu et al., 2008). In conclusion, the majority of these websites still have low levels of usability, and some of the traditional evaluation methods are not applicable on them. Consequently, it is extremely important to develop a new method for addressing and assessing their quality, and this includes classifying suitable criteria for identifying usability problem areas for these websites (Fox and Naidu, 2009); (Stracke and Hildebrandt, 2007).

In addition, the spread of many modern technologies is now exceeding the traditional range of computers, creating many different types of interface, which in turn affect the user's experience; this change in interface boundaries is what one might call 'the end of interface stability', making old notions of the term 'interface' obsolete (Harper, 2008). Thus, the future of computing (for more people than ever) will reveal different emergent patterns of use and many different interfaces. Accordingly, there is a need to be better understand the extent to which these technologies and their interactive capabilities are 'usable' for end users and how they impact on user experience; this will entail determining how these technologies should be continuously evaluated and monitored to measure their efficiency, effectiveness and level of user satisfaction, and ultimately to improve their quality. These concerns create more challenges for the future of HCI, and the current literature emphasizes the importance of developing UEMs as a matter of priority, in order to increase their effectiveness and to identify the most acceptable approach to assessing such interactions (Hertzum, 2006).

To address these challenges, many frameworks and models have been published to update UEMs; for example, Gutwin and Greenberg (2000) proposed a conceptual framework to develop discount usability evaluation techniques for defining groupware usability for shared-workspaces. They gave as the reason for the development of the new method as being because traditional laboratory methods may deliver simplistic results that do not generalize

well to real-world situations. Furthermore, Zaharias and Poylymenakou (2009) developed a questionnaire method based on a UEM for e-learning applications, and relied upon a conceptual framework. This framework combined Web and instructional design parameters. Their developed method extends current practice by focusing not only on cognitive but also on any affective considerations that may influence e-learning usability. Additionally, Nayebi et al. (2013) developed a framework for evaluating the usability of Apple's iOS applications. The motivation for this development was that they believe that mobile devices and their operating systems have their own characteristics, and that these should be considered during any app design and development process, and later for usability evaluation.

The above frameworks and models, however, are not applicable to all domains because they were developed to deal with certain aspects of usability in certain areas. Also, some of the new methods or the extended methods have not been developed systematically. For example, some researchers state that they have developed their new method based on a literature review (Alsumait and Al-Osaimi, 2009), and some of them lack any information on how they arrived at developing their own methods (Cairns and Cox, 2008). Consequently, there is a need for an adaptive framework that outlines clearly the systematic steps for developing methodologies that can be tracked to generate a context-specific method.

Finally, Hollingsed and Novick (2007) asked a key question, which relates to why usability specialists practically rely upon single-perspective methods by recruiting expert evaluators or end users but not relying on both. Consequently, any new evaluation method should be adaptive so that it can be used to assess new technologies, and to more clearly understand the perspectives of both experts and end users. This research aims to combine these multiple viewpoints in a unified usability context inspection method through the systematic adaptive framework.

1.5 Research Scope

The growth of the Internet has led to an explosion in dynamic website content, rising in accordance with demand, particularly after Web 2.0. For this reason, the Web, and in particular websites, have been chosen as the target product in this research. The websites can be divided into two types: static and interactive/dynamic. The former does not allow engagement with users, whereas the latter is part of Web 2.0 which allows engagement with

users. There are numerous varieties of website domains which are classified based on their use or content. The following are examples of three classifications of these domains:

Table 1.1: Classifications of website domains

Type of domain	**Description**
Educational	There are two types of educational products in this domain. The first product is educational software such as Learning Management Systems (e.g. Blackboard and Moodle) (Machado and Tao, 2007). The second product is free educational websites such as Academic Earth, CosmoLearning, and iTunesU (Wikipedia, 2016). The free online educational websites are websites made for the purpose of education and they can be used to learn almost anything for free. They are utilised to help students improve their knowledge by watching lessons or interactive learning material, playing games, taking courses, assessment tests etc. (Cook and Dupras, 2004; Oren et al., 1998).
Social networking	Lenhart and Madden (2007, p.1) defined social network as "an online place where a user can create a profile and build a personal network that connects him or her to other users". These websites have some features such as sharing ideas, profiles, photo/video sharing, posts, music, blogs, activities, events, forums, searching, video/voice call, making friendships, private messages, groups, crowd sourcing, privacy and security, NewsFeed, mobile connectivity, and sharing interests with people in a personal network (Ellison, 2007; Thelwall, 2009). These websites can be classified into different types. For example, they can be classified as multimedia sharing, tagging and social bookmarking, RSS, blogs, audio blogging and podcasting, wikis and social networking (Shrivastava et al., 2011). Also, they can be classified into blogging, collaborative authoring, scheduling and meeting tools, microblogging, conferencing, image or video sharing, social networking, social tagging and bookmarking (Rowlands et al., 2011). The most popular and well-known social networking sites are Google+, Myspace, Digg, YouTube, Flickr, Twitter, Facebook, LinkedIn, Blogger, and Stumbleupon (Al-Badi, 2014).
E-commerce	Kalakota and Whinston (1997, p.3) defined e-commerce as "the buying and selling of information, products and services via computer networks". These websites have some features such as customer accounts, privacy and security, profile pictures, searching and sorting, managing categories and products, adding and deleting from basket, set messaging, reviewing and comment, view shipping and billing address, tracking and updating order status, discount codes and promotions, and payment gateway (Kalakota and Whinston, 1997; Huang and Benyoucef, 2013).

In addition, Web 2.0 technologies have led to the appearance of websites that adopt tools from different domains. For example, government websites adopt some social networking tools for supporting communication, interaction between citizens, reserving budgets, and improving services (Al-Badi, 2014). Also, e-commerce websites adopt some social network tools for communications with their customers (Mata and Quesada, 2014). Another example is that educational websites and Learning Management Systems (e.g. Blackboard) also adopt

some social network tools such as the discussion board (Brady et al., 2010). In this regard, these types of websites are not considered as social networking websites, and they are thus out of the scope of the social network domain because they do not fall within the definition of the social network domain, as mentioned in Table 1.1 by Lenhart and Madden (2007) and Shrivastava et al. (2011).

Furthermore, this research in terms of Web usability is related to others research domains such as web accessibility, user experience (UX), and human factor. First of all, Web accessibility is defined by Sierkowski (2002) as ‘’the ability for a person to understand and fully interact with a website’s content’’. Based on this definition, the accessibility feature will be considered in this research, however, this research will not consider the accessibility area in terms of removing barriers that prevent interaction with, or access to websites, by people with disabilities. This is out scope of this study because it needs specific tools and medical knowledge. In terms of user experience (UX), it is defined by Usability.gov (2016) as ‘’it focuses on having a deep understanding of users, what they need, what they value, their abilities, and also their limitations to meet the exact needs for the usage of a product or a service, without fuss or bother’’. This research will consider the users’ experiences and understand their requirements by involving them during developing the DSI method through Step Two (User Input) at the adaptive framework. With regarding to human factor are, it is defined by Wikipedia (2016) as ‘’it is the scientific discipline concerned with the understanding of interactions among humans and other elements of a system, and the profession that applies theory, principles, data and methods to design in order to optimize human well-being and overall system performance’’. This seems that the human factor is an umbrella term for numerous areas of research that include human performance, technology, design, and human-computer interaction (Human Factors Society, 2016). This study does not aimed to design a product or improve human performance. Thus, the human factor are is out scope of this study. This study will involve experts during developing the DSI method through Step Three (Expert Input) at the adaptive framework, and it will conduct consultation sessions with experts during choosing the domains and identifying their scoping. Due to the broad nature of this subject and how it relates to others research domains, the researcher excluded the accessibility area (in terms of removing barriers that prevent interaction with, or access to websites, by people with disabilities) and human factor area during the first step

and the fourth step of the adaptive framework. Also, the experts (during the focus group in the step three in the adaptive framework) excluded any related data to the research domains that have been excluded due to that they are out of scope this research. The user experience (UX) was addressed through involving real users in the step two in the adaptive framework. Then, the research conducted a context meeting with real users to meet the particular needs for the usage of a website without fuss or bother, and to design real tasks for the mini user testing in the step two of the adaptive framework as mentioned in Section 4.4.6.5. The researcher also observed the testing sessions with real users and took a notes. After that, the researcher discussed the results of user experience in the step two with experts during the focus group in the step three in the adaptive framework. For example, when scoping Web accessibility domain; the researcher took the definition as mentioned by Sierkowski (2002). In this regard, the literature related to Web accessibility guidelines and how make a website and web tools accessible were addressed. However, the literature related to Web accessible to persons with disabilities are not addressed. After that, the researcher conducted the context meeting with real users for the educational websites as example. During the meeting, the interview agenda was discussed with them, as shown in Appendix B10. Subsequently, the tasks were designed and then ten users were recruited, who had not any disability for performing the mini-user testing. The overall aim of this step of the mini-user testing was to obtain the Web accessibility requirements, and then to formulate focused heuristics to evaluate these requirements based on the users' experiences. Next, the focus group session was conducted to discuss all data obtained from the above process, including literature review, context meeting, and mini- user testing. The overall aim of this step was to formulate focused heuristics to evaluate Web accessible tools in the educational websites, and to identify the usability problem areas related to the Web accessible.

In conclusion, the domain of free online educational and social network websites is identified based on the above definitions by pioneers from both domains, as shown in Table 1.1. Consequently, a domain-specific inspection method (DSI) will be built to evaluate the free online educational websites such as Academic Earth (Wikipedia, 2016; Cook and Dupras, 2004; Oren et al., 1998), as shown in Table 1.1. Also, a DSI will be built to evaluate the social network websites such as Google+, Myspace, Digg, YouTube, Flickr, Twitter, Facebook, and LinkedIn (Al-Badi, 2014; Rowlands et al., 2011), as shown in Table 1.1.

1.6 Research aim and objectives

The overall aims of this research are to propose a systematic adaptive framework for generating an evaluation method for assessing the usability of a product, which is called Domain Specific Inspection (DSI), to test this framework practically by generating two DSI methods for two different domains in order to evaluate them against well-known evaluation methods in those chosen domains which are heuristic evaluation (HE) and user testing (UT), and to quantify the required number of evaluators and users to achieve good evaluation results for the DSI, HE and UT methods. These aims will be achieved through meeting the following objectives:

1) Review the current issues in usability evaluation methods (UEMs) on dynamic websites.
2) Construct an adaptive framework that will be used to generate the new method (DSI).
3) Test the practicality the efficiency of the adaptive framework by following its steps to generate the new evaluation method (DSI) for two domains, which are the educational and social network domains.
4) Validate the outcomes of the adaptive framework, which will entail analytically assessing the DSI method and also through empirical process; this will be achieved in each of the two domains through applying the three UEMS (HE, UT and DSI) on three websites in each domain.
5) Identify the usability problem areas for the educational and social network domains.
6) Explore the effect of sample size on the usability evaluation, and quantify the sample size required for usability for DSI, HE, and UT.
7) Explore further the correlation among UEM measurements in this study.
8) Proposes a set of recommendations and suggestions in order to improve the usability of the chosen domains.

Having extensively reviewed the existing literature on Web usability evaluation methods, the author can claim that, to the best of his knowledge, this research is unique in systematically constructing an adaptive framework that involves the advantages of both the heuristic evaluation (HE) and user testing (UT) methods but avoiding their drawbacks, and that this framework is applicable across numerous domains. This framework generates a domain-

specific inspection (DSI) method that does not need to be conducted together with any other method.

1.7 Significance of this study and motivation

From the Problem Statement section, it is clear that the motivations for the development of a new method reflect the fact that the traditional usability measures of effectiveness, efficiency and satisfaction are not adequate for the new contexts of use, such as dynamic websites, home technology, ubiquitous computing and technology supporting learning (Mankoff et al., 2003) (Zaharias and Poylymenakou, 2009);(Monk, 2002). These methods should assist practitioners in taking the correct decisions when designing the interactive aspects of a product. The traditional methods have been developed over many years to meet certain goals. However, Harper et al. (2008, p.54) argue, "a quite different mindset is needed for thinking about how to design for, how to control and how to interact with emerging ecosystems of technologies". Thus, HCI UEMs need to be improved so that they can be adapted for evaluating different interface types and technologies in order to achieve the objectives for which they were created. There is needed to build a framework that enable scientists, designers and users to share and communicate their expertise across disciplines, and to generate domain-specific inspection methods. Designing a DSI method that is effective will depend on understanding the nature of their expertise; this will be based on their qualifications, experience and demographics.

This research evolves through a clear series of steps, each one contributing to the advancement of knowledge in the HCI field. First of all, the current issues and publications is explored and is provided comprehensive analyses of the efforts of pioneers in this field, such as Jakob Nielsen, Sherry Chen, Robert Macredie, Gitte Lindgaard, Jarinee Chattratichart, Jeff Sauro and Kasper Hornbæk. Other valuable contributors are not overlooked and are included to offer further understanding of this field. The second contribution of this research is that it presents a methodological framework that is applicable across numerous domains (an adaptive framework) to generate a domain-specific inspection (DSI) method that can be used to assess the usability assessment process for an application in any chosen domain. This framework is derived from an extensive and in-depth review of the existing literature. It seeks to identify the users' requirements and to share the expertise

of usability experts in order to design the DSI method. The third contribution of this research is that it validates this framework and evaluates its practicality by building two DSI methods for two emerging domains, which are the educational and social network domains. The fourth contribution of this research is that it provides a comparative study by applying the two new DSI methods against two well-known UEMs in the chosen domains, in terms of a number of usability performance metrics, the number and severity of real problems discovered, their relative efficiency in discovering these problems in each usability problem area, the costs associated with employing them, and other usability measurements. These tasks are achieved through analytical and empirical processes, which are adopted in order to investigate the efficiency and practicality of the adaptive framework. The fifth contribution is that it identifies the effect of sample size (evaluators and users) on the evaluation results. The research findings should prove invaluable in enhancing evaluation methods, with particular reference to the adaptive framework and its generated method (DSI).

1.8 Research Question

Taking into consideration the significance of this study and the motivations, which were mentioned in Section 1.6, and the aim and objectives which were outlined in Section 1.5, this research addresses a main research question, which is ‘’Do the Domain-Specific Inspection (DSI) method, Heuristics Evaluation (HE) method and User Testing (UT) method differ in terms of time spent, cost of employment, numbers and types of usability problems detected, usability metrics, and usability problem areas?’’ This question obliged the researcher to conduct exploratory experimentation and in-depth analysis in order to identify answers.

1.9 Research methodology overview

The experimental approach is selected as the most effective method for achieving the research objectives mentioned above. Two domains are chosen, three websites in each domain are evaluated, in each experiment three evaluation methods are employed, which are heuristic evaluation (HE), user testing (UT) and the new method (DSI), in order to make comparisons between them. The research process as a whole could be used to strengthen these methods; it will also validate the findings, results and interpretations, as discussed in more detail in Chapters 4.

The first experiment investigates the usability of the educational domain. The steps of the adaptive framework are executed, and the DSI is accordingly constructed for this domain. Next, we apply the new DSI against the HE and UT methods on three different websites in the same domain. Eight expert evaluators (a mix of 'single' and 'double' evaluators) and 60 users are recruited to perform a series of tasks within each method. The aim of this experiment is to compare the results of three methods in terms of the number of real problems discovered (unique and overlapping), their severity ratings, the areas of any discovered problems, and identifying which problems might be discovered by HE and DSI and not discovered by UT, and vice versa. Also, the UEM performance metrics of each method will be measured, and other measures, such as cost, reliability and questionnaires, will be included. Moreover, this experiment seeks to prove or refute any recommendations associated with conducting UT with respect to the new method. Also, it incorporates statistical analyses. A more detailed discussion of this experiment can be found in Chapter 5.

The second experiment investigates the usability of the social network domain. The steps of proposed framework are again executed, and the DSI is accordingly constructed for this domain. Next, we apply the new DSI against the HE and UT methods on three different websites in this domain. Six expert evaluators (a mix of single and double) and 75 users are recruited to perform tasks within each method. The aim of this experiment is to compare the results of the three methods in terms of the number of problems discovered (unique and overlapping), their severity ratings, the areas of any discovered problems, and identifying which problems might be discovered by HE and DSI and not discovered by UT, and vice versa. Also, the UEM performance metrics of each method will be measured, and other measures, such as cost, reliability and questionnaires, will be included. Moreover, we again seek to prove or refute any recommendations associated with conducting UT with respect to the new method, and again it incorporates statistical analyses. A more detailed discussion of this experiment can be found in Chapter 5.

The findings of these two experiments will offer a number of recommendations and suggestions for usability engineers, Web designers and developers, and business managers in order that they may benefit from the knowledge gleaned from this research, in terms of the proposed framework and its generated context-specific method. This is discussed in more depth in Chapter 6 and Chapter 7.

1.10 Thesis structure

This PhD thesis is divided into seven chapters, and a brief overview of the main topics addressed in each is as follows.

Chapter 1: Introduction: this chapter starts with a brief introduction to the study, and includes sections on the background to the research, a discussion on the problems relating to this research field, the research aims and objectives, the research methodology, the significance and motivation of this research, and finally an overview of the thesis structure.

Chapter 2: Literature Review: this chapter is the foundation of this study; it contains the literature review on usability evaluation. It starts with a definition of usability and its attributes, and enumerates the usability evaluation categories. This chapter then presents, at length, the current issues raised in usability evaluation methods, as well as how to conduct heuristic evaluation and user testing generally. Finally, it explains how usability evaluation methods are used to examine website usability.

Chapter 3: The Proposed Adaptive Framework: this chapter discusses the adaptive framework for generating a domain-specific inspection (DSI) evaluation method. It also explains the components of this framework.

Chapter 4: Research Methodology: this chapter explores the set of research methods used in the HCI and IS fields, and identifies the most appropriate research methods to adopt in the present research. It also describes the procedures for preparing and conducting the research experiments, and for collecting and analysing the data. Moreover, it presents the development of the instruments, an overall research design for this study, and a framework design for the experiments.

Chapter 5: The First Experiment: this chapter presents details of the first research experiment. It describes the first chosen domain, which is the educational domain, then the justification for selecting it. Then, it describes how we apply the adaptive framework for generating a domain-specific inspection (DSI) evaluation method for the educational domain. After that, this chapter explains the approach taken to achieve the experiment's particular objectives, including the preparation and actual testing procedures involved, comparing the generated DSI method against the user testing and heuristic evaluation methods. Moreover, it discusses

the data analysis and concludes with recommendations and a description of the lessons that have been learned from this experiment.

The same chapter presents details of the second research experiment. It aims to be the second validation step for testing the ability of the adaptive framework to generate a DSI method. It describes the second chosen domain, which is the social network domain, then the justification for selecting it, and proceeds to employ the adaptive framework for generating a DSI evaluation method for the social network domain. After that, this chapter again explains the approach taken to achieve the experiment's particular objectives, including the preparation and actual testing procedures involved, comparing the generated DSI method against the UT and HE methods. Moreover, it discusses the data analysis and concludes with recommendations and a description of the lessons that have been learned from this experiment.

Chapter 6: Discussion and Recommendations: It contains comparisons and highlights the main findings obtained from the two experiments. Moreover, it outlines a set of recommendations for usability evaluation methods and for dynamic website usability.

Chapter 7: Conclusion: this chapter presents the conclusions of this study, and highlights the contributions that have been achieved. It includes an outline of the research findings, limitations, the personal benefits, and recommendations for further research.

Chapter 2: Literature Review

2.1 Introduction

Any company that desires to produce competitive products to sell in domestic or global markets must take into account a set of requirements and measures when developing that product, and must satisfy these in an effective and efficient manner. Quality must be the primary objective for such a company; quality has been defined as, "the totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs"(ISO, 1994). Over the past few decades, various methods to guarantee and improve product quality have been developed. Also, many quality attributes have been identified, and one of these attributes is usability. The main goal of this chapter is to include a clarification of the relationship between usability and Human Computer Interaction (HCI), definitions of the key terminologies, explanations of the usability evaluation methods (UEMs) that are more commonly used to evaluate user interfaces, and a description of the current issues in conducting usability evaluations by using UEMs.

2.2 Human computer interaction

Human Computer Interaction (HCI) is a multidisciplinary topic; it includes the study, planning, design and uses of the interactions between users and computers. It is a major topic in computer science and information systems but no definition for HCI has been agreed upon amongst researchers because it is multidisciplinary field. Thus, HCI is the study of the issues that arise when people encounter computer-based technology, and the way this understanding can aid in the design of technology that is better in various ways. Also, it is concerned with the design, implementation and evaluation of interactive products in order to support their use by the general public (Hooper and Dix, 2012) (Rogers et al., 2011). It has historically grown out of both computer science and psychology but, in addressing the full complexity of how people use computers, it has also grown to encompass social sciences and organizational theories, etc. These areas have their own traditional means for how to make a positive contribution to HCI knowledge. However, the range of currently published research

methods may not be most appropriate in providing a substantial contribution to the HCI area (Cairns and Cox, 2008). This research aspires to provide a useful contribution to HCI knowledge.

There are a number of goals for HCI; for example, ensuring the safety, utility, effectiveness, efficiency, accessibility and usability of such systems and websites. Moreover, HCI seeks to improve user interfaces by presenting ergonomic properties as well as developing and designing new interfaces, methodologies, frameworks or models that are related to evaluating a website and improving its quality in terms of usability attributes (Stephanidis, 2001); (Johnston et al., 2003). Consequently, a major part of HCI field is related to usability attributes.

2.3 Usability

The reviewed literature shows that the techniques for measuring the quality of the user experience have been classified under the heading 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 important aspects affecting the quality of a website and its user experience. In fact, the expression for describing the usefulness of a website design is usability. Consequently and in terms of usability, poor websites may have a negative impact on various aspects of public or private organizations, resulting in users being unable to accomplish their tasks or to achieve their goals efficiently, effectively, and with a high degree of satisfaction (ISO, 1998a). In the commercial world, usability is a necessary condition for survival, and it should have a role to play in each stage of the design process, i.e. from the initial design stage, passing through the prototype stage, and during and after the implementation stage (Blomkvist and Holmlid, 2011).

There are many definitions for the term usability. The International Organization for Standardization ISO (1998a) defines standard usability as, "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use". Nielsen (2012) indicated that ''usability is a quality attribute that assesses how easy user interfaces are to use''. Furthermore, it refers to methods for improving ease-of-use during the design process (Nielsen, 2012). 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.

2.3.1 Usability attribute measures

Usability attributes represent the most important factor in UEMs because any failure to measure these attributes can lead to failure in the evaluation process as a whole. Measuring these attributes delivers both quantitative and qualitative data. There are many and various usability attributes that should be taken into account and measured. Shackel and Richardson (1991) proposed four dimensions of attributes that influence the acceptance of a product, which are effectiveness, learnability, flexibility and attitude. According to Rogers et al. (2011), Utility and Usability are classified as sub-categories under Usefulness. The former is used to describe the extent to which the product provides the right kind of functionality to help users perform relevant tasks to do what they need or want to do, while the latter analyses the question of how well users can use that functionality (Rogers et al., 2011). Nielsen introduced six major attributes of usability based on a System Acceptability model (Nielsen, 1994c), and they are as listed below;

1) Learnability: a system should be easy to learn for the first time in terms of time and effort.
2) Efficient to use: the relationship between accuracy and time spent to perform a task. In other words, it refers to how a product supports people in accomplishing their tasks quickly.
3) Effectiveness: how well can a user achieve his/her goal using the system? This refers to the quality of the product and what it is supposed to do.
4) Easy to remember: a user should be able to use the system after a period of not using it without spending time having to learn it again.
5) Few errors: the system should prevent users from making errors; this also addresses how easy it is to recover from errors.
6) Subjectively pleasing: this addresses the user's feelings towards the system.

Furthermore, Table 2.1 describes the more commonly used measures and how they might be assessed as they were derived from 180 studies (Hornbæk, 2006). In conclusion, these

usability attribute metrics can be measured by employing UEMs to judge a product's overall usability (Hornbæk, 2006).

Table 2.1: Usability attributes of various standards or models (Hornbæk, 2006)

Memorability	Learnability	Satisfaction	Effectiveness	Efficiency
Retention over time	Time to learn	Comfort and acceptability of use	Speed of performance	Task completion
Rememberability	Easy to learn	Subjective	Rate of errors by user	Time in mode
		Attitude, Preference, Opinion	Accuracy, Task success rate	Input rate
		Open/closed questionnaire	Quality of outcome	Mental effort

2.3.2 Usability evaluation methods

Usability evaluation methods (UEMs) are a set of techniques that are used to measure the usability attributes, as mention above. Also, they are used to assess usability by identifying design usability problems. The aim of these methods is generally to discover as many usability problems as possible on the targeted system, searching for those that seriously effect user performance; some may remain within the system but they may be resolved in a later version (Lindgaard, 2006). Regrettably, some UEMs suffer from a variety of shortcomings and need enhancement. Notwithstanding this, UEMs are classified in different ways according to their goals, type of users involved, evaluation location and cognitive model (see Table 2.2).

Table 2.2: UEM classification

Method	User involved type	Physical location needed?	Process type
Inspection	Expert	No	Analytic
Testing	Real user	Laboratory	Empirical
Inquiry	Real user	No	Empirical
Model	No	No	Analytic
Software tools	No	No	Empirical

For example, they can be divided into formative methods and summative methods, they can be divided into inspection, testing and inquiry methods, they can be divided into analytical or empirical methods, they can be divided into intrinsic evaluation and pay-off evaluation methods, and finally they can be divided into expert evaluation, user evaluation and software

evaluation methods. Hartson et al. (2003, p.149) defined the formative and summative classification thus, "formative evaluation is evaluation done during development to improve a design, and summative evaluation is evaluation done after development to assess a design". Carroll et al. (1992, p.1) defined intrinsic and pay-off evaluation thus, "intrinsic evaluation is accomplished by way of an examination and analysis of the attributes of a design without actually putting the design to work, whereas pay-off evaluation is evaluation situated in observed usage". Thus, the best example for the former is heuristic evaluation, and user testing for the latter (Hartson et al., 2003). In this regard, Nielsen (1994d) summarized four basic ways for evaluating user interfaces. Firstly, they can be evaluated automatically by running a user interface specification through some program. Secondly, they can be evaluated empirically by testing the interface with real users. Next, they can be evaluated formally by using models and formulae to calculate usability measures. Finally, they can be evaluated informally based on rules of thumb and the general skill and experience of the evaluators.

2.3.2.1 Inspection (evaluator-based) UEMs

The objective of these methods is to identify usability problems to improve the usability of an interface design; they can be used early in the usability engineering lifecycle (Nielsen, 1992b); (Holzinger, 2005). Nielsen (1995c) defined the usability inspection method as, "the generic name for a set of methods, which are all based on having evaluators inspect a user interface, which aims to find usability problems and their severity for the design". They can be divided into two sub-categories: firstly, design principles, such as heuristic evaluation, and secondly, design task analysis, such as cognitive and pluralistic walkthrough (Nielsen, 1994c). Sears and Hess (1999) claimed that the inspection method saves time and money, and can identify a variety of usability problems, compared with usability testing. Lindgaard (2006, p. 1070) points out that "most authors tend to support the use of these methods, some very strongly, others with some reservation". However, inspection methods have two main drawbacks. Firstly, they focus on surface-oriented aspects, such as the graphical interface. Secondly, few of them address the usability of the application structure, such as content or navigation patterns. Finally, the reliability of the results is often entirely dependent on the individual expertise and skills of the inspectors (Triacca et al., 2004). This method includes, but is not limited to, the following methods.

- Heuristic evaluation (HE)

HE is the most popular inspection method. It was developed by (Nielsen and Molich, 1990), guided by a set of general usability principles or 'heuristics', as shown Table 2.3. 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., 2003b). It can be used early in the development process, and may be used throughout the development process. It can be conducted by three to five evaluators (Nielsen and Molich, 1990).

Table 2.3: Heuristics evaluation (HE) (Nielsen and Molich, 1990)

Nielsen's heuristics
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 is no specific procedure for performing heuristic evaluation. However, Nielsen (1995a) suggested a model procedure with four steps. Firstly, a pre-evaluation coordination session is very important; before the expert evaluators evaluate the targeted website, they should take a few minutes browsing the website to familiarize themselves with it. Also, they should take note of the actual time taken for familiarization. If the domain is not familiar to the evaluators, this session provides a good opportunity to present the domain. Also, it is recommended that in this session, the evaluators evaluate a website using the heuristics in order to make sure that the principles are appropriate (Chen and Macredie, 2005). Secondly, in the actual evaluation, each evaluator is expected to take around one to one and a half hours to list all the usability problems, and the actual time taken for the evaluation should always be noted. Next, a debriefing session should be conducted primarily in a brainstorming mode, and should focus on a discussion of possible redesigns to address the major usability problems and the 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 actual evaluation tables, and then combined into a single table after removing any redundant data. After the problems are combined, the evaluators should agree on the severity of each individual problem (Nielsen, 1995a).

This method has advantages and drawbacks. For example, it can work more efficiently (in terms of uncovering real usability problems) than other UEMs such as user testing (UT). Also, it is cheap and (more importantly for businesses) quick; evaluators do not need to attend a training session, and it is attractive to small start-up companies with low budgets (Jeffries and Desurvire, 1992); (Nielsen, 1994d); (Cockton and Woolrych, 2002). Cockton and Woolrych (2002, p.14) go on to explain how it cut costs by saying, "it reduces demands on the critical resources of time, facilities, cash, and skill". Also, Sauro (2012a) states, "one of the advantages of an expert review is that it uncovers issues that are harder to find in usability tests because users rarely visit enough parts of a website or software application outside of the assigned tasks". However, it is claimed to be a general, subjective assessment, and to require more evaluators than some other methods; also, it depends on the evaluators' experience (the 'evaluator effect'), does not guide evaluators during the evaluation, can produce a large number of false positives (which are not usability problems at all) and can miss some of the real problems (Molich and Dumas, 2008); (Chattratichart and Lindgaard, 2008). Cockton and Woolrych (2002, p.15) stated, "it does not encourage analysts to take a rich or comprehensive view of interaction, and it does little to support analysis of problem causes, leading to inappropriate solution generation". One example of the evaluator effect is that involving multiple evaluators can lead to very different lists of usability problems being generated on the same website, and the level of agreement between those evaluators over the lists could be quite low. It has been found, based on many studies on this topic, that the level of agreement between different evaluators can range from 5% to 65% (Holzinger, 2005);(Nielsen and Loranger, 2006);(Chattratichart and Lindgaard, 2008).

- Usability checklist

Usability checklist and heuristics evaluation are similar in terms of using evaluators. Another similarity is summarized by Johnson (1996, p.183) when they stated that " in terms of content, it is not surprising that there are large overlaps between guidelines, heuristics as used in evaluation, and the criteria and items that appear within the usability checklists. As they are generally derived from the same or similar sources, then we would naturally expect the same main criteria to appear''. In contrast, heuristic evaluation offers a general level of description and little detail, but a checklist offers more details which can help to discover different usability problems. Also, the checklist is easy to learn and apply and thus can be used effectively by non usability experts, and it help to improve the effectiveness of heuristic evaluation. In fact, the heuristics evaluation itself has been integrated into usability checklists in many studies, such as Alsumait and Al-Osaimi (2009), to facilitate these heuristics and formulate them in context of the tested system. Chen and Macredie (2005) described a heuristics checklist as fast, learnable and do not needed to have training before the actual evaluation. Many researchers developed a usability checklist when they evaluated a certain product to facilitate the evaluation process such as eLearning systems, academic library websites, and mobile phone user interface (Oztekin et al., 2010); (Johnson, 1996); (Raward, 2001). In this research, the checklist for the Domain-Specific Inspection (DSI) heuristics will be used to support the use of the DSI method and to facilitate the evaluation process. It is different from DSI heuristics. DSI heuristics are general heuristics with their explanation, whereas their checklist includes most elements of the chosen domain in order to provide a wide range of evaluation of websites in the chosen domain.

- Cognitive walkthrough (CW)

CW was proposed by Lewis et al. (1990). Hollingsed and Novick (2007, p.250) define it thus, "it is a usability inspection method that evaluates the design of a user interface for its ease of exploratory learning, based on a cognitive model of learning and use". It links the interface walkthrough with a cognitive model as the evaluator explores each aspect of the interactive system to identify usability problems, focusing on how easy it is for new users to accomplish tasks with the system. It uses a more explicitly detailed procedure to simulate a user's problem-solving process at each step through the dialogue,

checking if the simulated user's goals and memory content can be assumed to lead to the next correct action (Mahatody et al., 2010);(Nielsen, 1995c). Polson et al. (1992) outlined the process of CW, which comprises a preparatory phase and an analysis phase. In the former, the evaluators check the interface to be used, the task scenarios, and the actions (or 'walk through') to be taken on an interface during each task. In the latter, the evaluators work through the four steps, which are: firstly, the user sets a goal to be completed within the system. Secondly, the user determines the currently available actions. Thirdly, the user selects the actions that they think will take them closer to their goal. Fourthly, the user performs the actions and evaluates the feedback given by the system (Polson et al., 1992). Indeed, this method has advantages and disadvantages. For instance, it helps find mismatches between the users' and the designers' conceptualization of a task, and it is inexpensive and fast; however, it needs extensive knowledge of cognitive psychology and technical details, it includes the possible danger of an inherent bias due to improper task selection, there is an emphasis on low-level details and non-involvement of the end user (Hwang and Salvendy, 2010);(Bernsen and Dybkjær, 2009);(Holzinger, 2005).

- Pluralistic walkthrough

Bias (1994) defined pluralistic walkthrough as, "a usability evaluation method that brings representative users and system designers together into a design session to evaluate each element of interaction based on their expertise". He defined five characteristics for this method, which are: firstly, this method needs users, system designers and usability experts in the same walkthrough session. Secondly, the interface screens are presented in the same order as they would appear in the system to the user. Thirdly, all members take the role of a user. Fourthly, the members note down the actions they would take to perform the given tasks before the group discusses the screens. Finally, the group discusses the solutions they have reached. The administrator first presents a correct answer. Then the users describe their solutions, and only after that, do the designers and usability experts offer their opinions (Bias, 1994);(Riihiaho, 2002). The efficacy of this method has been measured against some UEMs, such as user testing. The results show that pluralistic walkthrough can provide more reliable data on a particular user interface than UT. Also, it is better at revealing uncertain decisions than

UT 'lucky guesses' (Hollingsed and Novick, 2007). Furthermore, this method has other benefits and limitations. For example, PW delivers responses from users with little effort, even if the interface is not fully developed, it enables rapid iteration of the design cycle (redesign), and it focuses on users' tasks. In contrast, it is not easy to group all the users at once, and then it works at the speed of the slowest, and the approach must be limited to representative rather than comprehensive user paths through the interface (Rogers et al., 2011); (Hollingsed and Novick, 2007).

2.3.2.2 Testing (user-based) UEMs

There are various testing (user-based) techniques, which can be used to find usability problems, as discussed below.

- Usability testing

Usability testing (also known as user testing, moderated testing or UT), 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 are 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, 1994c). Dumas and Redish (1999) stressed that a fruitful usability testing session needs careful planning and attention to detail. Accordingly, there is a general procedure for conducting UT, 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 (Dumas and Redish, 1999). There are two types of UT. The first is for problem discovery and it is called formative user testing. This aims to discover the usability problems in the targeted product and fix them. The second type is for benchmarking and it is called summative user testing. This aims to verify the

usability of the targeted product based on the efficiency and effectiveness of real users in completing the tasks that they perform during the test session (Sauro, 2011b).

In addition, the UT method has many advantages. For example, it takes place in controlled environment. Also, the conditions for conducting the experiment can be controlled. Additionally, all the users experience the same setting, leading to higher quality data (Trivedi and Khanum, 2012). However, it can cost a great deal and be time consuming for users and the test facilitator, particular when a large set of users is involved in a series of experiments (Sauro, 2009);(Skov and Stage, 2005). Dykstra (1993) stated that '' unfortunately, usability testing is not always a feasible alternative. Facilities for testing may not be available. It may be too early in the development cycle to test, or a software designer may need feedback sooner than is possible with usability testing''. Also, there are various factors affecting UT and its results. These factors include the users' varying characteristics ('user profiling'), the formulation of the task or set of tasks, the test environment, and others. The users' characteristics are important and user profiling should be considered from all angles. Users differ in terms of age, nationality, background, gender and, crucially, computing skills. Each one of these factors may play a substantial role in UT. Molich et al. (2004, p.73) stated, "usability testing effectiveness is dependent on tasks, methodology and users' characteristics". Sauro (2010) recommends testing with actual users, for whom the product was originally designed, and also testing with users that may perform the evaluation in different ways, such as different types of user within an organization. Task design is an important factor in the design of adequate Web usability tests. The tasks designed for Web UT 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. Also, they suggested that the tasks 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 and Redish, 1999). Sauro (2010) points out that the task number should be a minimum of 3 to 5 tasks.

Also, he advises selecting representative tasks, i.e. those that are typical of the tasks that the users would implement on an interface; this is to maintain data accuracy and validity. Also, tasks should be designed in terms of the functionality and features of the chosen website, taking into account the time for each task. Furthermore, it may be necessary to break down the tasks into smaller segments, particularly for complex activities (Hanna et al., 1997). There are several concerns regarding the task scenarios, for instance, task coverage, task number, task selection and formulation, and task order. Wilson (2007) warns of the risks of selecting inappropriate tasks, which can lead to complaints about the product and/or usability testing. Also, he clarified that one of the reasons why unsuccessful tasks are proffered to users is because they are designed without considering the real users of the product.

In terms of testing environment, UT takes place in a controlled laboratory. Tullis et al. (2002) found several cases where the product had worked fine in the laboratory, but not in the real word. They discovered that the conditions under which the product's use had been tested were different to the conditions for actual use. Nayebi et al. (2012, p.2) justified this when he said, "isolating users from environmental factors that can affect usability may cause differences in the user experience, and the effect of environmental factors prevalent in the real world may not be felt". Wolf et al. (1989) and Dix (2009) listed four aspects when seeking to understand why a laboratory experiment sometimes fails: 1) The users' motivation can be greatly diminished or destroyed by the atmosphere of a controlled laboratory; 2) A laboratory does not take into account the social context (that supports and motivates the users if they need it); 3) A laboratory setting does not consider the time context, where, in reality, users may leave their work and resume it later; and 4) A laboratory does not take into account the user's work context (users may feel disinclined to invest time and effort in something that they see as someone else's job). However, conducting experiments in a laboratory can increase their validity, can facilitate system comparisons, and can offer a controlled area where all interactions with the system can be closely recorded and monitored (Wolf et al., 1989); (Nielsen, 2005) (Feng et al., 2010).

- Think-aloud (TA) protocol

There are various techniques that can be used to supplement UT, and the 'think-aloud' protocol is the most widely used. It is employed during the test when the users are asked to think out loud whilst performing their tasks, and in this, it is important to record the users' thoughts, feelings, and opinions. This technique can effectively help evaluators to capture how users interact with an interface and what is happening on the screen (Rubin and Chisnell, 2008). It has been claimed that one-third of 'severe' usability problems can be discovered through this technique (Ebling and John, 2000). However, the setting of the usability test can sometimes influence the effectiveness of the 'think-aloud' protocol, and it does not always help when the users are not in their natural surroundings; this means that users may not feel as at ease and may feel unable to talk or express their thoughts and ideas freely in a restricted and unfamiliar laboratory environment (Van den Haak et al., 2004). Furthermore, Rubin and Chisnell (2008) suggested that if the tasks are designed to assess the efficiency of a system (i.e. measuring time spent on tasks), then TA should be avoided, as it may negatively impact on the performance of the users. TA has been generally used to achieve three types of goal; firstly, to find evidence for models and theories of cognitive processes; secondly, to discover and understand general patterns of behaviour in the interaction with documents or applications, in order to create a scientific basis for designing a new product or service; and thirdly, to test specific new documents or applications in order to troubleshoot and revise (Krahmer and Ummelen, 2004).

The observer does not know what the users are thinking while they are performing their tasks, so this is a big problem with observation methods. TA is a useful way of understanding what is going on in a user's head (Rogers et al., 2011). There are three types of TA, which are concurrent, retrospective and 'constructive interaction'. The concurrent TA type is the most common; this involves participants verbalizing 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, verbalizing their thoughts

through interacting (Van den Haak et al., 2004). On the one hand, it has been argued that TA should be avoided in certain circumstances, as mentioned above, but on the other hand, Albert and Tullis (2013) assessed the degree to which it can actually influence users' performance, as they concluded that this technique, in fact, can enhance performance because it helps users to focus more. However, some researchers, when employing the concurrent type, have expressed concerns about reactivity, i.e. the possibility that the act of speaking concurrently may influence user performance through distracting their attention and concentration; the effort to fully verbalize the steps taken in the task may change the ways that users attend to the task components. For this reason, the retrospective TA type was proposed to avoid the problems of concurrent TA; it is assumed to be the most fruitful in terms of problems reported per participant (Van den Haak et al., 2004). Furthermore, Co-Discovery Learning (constructive interaction) has been claimed to be the most suitable method for evaluating collaborative systems, and to be the most appropriate method for usability testing with children (Felder and Silverman, 1988).

- Remote testing

Remote testing or un-moderated testing is a distinct method within UT; it occurs when users are separated physically in space and/or time from their evaluators during the testing period. This method is the only choice when users and evaluators are located far from each other. Also, it is more appropriate for large sample sizes. Some software tools are available for facilitating observations, such as WebEx, Morae, CU-SeeMe and certain Open Source Software (OSS) developments (Andreasen et al., 2007);(Castillo et al., 1998). This method can be generally classified into two main categories: synchronous remote usability testing (moderated), and asynchronous remote usability testing (unmoderated) (Dray and Siegel, 2004). This method has been developed over many years, so there are various different approaches, but generally five main types have been frequently referred to in the literature, namely, instrumented remote evaluation, user-reported critical incident, remote questionnaire or survey, third-party services, and workflow logging (Petrie et al., 2006).

- Eye-tracking

This method provides information in real time and to a high level of detail on users' visual patterns whilst using an interface. Thus, this method gives researchers an insight into how users think with a deep understanding of what users ignore. This method is helpful in many areas and more especially in cognition, such as in assessing attention. This method is an integration of two main behaviours: first, fixations, where the eye is relatively still; and second, saccades, where the eye moves rapidly between fixations. This method is claimed to be quite complex in terms of its process and interpretation, and needs a strong study design to answer research goals (Granka et al., 2004);(Salvucci and Goldberg, 2000);(Cairns and Cox, 2008).

- Software (tools-based) UEMs

Many software tools are available to reduce the human workload and to assess whether an interface conforms to a set of specific usability guidelines, e.g. automatic usability evaluation tools, and transaction log file and web analytics tools such as Google analytics. These tools differ from one another in terms of aim, type of use, collected data, and where installed (Jain et al., 2012). These methods can be used with other methods such as HE, UT, and remote testing for recording the behaviour of evaluators or users.

2.3.2.3 Model-based UEMs

Over the last thirty years, the rapid developments in computers and complex systems have presented substantial challenges to interface designers. For this, analysts have utilized the theories and methods of cognitive psychology to construct cognitive models (Cairns and Cox, 2008). These models take several forms to explain and predict user behaviour. For example, Card et al. (1983) proposed a task analysis method which was called GOMS (Goals, Operators, Methods and Selection rules). This method was used to predict certain aspects of human performance, i.e. how to perform a task with an interface from the first time of use, from four elements of analysis. Subsequently, the GOMS method was developed into several analysis techniques, such as the Keystroke-level Model (KLM), Cognitive-Perceptual-Motor GOMS (CPM-GOMS), and The National Institute of Standards and Technology (NIST) Web Metrics (John and Kieras, 1996). On the other hand, these methods generate numbers (i.e.

quantitative data) and these numbers may not correlate to the actual website usability (Zaphiris and Kurniawan, 2007).

2.3.2.4 Inquiry UEMs

Inquiry methods gather subjective inputs from participants through observing and asking them, and they are likely to be combined with other methods. These methods are most commonly used after a usability evaluation session. The results of these methods can be used as benchmarks for comparing the currently tested design with future design iterations (Sauro, 2011c). The following are representative of the inquiry techniques.

- Focus groups

  These are most commonly used with qualitative approaches in marketing, political campaigning, and social sciences research. Krueger and Casey (2000, p.6) defined a focus group as, "a carefully planned discussion designed to obtain perceptions on a defined area of interest in a permissive, non-threatening environment". Rogers et al. (2011) described it thus, "it assumes that individuals develop opinions within a social context by talking with others". This is an informal technique that allows the researcher to explore the considered judgments of a few participants (normally 3 to 10), who meet for a period of around an hour and a half to two hours, in great depth and to learn something about how end-users think about an interface. A focus group allows each participant to put forward their own opinions in an encouraging environment and to ask questions of each other, but each one must also hear from the other participants and is then encouraged to comment on what has been said, prompting others to offer more clarification on the subject in question. A focus group can highlight the users' spontaneous reactions, comments and suggestions through their interaction. It involves recruiting a number of users/experts, who are a representative sample of the target population. The advantage of this method is that it explores and seeks to understand the views and attitudes of the participants in an efficient manner. On other hand, it relies on the ability of participants to raise issues for discussion, which means that they must have a significant amount of information and the willingness to interact, there is less experimental control, and it needs an environment that encourages conversation (Kitzinger, 1994). A moderator or the researcher should prepare the system usability

issues together with an agenda to guide the discussion on the topic of interest, ensuring that the participants are able to contribute fully to the developing discussion (Cairns and Cox, 2008); (Rogers et al., 2011); (Freeman, 2006).

- Interviews

This technique is also most commonly used with qualitative approaches; it is used, for example, in case studies and in action research. An interview involves asking subjects for their opinions within a limited timeframe. In other words, it comprises a list of (usually) open questions designed to encourage the respondents to deliver pertinent information; the questions should be prepared by the evaluator but, preferably, two evaluators should be engaged in the process, one to ask the questions and the other to note down the interviewee's responses. Interviews can be unstructured (or open-ended), semi-structured, or structured (Fontana and Frey, 1994);(Myers and Newman, 2007). The following is explanation of these three types.

i. Unstructured interviews

These entail a set of open questions that are posed by the interviewer through conversation on a particular topic to obtain in-depth information and to explore a range of opinions without expectation about the format or content of the answers. The one advantages of this method is that it generates rich information that can offer a deeper understanding of the topic at hand. However, they need a considerable amount of time to analyse the data, which are sometimes complex and interrelated; this adds to the overall costs of the research. Also, the interviewer cannot repeat the process (Rogers et al., 2011).

ii. Structured interviews

This method is similar to the questionnaire method in terms of directly asking predetermined questions. A structured interview is a complete script without need for improvisation, which is often used in surveys, and the interview is not necessarily conducted by the researcher. This method can be used when the research goals are clearly understood. The questions should be clear and short, and repeated in the same

order with each interviewee (Rogers et al., 2011); (Newman, 1998); (Myers and Newman, 2007).

iii. Semi-structured interviews

This method integrates features of both previous methods. They are a mix of open and closed questions, more flexibly worded, and they are modelled more closely on the unstructured than the structured. They are broadly replicable and the interviewee should be given time to speak and not move on too quickly. They help to increase the reliability and validity of a research (Klenke, 2008); (Rogers et al., 2011); (Hersen et al., 2011).

- Questionnaires

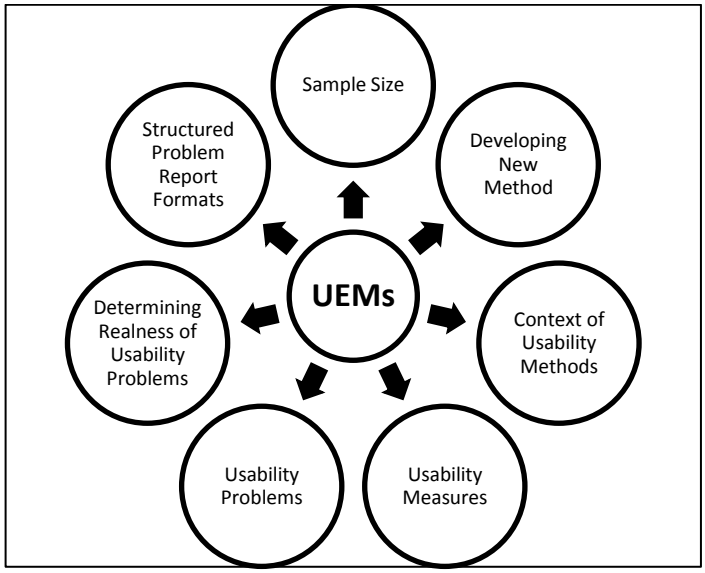
This is a technique that uses a set of written closed or open questions for gathering demographic data and opinions from volunteers in order to measure interface usability, as an example. In other words, they are indirect usability measures that collect data on issues such as user satisfaction, user preferences, user attitudes and others. There are many types of question format, such as rating scales, check boxes and ranges, and postal and email questionnaires. These techniques are mostly conducted after employing UEMs such as the UT, TA and HE methods, seeing them as complementary, or they could be used alone at the same time as indirect usability methods. The questions should be written clearly and the layout should be constructed to encourage the respondents to complete them and to avoid any ambiguities that may result in failure to achieve the desired goals. Questionnaires have advantages and drawbacks. For example, they can be used to collect data from a large sample across a wide geographical area; they are cost-effective and quick, particularly when compared with other more resource-intensive methods such as observation and interviews. Nonetheless, they are time consuming and need a sufficient number of responses to obtain significant results; thus, interviews are more in-depth and flexible than questionnaires. Also, designing negative questions can lead the users into giving false data (Cairns and Cox, 2008);(Holzinger, 2005); (Dumas and Redish, 1999); (Zhu and Liao, 2007); (Hornbæk, 2006); (Cho and Park, 2001); (Gillham, 2008); (Tuckman and Harper, 2012); (Cairns and Cox, 2008).

It can be concluded that the above UEMs have advantages and disadvantages. The rapid evolution of Information Technology (IT), especially with the coming of the Internet and other issues has challenged these methods. So, there is need for an effective and appropriate methodology for evaluating the emerging domains/technologies to measure their levels of efficiency, effectiveness and satisfaction, and ultimately to improve their quality. In order to develop new UEMs, certain factors should be considered carefully, as the present research confirms their impact on usability evaluation results. The following section sheds light on these issues.

2.4 Current issues in usability evaluation methods

The above UEMs have many issues that have impacted on their efficiency levels and results over the years. These issues are summarised in Figure 2.1, and more information is presented in the following sections.

Figure 2.1: Set of issues impacting on usability evaluation results



2.4.1 Sample size

A crucial aspect when planning UEM sessions is establishing the sampling size, as each method entails certain implementation costs; there is needed to balance costs against benefits. Throughout the literature, one of the most frequently asked questions in the usability field, and one whose answer is very important for developers, designers, market researchers and

usability practitioners and experts, is "how many users are really enough?" This question has challenged researchers and professionals in the field of usability engineering and Human Computer Interaction (HCI) because it has consequences for any evaluation results. Although this issue has been hotly debated amongst researchers for many years, there is no consensus on any rule that could be relied upon to determine this number because all usability practitioners seems to have a different opinion. This is a major challenge in HCI because no one can know in advance how many problems exist. Thus any estimation of how many participants are required to find a certain percentage of interface problems is based on an assumption (Lazar et al., 2010).

There are thorny issues making an unequivocal answer to this question almost impossible; for instance, the aim of researcher's study, the size of the project, the accuracy of the uncovered usability problems, and the design and scope of the tasks (Lewis, 2006). For example, if the aim of an evaluation is to identify the major usability problems in a small part of a system, the researcher may recruit a small sample. Should the researcher wish to evaluate the whole system with many task scenarios, a large sample size would be needed to identify the remaining issues (Albert and Tullis, 2013). Over the years, two different viewpoints have emerged on this topic; one that believes that five users are enough to identify most of the usability problems, and another that believes that this number is nowhere near enough (Nielsen, 2000b); (Woolrych and Cockton, 2001). These are discussed in the following.

- Why are only five users needed?

  The pioneers of the first viewpoint, such as Nielsen, Lewis and Virzi, believe that 80% of usability problems can be identified with a sample of five users, which is known as the 'magic number'. They arrived at this conviction after analysing the results of many empirical studies. They find that observing five users allows them to discover 80% of a product's usability problems (Turner et al., 2006). More specifically, they find that the first user discovers almost one-third of all usability problems; the second discovers many repeated problems but new ones appear; the third user discovers a small number of new problems; and the fourth and fifth users also find a small number. After the fifth user, many problems are merely repeated, and fewer and fewer new problems are revealed (Zapata and Pow-Sang, 2012). Nielsen (2000b) summarized this issue eloquently when

he said that ‘’add more and more users, fewer and fewer new problems will appear’’. After the fifth user, many problems are repeated and few are incorporated. At the Conference on Human Factors in Computing Systems 2003 (CHI’03), the panels discussed this issue, by then called ‘the magic number 5’ (Bevan et al., 2003). Nielsen defended his original theory of using only 5 users, and clarified the reasons for this by saying that this is ‘discount usability’, and that resources and time are wasted if the recruitment process engages more than five users. Virzi (1992) argued that the optimal sample size in terms of commercial cost-benefit may be as low as three users. This viewpoint uses the following formula to estimate the problem discovery rate (*p*). Lewis (2006, p.30) defines it as, “the average of the proportion of participants experiencing each observed problem”.

- Proportion of unique problems found (P) = $(\mathbf{1} - (\mathbf{1} - \boldsymbol{p})^{\boldsymbol{n}}) \times \mathbf{100}$)

  Where $\boldsymbol{p}$ is the average problem discovery rate computed across subjects/problems, $\boldsymbol{n}$ is the number of subjects, and P is the percentage of problems that can be discovered. $\boldsymbol{p}$ can be computed by listing all the usability problems identified during the test. Then, for each user, mark all the usability problems, add the total number of the usability problems identified by each user, and finally divide by the total number of problems.

- Why and when five users are not enough

The pioneers of the second viewpoint, such as Lindgaard, Chattratichart, Spool, Schroeder, Hwang and Salvendy, disagree with the above assertion. They criticise using a small number of users arguing that reliability may be lost and usability problems may be missed. Also, employing only a small number of users ignores the individual differences between them, and yet this aspect underpins the relatively straightforward studies utilizing quite closed/specific tasks. Accordingly, they recommend recruiting more than five users. For example, Spool and Schroeder (2001) evaluated four different electronics websites, and they found that five users discovered only 35% of the usability problems. Lindgaard and Chattratichart (2007) conducted nine usability tests; they compared the results of two teams, where team A consisted of six users and team B consisted of twelve. The analyses showed that the teams discovered 42% and 43%, respectively. Law and Hvannberg (2002) reported that five users failed in reaching the 80% overall usability

discovery rate and that eleven users were needed to achieve this percentage. Hwang and Salvendy (2010) analysed the quantitative data of 27 experiments. Those 27 studies all employed three evaluation methods, and linear regression was applied to determine the samples for each. They found that Think Aloud (TA), Heuristic Evaluation (HE) and Cognitive Walkthrough (CW) required nine users, eight evaluators and eleven evaluators, respectively, to discover 80% of usability flaws. As a result, they proposed a new rule for optimal sample size, which is 10 ± 2, recommending its application under general evaluation conditions. In this regard, Faulkner (2003) found that a sample size of ten participants will most likely reveal a minimum of 82% of the problems. However, Schmettow (2012) doubts the ability of ten users or experts to find 80% of usability problems; also, this rule ignores usability practitioners who test with only a few participants in iterative design cycles. Jabbar et al. (2007) developed an adjustable sample-size estimation model for usability assessments by using two factors: Beta (β) and Alpha (α); they found that the best estimation for sample size is about eight users. Later, Turner et al. (2006) improved the small-sample estimation of p by using statistical technique that is called 'GOOD-Turning', and they applied this on eight users. They found that the appropriate sample size would be seven users, even where the study is quite complex in nature. Perfetti and Landesman (2001) argued that twenty users are suitable for many commercial studies. Macefield (2009, p.41) found that "8 to 25 participants per team is a sensible range to consider and that 10 to 12 participants are probably a good baseline range".

Overall, it can be seen that this issue can impact on evaluation results, and so it should be considered before starting any usability studies. Moreover, it will be necessary to examine this issue whilst developing the new method in order to identify the most appropriate sample size.

2.4.2 Developing a new method

Since the 1980s, the user testing method has become the major UEM for evaluating a new and improved interface. Hartson et al. (2003, p. 374) described it thus, "user testing was seen by developers as a way to minimize the cost of service calls, increase sales through the design of a more competitive product, minimize risk, and create a historical record of usability benchmarks for future releases". However, it is too expensive, it is difficult to design the

most appropriate tasks, and it is used only in the latter stages of the design process. It involves users in measuring their speed, accuracy, error rate, and user subjective evaluations. Rubin and Chisnell (2008) pointed out four different shortcomings in UT; the first limitation is that the testing session is always a fabricated circumstance (i.e. not real); the second limitation is that the results of UT do not mean that the product works; the third limitation is that the sample of users may not fully represent the target population; and the fourth limitation is that choosing UT is not always the best approach. Molich and Dumas (2008, p. 280) conducted a comparative usability evaluation of Hotel Pennsylvania's website (CUE-4), and nine teams used the UT method. They found that "usability testing is not the 'high quality gold standard' against which all other methods should be measured. CUE-4 shows that usability testing - just like any other method - overlooks some problems, even critical ones".

There are two prerequisites for any usability evaluation, which are valid and useful results. In terms of the first prerequisite, Macleod (1994, p.2) stated, "it was recognised in the 1980's that usability testing has often failed to meet the first prerequisite". This does not stop just at the reasons relating to the physical and organisational setting of the evaluation; 'context of use' also plays a substantial role, as Section 2.4.3 will explain in detail. For those reasons, developers in the 1990s started to search for other methods that are low in terms of cost, consume less time, and can be used in the earlier stages of the design process. As a result, expert-based inspection methods grew in popularity to fulfill those requirements. Some of these methods are still popular, such as Heuristic Evaluation (HE) and Cognitive Walkthrough (CW). In practice, HE appears to be the most popular form of inspection method (Hollingsed and Novick, 2007).

Furthermore, the growth of the Internet has created a new breed of dynamic websites, which are becoming increasingly interactive in the midst of ever-improving information technologies. Hence, studies have sought to compare and contrast the efficiency of different UEMs in order to find which method is the most adequate for assessing website usability. Huge studies, such as (Hartson et al., 2003), went on to assess and compare UEMs in order to understand the capabilities and limitations of each because some developers have recently questioned the effectiveness of UEMs in terms of their ability to predict problems that users actually encounter (Hvannberg et al., 2007). These studies emphasize that testing web usability appears to be more difficult than testing other systems. For example, various types

of hardware and software are used daily by users to access the Web. This problem has created other difficulties, such as the presence of a large number of websites located all over the world, with different goals, cultures and levels of quality; also, a large number of users access them concurrently (Di Lucca and Fasolino, 2006). Nielsen investigated several mid-sized e-commerce websites. He found that most e-commerce websites comply with only a third of documented usability guidelines. This causes websites to lose almost half of their potential sales because users cannot use the site (Nielsen, 2001).

Thus, the findings of many studies have confirmed that there is a substantial need for enhancing the current UEMs and their processes. Some researchers have found that a number of UEMs are not stable and not readily applicable to many 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, 2005b). Also, other scholars have emphasized that "UEMs continue to change because human-computer systems, their interaction components, and their evaluation needs change rapidly, requiring new kinds of UEMs and constant improvement and modifications to existing UEMs" (Hartson et al., 2003); (Di Lucca and Fasolino, 2006). For example, Silva and Dix (2007) used HE and UT to evaluate YouTube. They found that only two heuristics out of ten were respected by YouTube, which means that it has low usability and thus it completely failed. Also, they used other usability metrics, such as number of errors, number of clicks, and task completion time. The result also confirmed that YouTube had failed. However, they stated, "YouTube's clear success means there must be something really good that makes users go back and back again". Also, they added, "YouTube appears to fail miserably when evaluated with a conventional usability evaluation technique". Silva and Dix (2007, p.106) pointed out that there are many new websites in which users play a major role in their success, even though these websites have low usability levels; thus, it is necessary to understand these Web phenomena in order to find a valid way for evaluating them. They further stated, "this new context requires new and sharp usability evaluation approaches". Hollingsed and Novick (2007, p. 249) reviewed research and practice in usability methods over a period of 15 years starting from 1992, and they found that the majority of studies only compared the effectiveness of the many different approaches to usability evaluation. Perhaps their most important finding was that "some researchers did not propose new usability

inspection methods". This finding was supported by some other researchers who argued that research-oriented endeavours should concentrate on improving and refining UEMs to provide better discount usability instruments for usability practitioners. In fact, both researchers and practitioners of usability engineering would benefit from new methods and tools if they are designed to support the evaluation of a product through context of use, thereby facilitating usability problem classification; they would also benefit from discount (but valid) methods for identifying real usability problems (Hartson et al., 2003); (Chattratichart and Brodie, 2004); (Lindgaard, 2006). Moreover, Cockton and Woolrych (2002, p.18) highlighted this by saying, "discount methods aren't very safe. They can and should be improved." Consequently, these methods need further research to increase their efficiency and effectiveness, in terms of expanding their capacity to identify new problem areas, specifically for emerging domains and technologies.

2.4.3 Context of usability methods

The second challenge in determining the usability of a product is the context of the designed UEM because product usability depends on the context in which it is used (Bevan, 1995). In other words, product usability can be measured by the quality of use in a particular context; thus, it is important to know "whom the product was designed for, what it will be used for and where it will be used" (Maguire, 2001a, p.454). The term 'context' encompasses many aspects, for example, user characteristics and the task goals they are trying to achieve, the environments in which they work, and the method employed for assessing the products they use within that context. Maguire (2001a, p.457) pointed out to another usability definition from the above example on contextual circumstances: " the usability of a product is affected not only by the features of the product itself, but also by the specific circumstances in which a product is used". Accordingly, the International Organization for Standardization ISO (2010) defines usability as, "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use".

In fact, there are two requirements for any usability evaluation, which are valid data and useful results. Thus, an appropriate method must be applied for analysing the data. As we mentioned above, UEMs such as UT and HE started to appear in the 1980's. Although UT is the most

commonly used method, it has failed to meet the first requirement (valid data). This is because some contextual factors can influence its validity; for example, the recruitment of inappropriate users, employing limited or low-quality tasks, evaluating isolated parts of a product, and conducting the evaluation in environments unrelated to real work goals or the conditions of the workplace. This in its turn generates invalid data, and may result in the researcher concentrating only on small issues (Macleod, 1994). Furthermore, HE is another most commonly used method, although it suffers from being too abstract to apply directly, too difficult to comprehend fully, and sometimes difficult to determine when a guideline has been violated (Thovtrup and Nielsen, 1991). Henninger et al. (1997, p.2) supported this claim by saying, " a guideline stating to "always keep users informed of system states" is abstract and open to a wide variety of interpretations in different contexts". Also, Henninger (2000, p.228) highlighted an important issue, which is that "contextualized guidelines are better than abstract or decontextualized ones… Less clear is how guidelines are created to meet diverse application needs, how one applies the guidelines to a specific context, and how guidelines can be refined to meet user task requirements".

To sum up, Mankoff et al. (2003) emphasized the need to propose a new method that would be able to uncover those aspects that are of particular importance in a given particular context because usability evaluation cannot be simply based on the evaluation results of a system that combines one or more of the traditional methods mentioned in the literature review. Various contextual characteristics must be taken into account, such as 'context method' which plays a vital role in influencing the evaluation results; there is also the need to move forward in developing a new usability evaluation framework that specifically encompasses context (Trivedi and Khanum, 2012).

2.4.4 Usability measures

Various types of data can be gathered during an evaluation session. These data can be quantitative or qualitative based on the goal of the usability study. Consequently, failure to measure these data leads to failure in achieving the goal of a usability study as a whole. UEMs have focused on users being able to interact with a system in a way that is efficient, effective and satisfying. These three represent important elements in the IOS 9241 standards

for measuring usability attributes; indeed, they are the most commonly used ones in usability studies, and are measured quantitatively (Hornbæk and Law, 2007). From these, Sauro and Kindlund (2005) proffered a high-level quantitative model of usability metrics, which they arrived at after analysing the data from four summative evaluations based on investigating the correlations amongst the above attributes in order to build it accurately and with equal weighting. For example, efficiency can be measured by the time performance or how long time is spent to complete a task, effectiveness can be measured by task performance or completed tasks and the number of errors, and satisfaction can be measured by questionnaires or a survey.

However, it has been found that these measures have an impact on usability studies. Cairns and Cox (2008) criticised these measures, saying, "what has also emerged is that these measures do not seem to provide insight into how interfaces work and whilst a design may be effective, efficient and satisfying, it can somehow still not be a good user interface". Hornbæk (2006) found that some usability studies faced difficulties in determining how to measure a system's usability, what aspects should be measured, and what the best ways are to measure them. In this regard, many attempts have been undertaken to improve usability measures by proposing additional metrics, such as the number of pages visited, the number of clicks needed to succeed in the tasks, mouse movements, task-difficulty, task-confidence and typing speed as well as theoretical models to predict in advance whether a system is usable or not (Lazar et al., 2010); (Lazar, 2005); (Sauro, 2012b). Furthermore, there is a need to develop UEM performance measures that are computed from the raw experimental usability data produced by each UEM in order to identify which methods are more effective and more valid in discovering real usability problems. However, there are difficulties in developing this because it requires usability researchers to comprehend the ability and shortcomings of each UEM. Also, there are no clear definitions, measures or metrics on which one can depend for this task. For this reason, some effective UEM evaluation criteria have been developed to facilitate a reliable comparison of various UEMs; these are thoroughness, validity, effectiveness and reliability (Hartson et al., 2003).The thoroughness criterion is a measurement that assists in identifying how a UEM can identify real usability problems. Hwang and Salvendy (2010, p.131) define how to compute this as, "the ratio of the sum of unique usability problems detected by all experiment participants' to the number

of usability problems that exist in the evaluated systems". Khajouei et al. (2011, p.345) also defines how to compute thus ''the ratio of the number of real usability problems found using each usability evaluation method to the total number of real problems existing in the user interface of the system (given by the standard-of-comparison usability problem set)''. The validity criterion is a measurement that assists in identifying how a UEM is able to discover usability problems accurately. Sears (1997, p.214) defines how to compute this as, ''the proportion of real problems found using a UEM to issues identified as problems". Khajouei et al. (2011, p.345) also defines how to compute thus "the ratio of the number of the real usability problems found by a method to the number of issues the method (correctly or incorrectly) identified as usability problems". The effectiveness criterion is defined as the ability of a UEM to identify usability problems relating to the user interface (Khajouei et al., 2011). If the level of thoroughness or validity is low, effectiveness will be low as well. All of these metrics have values that range from '0' to '1' (Hartson et al., 2003). Moreover, these metrics are affected by the set of concepts detailed in Section 2.4.6 below, but they can be calculated as follows.

- $\text{Thoroughness} = \dfrac{\textbf{No.of real usability problems found}}{\textbf{Total no.of real usability problems existing}}$
- $\text{Validity} = \dfrac{\textbf{No.of real usability problems found}}{\textbf{No.of issues identified as a usability problem}}$
- Effectiveness = Thoroughness × Validity

In terms of reliability criteria, Hartson et al. (2003, p.167) defined it as "a measure of the consistency of usability testing results across different users (developers)". It can be said that a UEM is reliable when a repeat of an experiment delivers similar outcomes across a range of minor differences in the experiment situation (Hertzum, 2006). Some differences have been investigated in terms of the reliability of UEMs over time. For example, some studies have examined the impact of involving different numbers of users, and others have examined the impact of individual and cooperating users on the results of some UEMs (Lewis, 1994); (Hackman and Biers, 1992). The evaluator effect has also been examined on some UEMs, for example, in the Think Aloud protocol it has been found that evaluators (during an analysis session for the same experiment) discover different sets of problems (Hertzum, 2006).

The reliability criteria can be computed in different ways. The first way is by using the average detection rate of a single evaluator to assess the evaluator effect. It can be computed by dividing the average number of unique problems detected by a single evaluator by the number of unique problems detected collectively by all the evaluators (see the equation below). However, it has been found that the sample size of evaluators affects the detection rate. Hence, whenever the sample size is small, the detection rate will be high, and vice versa; for example, if you have one evaluator, the detection rate will constantly be 100%. In the case of having two evaluators (with no overlap between them), the detection rate will be 50% (Hertzum and Jacobsen, 2001).

- Detection rate = Average of $\frac{|\mathbf{Pi}|}{|\mathbf{P\ all}|}$ over all evaluators

where '**Pi**' is the set of problems identified by evaluator 'i', and '**P all′** is the set of problems identified together by all 'n' evaluators (Hertzum and Jacobsen, 2001).

In the second way, and based on the above drawback, the same scholars developed another measure, which is called Any-Two-Agreement (see the equation below). It aims to measure the extent to which two evaluators agree on the problems they have discovered. It can be computed by the number of problems they have in common divided by the number of problems collectively discovered. The result ranges from 0%, which indicates that the evaluators have not agreed on any problem in common, to 100%, which indicates that all the evaluators have agreed on the same set of problems (Hertzum and Jacobsen, 2001).

- Any-Two-Agreement = Average of $|Pi \cap Pj| / |Pi \cup Pj|$ over all $\frac{1}{2}$ n (n-1) pairs of evaluators

where 'Pi' and 'Pj' are the set of problem discovered by evaluator 'i' and the other evaluator 'j', and 'n' indicates to the number of evaluators (Hertzum and Jacobsen, 2001).

Another way is by using Cohen's Kappa measure. This is a statistical analysis of reliability for measuring the ratio of agreement between two nominal variables. It has often been used in usability studies, and it is a reflection of the detection rate (Mendoza and Novick, 2005). However, it is limited to assessing two subjects only, which is not the case for most usability studies which involve a few evaluators. Consequently, using Any-Two-Agreement has been recommended (Hertzum and Jacobsen, 2001). Finally, the level of reliability can be measured

by using the mean number of evaluators finding a real problem, as proposed by (Chattratichart and Lindgaard, 2008) (see the equation below).

- Reliability = $\frac{\textbf{Number of evaluators}}{\textbf{Number of real problems identified}}$

In addition to above usability measures, cost is another important issue in employing UEMs during an evaluation phase within the usability engineering lifecycle. On this point, some designers believe that no UEM is able to discover all the usability problems, and they are not prepared to extend their projects (adding further expense) indefinitely looking for a definitive solution to the problem of usability. Thus, there is a need for a low-cost method (particularly in terms of time). For these reasons, Nielsen (2009) made a case for discount evaluation, known as the heuristic evaluation (HE) method. However, it has been criticized for being not sufficiently effective (Cockton and Woolrych, 2002); (Hertzum, 2006). Cockton and Woolrych (2002, p.18) highlighted this, saying, "discount methods aren't very safe. They can and should be improved." It is clear that managing a group of several analysts and users is time consuming. Also, some of the problems discovered may be vague or conflicting, and consequently, analysing the test data to understand, match and merge them takes time. Furthermore, the need to conduct extensive causal analysis in order to understand user difficulties during a test (to produce useful recommendations) is time consuming (Cockton et al., 2004a). In this regard, some studies have compared the cost employing UEMs, in particular, UT and HE. Generally, all such studies have proved that UT is more expensive than HE in terms of time taken in designing, conducting and analysing the process. For example, it is worth mentioning, Jeffries et al. (1991), who conducted a comparative research on three different inspection methods, which were HE, cognitive walkthrough and UT in terms of the hourly cost for each participant for each method. Their results showed that HE delivered the lowest score, taking only 35 hours, followed by cognitive walkthrough (43 hours), and lastly UT (at 199 hours). Also, Law and Hvannberg (2002) found HE to be less expensive than UT. The former needed 9 hours, whereas the latter needed 200 hours; the hour counts included designing and conducting those two methods. Additionally, Molich and Dumas (2008) conducted the same comparison, and their results proved that HE is cheaper than UT, requiring 67 hours and 199 hours, respectively. Furthermore, Hasan (2009) compared the cost employing three UEMs in evaluating ecommerce websites, which were

HE, UT and Google analytical software. She confirmed the previous findings by revealing that HE required 247 hours, whereas UT required 326 hours. The Google analytical software was the most costly, requiring 360 hours. The aforementioned hours included the set-up and design of the research tools, and collecting and analysing the data. Similarly, Martin et al. (2014) compared the cost-effective benefits of UT and remote asynchronous testing in terms of the time and effort consumed during each stage of designing, implementing, piloting, conducting and evaluating the two usability tests. They compared the time spent on each stage of the usability evaluations with the time-scale used and provided by the Usability Company (UC) for usability consultants' working hours (7.5 hours on five working days per week). This company is a usability consultancy company based in the UK. They find that remote asynchronous testing needs less time and effort than UT; the latter took 35 days, whereas the former took 30 days.

However, the above studies did not take into account the cost of these methods in terms of fixing the discovered problems. This matter was investigated by Jeffries and Desurvire (1992), and the results were impressive; HE was more costly than UT because the former revealed a substantial number of problems but the majority of them ware minor. Redish et al. (2002) mentioned the high cost of false positives that might be entailed by HE; examining these problems might take yet more time or may lead to new usability problems in the modified interface. As a further example, Martin et al. (2014) calculated the financial costs of employing UT and remote asynchronous testing based on the daily rate of consultants at UC, which is £800.00 per 7.5 hour day, i.e. as if this evaluation were conducted in a business environment. Also, they compared the costs in terms of the number of problems discovered by each method. UT discovered a total of 7 problems (including 4 critical and 3 minor ones), whereas remote asynchronous testing discovered a total of 10 problems (including 6 critical and 4 minor ones). The results show that the total financial cost of UT was £28,000, which means £4,000 per problem, and that the total financial cost of remote asynchronous testing was £16,000, which means £1,600 per problem. This represents a saving of £2,400 (60%) compared to UT. Thus, remote asynchronous testing is a more cost-effective method than UT (Martin et al., 2014). In this regard, Nielsen (2003) indicated that the cost to recruit test participants for usability studies by stating '' the average per-user cost is $171''. Moreover, the cost of employing an evaluator (or the cost-effective benefits) can be computed 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, 1994a), as the formula below shows:

- Cost of Employing an evaluator for a UEM = (Number of evaluation hours) × (Estimate of the loaded hourly cost of the participants'$100')

2.4.5 Usability problems

Essentially, all UEMs discover a number of usability problems. The word problem means something that prevents a user from achieving a goal or that causes some difficulty (which can or cannot be overcome). However, not all discovered problems are usability problems. A usability problem is defined as a flaw in the design of a system that makes the attainment of a particular goal (through the use of the system) ineffective and/or inefficient, and thus lowers the user's level of satisfaction with its usage (Polson et al., 1992); (Albert and Tullis, 2013). There are many terminologies for describing usability problems. It can be described as 'problem types', which refers to unique problems, or it can be described as 'problem tokens', which refers to duplicate violations of the same problem type. Thus, the list of problem types is usually shorter that list of problem tokens (Lindgaard, 2006). In this regard, there are other usability measures that can be used to assess the effectiveness of UEMs; they are severity assessment, counting the number of discovered problems, and matching the discovered problems.

- Severity assessment

To fix a usability problem in a system, it should be rated in terms of its severity by expert evaluators; this rating aims to given a priority for fixing the usability problem. Nielsen (1995b) stated, "severity ratings can be used to allocate the most resources to fix the most serious problems and can also provide a rough estimate of the need for additional usability efforts". Also, (Hertzum (2006)) points out that the severity ratings of discovered usability problems must be reliable, valid and sufficiently convincing to justify the cost of fixing the problems; these ratings help designers through offering guidance on the order in which the problems should be addressed (from 'disasters' down to 'cosmetic problems') (Hertzum, 2006). In this regards Jeffries et al. (1991) defined three different usability problem categories, which are consistency, recurring and general problems. Furthermore, Nielsen

(1992a) described three basic elements that can assist in measuring and evaluating the strictness and seriousness of the level of discovered usability problems. They are repetition of events, influence on users and constancy. However, these three categories differ in terms of severity. Some of the problems may be superficial and frustrating at best, while others may be functionally debilitating. Also, Dumas and Redish (1999) suggested two classifications for usability problems in terms of scope and severity. They further suggested two classes for scope problems, which are global and local; the former is defined as a problem occurring on a certain page, which should be given priority in fixing, and the latter is defined as different problems appearing on several pages.

In addition to this, it has been suggested that three factors play crucial roles in determining the priority of a usability problem and in evaluating its severity. These factors are: firstly, the frequency of the problem that occurs: is it common or rare, or how many users will be affected by the problem? Secondly, the impact of the problem that occurs: is it easy or difficult for the users to overcome, or how much trouble will it make and how far will it affect the user's experience? Lastly, the persistence of the problem: is it a 'one time' problem that users can overcome once, or will users repeatedly be bothered by the problem, or how many times will a user experience the problem? (Hertzum, 2006). Furthermore, these problems are classified into different groups to which a numeric scale is used to measure the severity of each problem (Nielsen, 1995b), and they are as follows.

- (0), this issue is not a usability problem at all.
- (1), this is a cosmetic problem that does not need to be fixed unless extra time is available on the project.
- (2), this issue is a minor usability problem; fixing this should be given low priority.
- (3), this is a major usability problem; it is important to fix this, so it should be given high priority.
- (4), this issue is a usability catastrophe; it is imperative to fix this before the product can be released.

- Problem counting

Many prior studies have been undertaken to compare the effectiveness of UEMs in terms of counting the number of problems discovered and listed by the evaluators. After that, the method that discovered the most usability problems was argued to be the most effective UEM (Mankoff et al., 2003); (Markopoulos and Bekker, 2003). Other researchers have measured the productivity of a UEM by counting the number of problems discovered for a system (Muller et al., 1993). Furthermore, Lindgaard (2006, p.1071) explains the importance of counting all problem tokens because "the resulting tally gives the development team some indication of weak aspects of the interface as well as pointing to individual problem tokens". However, the above approach has many limitations. For example, counting the number of problems conflates potential problems with 'not real' problems. Also, the problem counting approach produces different types of problems that are given the same weight when counted. Additionally, the method of counting problems might include overlapping and duplicated problems. Thus, this approach can have serious ramifications if a large number of problems are counted that cannot be fixed to improve the tested product, or if evaluators refrain from reporting problems because they are unable to think of a solution (Hornbæk and Frokjaer, 2004); (Hornbæk, 2010); (Hertzum, 2006).

- Problem matching

The most common assessment approach for UEMs is through conducting comparisons in order to identify similar and dissimilar problems (found by evaluators). This approach is used in many studies (Molich et al., 2004). The problems found by different UEMs should be matched to identify any overlapping or duplicated problems. Also, matching should be done between the results of the different UEMs, through master usability problem lists, to determine the 'realness' of the discovered problems (Gray and Salzman, 1998). Four different techniques can be used for matching usability problems, which are: similar changes, practical prioritization, Lavery's model, and User Action Framework (UAF). Nevertheless, they are used rarely or implicitly without mention in some studies (Lavery et al., 1997);(Molich and Dumas, 2008); (Andre et al., 2001); (Hornbæk and Frøkjær, 2008).

However, this approach has many limitations. For example, there are no procedures or rules on how to match lists of known problems in many studies, which leads to many studies

having little or no clear procedure on how matching was used, thereby generating unreliable findings. For example, Nielsen (1993) applied the HE method on a telephone operator interface to investigate the effects of different levels of experience on the part of evaluators in discovering usability problems. Hornbæk and Frøkjær (2008, p.505) criticized Nielsen's procedure on the above study by saying, "he does not explain the procedure he followed for matching descriptions of usability problems or the criteria for treating two descriptions as similar". Also, Mankoff et al. (2003) developed a technique for evaluating the usability and effectiveness of ambient displays. They claim they used solid procedures for matching the problems between their developed heuristics and traditional HE with a list of known usability problems. However, they too did not explain their procedures (Hornbæk and Frøkjær, 2008). Overall, (Hornbæk, 2010, p.100) commented, "it seems that matching of usability problems is generally considered straightforward". The second limitation is that many evaluators and observers write down the discovered problems in brief descriptions. This leads to having many considerably different descriptions for the same problems. As a result, problems that are considered similar are replaced with a general description of the problem (Woolrych et al., 2004). The third limitation, as stated by Hornbæk and Frøkjær (2008, p.505), is that "matching has received scant attention in usability research and may be fundamentally unreliable". Thus, it seems that the matching technique has the potential to affect research findings. However, using structured usability problem reports will solve this problem (Lavery et al., 1997), and this is described in detail in Section 2.4.7.

2.4.6 Determining the realism of usability problems

From reviewing the literature, many studies have compared the relative efficiency of UEMs based on the criteria mentioned in Sections 2.4.4 and 2.4.5 (Jeffries and Desurvire, 1992); (Molich and Nielsen, 1990). However, there are risks in terms of producing a set of problems from some UEMs; for example, some problems produced by a UEM would not be actual problems in a 'real work' context of use. Also, some problems that are faced by users in a real work context may not be apparent during an actual evaluation session. Substantively, comparisons between UEMs should be conducted only on the 'real usability problems' found in the target system. Accordingly, four concepts have been proposed for creating reliable UEM assessments; these are miss, hit, false positive and false negative (Lindgaard, 2006). Cockton and Woolrych (2002) defined a missed problem as a known problem not discovered

by participants, or a known problem not revealed by the UEM in question. A hit problem is defined as one that was successfully predicted and subsequently discovered by the user or one that is revealed by the UEM in question. A false positive problem is defined as an unsuccessful expectation, i.e. one that represents not a real problem in the tested interface. Lindgaard (2006, p.1069) gave another definition for false positive: "the dismissal of a problem that turns out to be problematic when tested by another method". A further definition was given by Chattratichart and Brodie (2004, p.1121), stating "some of the problems predicted by a UEM that should have been hits could be mistaken as false positive just because they were not found by user testing". A false negative problem is defined by Lindgaard (2006, p.1069) as ''the dismissal of a problem that turns out to be problematic when tested by another method''. In this regard, high validity means a low proportion of false positive, and vice versa. Also, discovering more real problems leads to a high level of thoroughness, and vice versa (Woolrych et al., 2004); (Cockton et al., 2004a).

To calculate the false and missing problems accurately, one must first correctly find as many known usability problems as possible in order to establish a standard list of problems for matching it with the candidate problem list (to decide whether a particular found problem is actually on the standard list, and thus, whether it is a real problem) (Hartson et al., 2003). Woolrych et al. (2004) outlined six techniques for establishing a standard list of problems, which are helpdesk logs, logging (via software), observation of real usage, user interviews, user diaries, and UT. The fact is that the final usability problem list is identified by end-users (not by expert evaluators), and therefore the realness of any usability problems needs to be established by the user. In this regard, UT is the gold standard for comparison, and it is used overwhelmingly in studies that evaluate the performance of UEMs (Hartson et al., 2003). Generally, the key purpose of employing UT in current usability studies is thus to confirm or expose that the problems discovered in HE truly cause users difficulty (Lindgaard, 2006). Accordingly, missed problems are defined as problems discovered by UT but not discovered by HE. A false positive is defined as a problem discovered by HE but not discovered by UT (Cockton and Woolrych, 2002). Sauro (2012a) defines hit problem as, "meaning the discounted method (HE) hit on the same issue as found in the traditional evaluation method of usability testing (UT)". Lindgaard (2006, p.1072) defined false negative as, "issues rejected by the HE but found to be problematic in the user test". In other words, problems

are classified as false positive because they are not discovered by UT; however, in fact they are real problems but UT failed to discover them. In this regard, the risk of depending on a small sample in UT was investigated, and the results show that approximately half of the identified problems could have been missed by relying on only five users (Faulkner, 2003); (Woolrych and Cockton, 2001).

In brief, there are many and various problems that can result in a miss, a false positive or a false negative, and these are not due to flaws in the method assessment. For example, there may be a misunderstanding on the part of evaluators in analysing a UEM; or it may be that UT fails to expose a predicted problem, which can lead to the incorrect scoping of an inspection method; there is also the effect of evaluators and user sample sizes; and finally, the quality of the structured task and of the problem extraction reports may be insufficiently good, as described in detail in the next section (Faulkner, 2003); (Cockton et al., 2004b);(Woolrych et al., 2004).

2.4.7 Structured problem report formats

An important phase in the evaluation process is reporting; it is incorporated into all the steps, from the pilot study, through the session notes and problem reports, to determining the overlapping problems amongst the UEMs and participants and making recommendations. In fact, there is a question that has been the subject of much debate amongst usability practitioners in assessing UEMs, particularly inspection methods. This question is "why some problems get missed but others are falsely predicted. Missed problems are either never found or are mistakenly dropped. False positives get found, but are mistakenly preserved!" (Woolrych et al., 2004). Consequently, a certain level of ability is needed to distinguish finding and collecting problems from identification and elimination. This is particularly so in the validation process in terms of matching the predicted problems found by analysts against the usability problems found by users. Capra (2006) asserted that poor documentation and communication of the usability problems identified can lead to reducing the effectiveness of a UEM, reducing the return on the effort spent in conducting the assessment, and reducing the number of problems that selected to fix. Thus, it is an important to use a structured usability problem report to compare the UEMs in question in order to facilitate their matching (Hornbæk and Frøkjær, 2008). Lavery et al. (1997, p. 247) highlighted the importance of this

by saying, "the content of usability problem reports and their matching is a major methodological problem for the scientific study of usability methods". This is because using the unstructured report leads to the incorrect identification of problems from within the empirical data, which is matched to an analyst's prediction, which in turn leads mistakenly to representing a false positive as a hit. Another example is when a real problem is wrongly eliminated from the empirical data, and is not found in UT, this leads (mistakenly) to representing it as a false positive. A further example is the absence of a description for a problem, leading to the incorrect or misleading merging of analysts' predictions. Also, using an unstructured report leads to an increase in the evaluator effect on the reporting of usability problems because each evaluator will report different problems. This makes the matching procedure (between their problems) more complicated, and thus each evaluator reports more unique problems (Woolrych et al., 2004); (Hornbæk and Frøkjær, 2008).

Furthermore, Skov and Stage (2005, p.2), in discussing analysts and users, declared, "they generally describe what they have done but less about the specific way in which they identified each individual usability problem". In this regard, three main criteria should be considered in generating a problem report, as advised by Lavery et al. (1997, p.251), which are: firstly, "problem reports must be of comparable granularity such as the same level of abstraction and/or generality". This indicates that using a general description report and comparing it with a specific report (between different methods) will lead to failure in meeting the first criterion; the report formats should be consistent across the methods used to obtain reliable validation. Secondly, "there must be explicit matching rules". Using a structured report would help to make the matching possible (between the predicted and the real problems), thereby meeting the second criterion. Also, a structured problem report would stop researchers from incorrectly matching the predicted problems to the real problems or from incorrectly merging the predicted problems (of all the analysts) to produce a single set of predictions (Cockton et al., 2004b). Thirdly, the good report format should be derived from a definition of 'usability problem'. Thus, Jeffries (1994) advised that usability problem reports should consist of three parts, which are a description of the problem (based on the users and their tasks in real usage), a description of the severity of the problem, and a solution to the problem. However, she later found difficulties in understanding the problem descriptions because some evaluators include the solution in their problem description, rather

than focusing on describing the effect of problems on users. Also, she "did not summarize the areas that the reports covered well" (Capra, 2006, p.10).

In this regard, Lavery et al. (1997, p. 258) developed a model for structured reports, which includes four components. These components are: context or design change, cause, breakdown and outcome. They described the goal of each component thus, "the cause describes what is wrong and needs to be fixed. The breakdown and/or outcome provide justification why the cause is problematic. The outcome suggests why the problem is severe and how it relates to any usability criteria. The context describes when the problem occurs, suggesting possible frequency and in some circumstances the solution". However, it was a relatively early paper in which to discuss matching reports; they gave some examples of usability problems and concluded that their model would facilitate collating them, but their paper did not describe any particular procedure for matching problems, did not document whether it improved the matching of problems, or whether evaluators found the format too laborious. This technique was examined practically and the results showed that the participants faced difficulties in describing the usability problems to the four components of the model. Also, the evaluators found that it was hard to interpret the problems reported by the participants, and this was clear in the low levels of agreement among them in terms of identifying the overlapping and unique problems (Hornbæk and Frøkjær, 2008).

Furthermore, Cockton et al. (2004b) described an extended reporting format using a heuristic method. This report consists of four main sections. The first section aims to describe the problem and associated user difficulties (by analysts) through using four elements, which are: problem description, likely/actual difficulties, specific contexts, and assumed causes. The second section addresses the discovery resources and methods (again by analysts), who should explain how they discovery and report problems; also, this section indicates if their method is system- or user-centred and unstructured or structured. This yields four categories which are: system scanning, system searching, goal playing and method following. The third section deals specifically with the application of heuristics to individual problems. Accordingly, analysts should provide evidence of conformance rather than just naming a heuristic. The fourth section requires analysts to explain any problem elimination, with specific reference to user impact and behaviour (Cockton et al., 2004b).Their results show how effectively their report was in helping evaluators to predict fewer false positive

problems, leading to an increase in the validity of the HE method, compared to the previous study using a simpler reporting format. However, "the study is indicative only, because the use of the extended reporting format is compared only to a previous study, with many differences to the study in which the extended reporting format was being used" (Hornbæk and Frøkjær, 2008, p. 101).

In conclusion, few studies have shed light on this issue. Having an arbitrator's report should facilitate the comparison of UEMs, defining overlapping problems and the realness of problems, eliminating ambiguous problem descriptions, distinguishing between different types of problems, and producing useful information on fixing a number of problems (which in turn should assist developers and designers to change and improve the system that was evaluated). Using various forms of reports will lead to generating different results on the relative merits of UEMs, which will influence the findings of usability research as well as influencing the effectiveness of evaluation methodology. Thus, there is a need for a standard approach for describing usability problems, so that they can be more easily and more directly compared (Andre et al., 2001); (Hornbæk and Frøkjær, 2008).

2.5 Conclusion

This chapter has provided a broad overview of usability evaluation methods for evaluating websites from different perspectives. Also, the current issues in usability evaluation were reviewed, so that researchers can better understand and thus tackle these issues. The literature showed that there has been a lack of research focusing on how to develop a method that combines the advantages of HE and UT but avoids their drawbacks. Many studies have developed usability evaluation methods; however, those that were assessed were designed to deal with certain aspects of usability, in certain areas of the product. Furthermore, it is clear from the literature that this field is still in need of more research in order to develop the most appropriate method for evaluating website usability in context. There are very few guidelines to help anyone who desires to develop a new method for assessing emerging products, or frameworks to overcome the defects of the traditional UEMs. In other words, there has been no research to the best of my knowledge that presents a framework that is readily capable of adaptation to any domain, and that combines the advantages of HE and UT and can help to generate an evaluation method for assessing the usability of products in a particular domain.

The next chapter provides further information (as it is a continuation of the literature review) by proposing an adaptive framework, which is the main aim of this study; this framework should help to solve the aforementioned problems, and should be invaluable to designers, developers, researchers and managers who wish to uncover usability problems related to specific usability areas. That chapter will describe the components of the adaptive framework as well.

Chapter 3: Research Adaptive Framework

3.1 Introduction

Chapter 1 introduced the problems that are now challenging the traditional usability evaluation methods (UEMs), and it highlighted the value of developing a new UEM for assessing the usability of the new breed of dynamic websites and applications that are growing rapidly in use and that have had a great impact on many businesses. Also, the urgent need for continuous assessment for these websites and applications (to measure their efficiency and effectiveness, to assess user satisfaction, and ultimately to improve their quality) require the design of a formative and summative evaluation method for achieving high levels of quality. Chapter 2 reviewed the literature for this topic from various different prespectives. It shed light on the current issues on usability evaluation, which motivated this research in order to improve usability testing. This chapter illustrates how the aim of this research is to be satisfied by developing a systematic adaptive framework to generate a context evaluation method for any product. This method is called the Domain Specific Inspection (DSI) method. It also explains in detail the components of this framework, how to utilize it, and the target products adopted to evaluate it.

3.2 Research adaptive framework

Rogers et al. (2011) defines 'framework' as, "a set of interrelated concepts and/or a set of specific questions that are intended to inform a particular domain area". Lazar et al. (2010) highlight the benefits of methodical frameworks by stating, "they can help you frame the research questions, decide on the specific research approach to adopt (e.g. survey, interview, focus group, etc.), and identify the concepts and questions to be included in each approach".

From reviewing the literature, several frameworks have been published in the HCI field that assist in building and evaluating an interactive design (within the scope of the user's experience) for products. These include a framework for helping designers think about

how to conceptualize and socialize a product, a cloud usability framework that offers a structure for evaluating the main attributes of the cloud user's experience, and a framework for the design and implementation of usability testing of mobile applications (Norman, 2002); (Rogers et al., 2011); (Stanton et al., 2014);(Chisnell et al., 2006); (Zhang and Adipat, 2005);(Neto and Campos, 2014). However, to the best of researcher's knowledge, there is no adaptive framework that is valid across time, that can assist a researcher/developer to generate a domain-specific inspection method, and that then uses it to evaluate the usability of products in a domain in the context of what it was built for.

The literature shows that many researchers have attempted to enhance the traditional inspection heuristics through assessing their heuristics to identify those heuristics that do not work, and remove them. Then they develop new heuristics to cover areas not covered by the traditional heuristics. Finally, these new heuristics are added to the ones remaining from the traditional heuristics. This technique is called extended or modified heuristics (Ling and Salvendy, 2005a). Other researchers went further than that through developing customized heuristics for areas such as games, e-learning systems, and mobile launchers for elderly people (Al-Razgan et al., 2012); (Alsumait and Al-Osaimi, 2009); (Pinelle et al., 2008). They used three kinds of techniques and each one has drawbacks. The first technique is thorough evaluation of the several existing sets of heuristics in literature review to determine which ones provides the widest explanatory coverage. For example, Nielsen and Molich (1989) examined the reported problems from eleven previous studies and rated their explanation by 110 previously collected heuristics, and they performed a factor analysis to find the heuristics with the most explanatory power. Finally, they arrived at a list of seven heuristics which later developed into ten heuristics. This technique's disadvantage is that literature on the usability of the targeted product might be limited or not reported in detail. Also, Brown (2009, p.17) criticized this technique by stating ''…the usability problems have to be discovered before heuristics can be created''. The second technique uses expert opinion and then develops heuristics from their opinion. For example, Federoff (2002) collected several heuristics from different sources and used expert opinion to choose those most useful for game design. The disadvantage of this technique is that it relies only on the expert opinion and ignoring the opinion of the real users of the targeted product. Also, Brown (2009, p.17) criticized this technique by stating ''this method is difficulty in selecting the 'best' heuristics. Using experts will create the same biases as expert created heuristics and most end users will not have sufficient knowledge to assess the heuristics''. Thus, this leads to the development of specific

heuristics that are not efficient for the chosen domain. Also, they are not valid to all products on the same domain, and they rely on expert opinions and do not necessarily linked to users' needs (Pinelle et al., 2008). The third technique uses a questionnaire to obtain users' feedback on the targeted products and then develops heuristics from their feedback. For example, Brown (2009) investigated the user's interaction through a web questionnaire to find the main usability issues in computer games. The data collected from the questionnaire was analysed using the content analysis method and the result was used to produce categories and themes that describe the main computer game usability issues. After that the results produced from users and the previous studies were drawn together to form focused heuristics. This technique is good but also has limitations, which are that the discovered problems might not cover all usability problems for the targeted product, or these problems might suffer from the lack of user input, and finally the targeted product does not represent the main product genres or non-traditional games in the same domain (Dykstra, 1993).

Considerations have been given to certain aspects, all of which are related to the roadmap designed to develop the adaptive framework in this research. Firstly, the researcher believes that it is important to understand the meaning of usability in both general and contextual terms before proposing any usability framework. The general meaning of usability refers to the usability attributes of a product that are classified from a 'software quality' perspective, and are defined by the International Organization for Standardization ISO (2011) as, "a set of attributes of software which bear on the effort needed for use and on the individual assessment of such use by a stated or implied set of users". The contextual meaning of usability refers to 'quality in use', which is defined by the International Organization for Standardization ISO (2010) as, "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use". Secondly, 'quality in use' is reliant upon 'context of use', and a successful level of quality in use will rely on the specific contextual circumstances in which a product is used; this refers to users, tasks and social environments. These circumstances were identified in the process of User Centered Design (UCD) (Bevan and Azuma, 1997). Accordingly, Maguire (2001a, p.457) developed from the above contextual circumstances another usability definition: "the usability of a product is affected not only by the features of the product itself, but also by the specific circumstances in which a product is used". Thus, these circumstances are essential for considering any usability framework. Thirdly, the researcher believes that

mixed methods are the most effective approach to find the root cause of interface problems for a new product, to discover usability problems for a new product, to propose changes for a new design for a new product, to provide recommendations for improving the user experience, and to understand and make the new approach successful (Sauro, 2012b). Sauro (2011c) states, "there is not a single silver bullet technique or tool which will uncover all problems. Instead, practitioners are encouraged to use multiple techniques and triangulate to arrive at a more complete set of problems and solutions." Also, Hornbæk (2010, p.106) asserts, "it is quite difficult to identify a single best UEM because none of the studies has looked at evaluators using combinations of methods". Hence, he argues in favour of not looking at individual techniques, but at combinations of techniques. Additionally, Hornbæk and Frøkjær (2008) and Hornbæk (2010, p.108) conclude, "usability testing by itself can't develop a comprehensive list of defects. Use an appropriate mix of methods". Finally, a multiphase design is one that uses common mixed-method designs to provide an overarching methodological framework in order to develop an overall programme of research. Tashakkori and Creswell (2007, p.4) defined the mixed-method as "research in which the investigator collects and analyzes data, integrates the findings, and draws inferences using both qualitative and quantitative approaches or methods in a single study or programme of inquiry". Creswell and Clark (2011, p.100) defines multiphase designs thus, "they occur when an individual researcher or team of investigators examines a problem or topic through an iteration of connected quantitative and qualitative studies that are sequentially aligned, with each new approach building on what was learned previously to address a central programme objective". Also, Sandelowski (2000, p.247) defines multiphase designs thus, "multiphase designs occur when the researcher alternates the quantitative and qualitative methods across three phases, such as qualitative then quantitative and then qualitative". Moreover, the core phases of development in a method presented by Cairns and Cox (2008) are taken into account. These phases are: 1) identification of an opportunity or need; 2) development of more detailed requirements (optional); 3) matching opportunities, needs and requirements; 4) development of the method; and 5) testing the method.

The above definitions and concepts, the problem statement, the research questions, the review of current issues on UEMs, and the researcher's experience; all together are adopted to represent a roadmap for researchers to devise an adaptive framework (Figure 3.1). It offers different perspectives (users and usability experts) and corroboration of

findings across techniques, thereby leading to more rigorous and defensible findings. The following is briefly explanation for the justifications of these components.

1. Development Step One (D1: Familiarization)

The literature review is the first step that is adopted in the adaptive framework. This step is started because it is most helpful to understand the chosen domain broadly. Also, to gather more data on the chosen domain from different angles that can help in scoping of the chosen domain (Janesick, 2000); (Lazar et al., 2010). There are some studies in literature review that have low external validity and/or low internal validity which they can not be sufficient resource for understanding the chosen domain (Gray and Salzman, 1998). Then, the next step is adopted by involving the real users of the chosen domain in building the new DSI method.

2. Development Step Two (D2: User Input)

This step helps in increasing the external validity for the new method through using random sampling to select participants during conducting the mini-user testing. Also, to develop more detailed requirements than that were found in the literature review (Cairns and Cox, 2008). To make the new DSI method applicable on many websites in the chosen domain through covering all features and areas in that domain, the experts should be involved as the third step. This step cannot be used to be the third step in the adaptive framework due to that this step should be completed and verified by perspectives of experts (Pinelle et al., 2008).

3. Development Step Three (D3: Expert Input)

This step is adopted to be after the previous steps to cover any missing information for supporting the development process for the new DSI method through taking advantage of their expertise. This step can not be used to be the second step in the adaptive framework due to that the essential requirements and needs for the real users are unknown yet. Also, the aim of the expert's step is to develop more detailed requirements that are linked to users' needs (Brown, 2009). Moreover, this step will help in the data revision of the previous steps to exclude any related data to the research domains that have been excluded due to that they are out of scope this research.

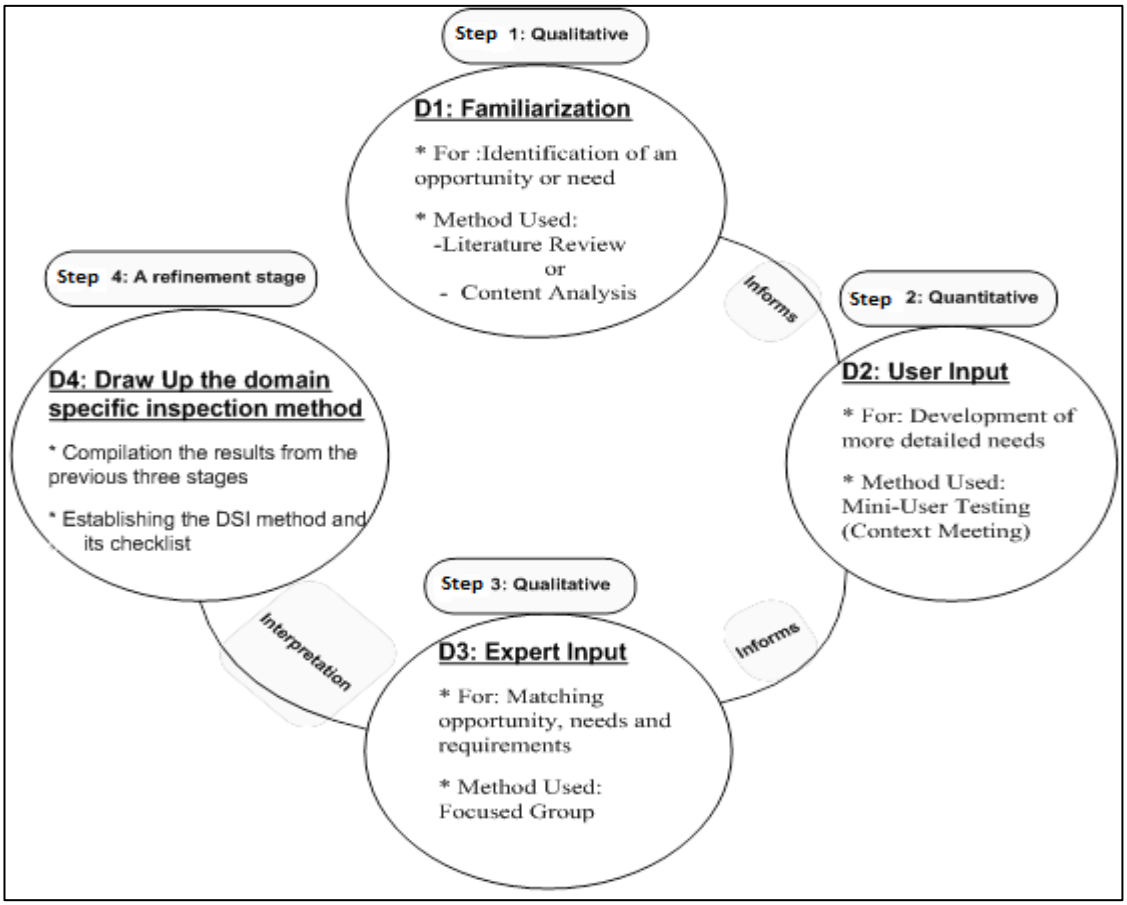
4. Development Step Four (D4: Draw up the DSI Method: data analysis):

This step aims to analyse the massive data that will obtain from the previous three steps (Ling and Salvendy, 2005a). This step can not be used to be the second step or the third step due to that the data will be analysed for example from the first step, however, a new

data will be obtained from the second or third steps which they need to be analysed and revised.

To sum up, the steps of the adaptive framework are fixed. Thus, the importance of this framework is not just its generated method, it is also the existence of the adaptive framework for use. The next section explains in detail the adaptive framework's components.

Figure 3.1: The adaptive framework for generating the DSI method



3.3 Components of the adaptive framework

The adaptive framework consists of four development steps, as outlined below, for gathering together suitable components to develop a domain context-specific inspection method (DSI). There is an adopted method in each step, and the limitations of each method are complemented by the strengths of the others, as follows, and see Figure3.2;

- Development Step One (D1: Familiarization): This step starts from the desire to develop a method that is context-specific, productive, usable, reliable and valid, and that can be used to evaluate a product. The literature review is an essential step to start

understanding, examining, and gathering data on the product that is under investigation. The literatures related to the usability for the chosen product in the chosen domain are examined. Also, published works related to developing heuristics for the chosen product are reviewed. If there is very limited literature or not enough information to build on (e.g. new product), a content analysis method is a good method in this situation. Lazar et al. (2010, p.289) describe the process to develop new information by using an emerging coding approach in the content analysis thus "multiple researchers first examine a subset of the data independently and each develops a list of key coding categories based on their interpretation of the data. The researchers compare their category list, discuss the differences, and reach a consolidated list that all agree upon. Then each of them applies coding independently using the consolidated list. In the next step, the codes of multiple coders are compared and reliability measures are compared''. In general, the content analysis is defined as using a quantitative method and/or qualitative method for an in-depth analysis of media content and audience content to generate new knowledge. This method entails reviewing the published material (e.g. book, journal papers, websites, video, and audio), interaction design, and more deeply in a specific focus on knowledge of the chosen product. Also, it helps to gain a high level of understanding of the chosen product and the reason behind its development. It seeks to examine the stated purposes in the selected product and the relationship between that purpose and the users' requirements (Lazar et al., 2010). Lazar et al. (2010) recommend three steps for coding in the published material which are: 1) look for specific items; 2) ask questions constantly about the data; and 3) making comparisons constantly at various level. Overall, this step helps to identify usability problem areas for the chosen product and to formulate specific heuristics for each usability problem area (see Figure3.2). The results of this step can be used as the starting point for next step.

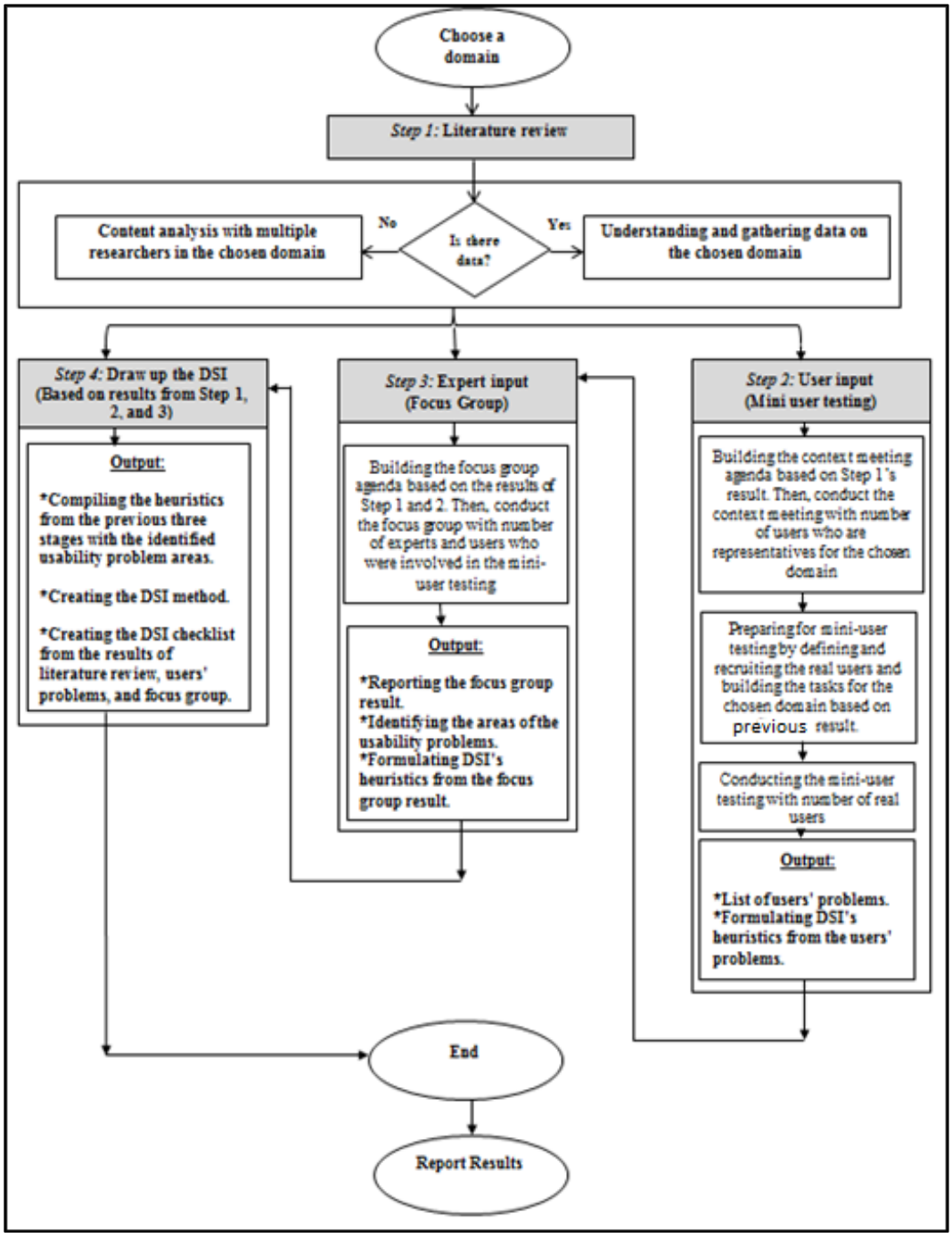
- Development Step Two (D2: User Input): This step consists of mini-user testing which includes a context meeting, context task scenarios, the Think Aloud (TA) protocol and questionnaires. The context meeting is conducted with a number of stakeholders, developers, designers and user representatives of the chosen product to understand the context of use, to specify the usability requirements, and to define the user types and tasks. Then, a number of users are recruited for mini-user testing; these individuals are chosen based on the result of the context meeting, and are asked to perform a set of context tasks that are formulated based on the results of the context meeting, to 'think

aloud' whilst so doing and then to fill out a questionnaire. The broad aim of this is to elicit feedback on a product from the real users in order to appreciate user perspectives, to identify user requirements, to establish extra user requirements for the benefit of user experience, to give an overview of how the product will be used, to provide reliable results, to set the overall usability goals, to learn from the errors made by the users. This allowed the identification of usability problem areas, and also the formulation of specific heuristics from the usability problems that were found for developing or designing an efficient product (that helps them to be highly productive in their business). Understanding the users in their contexts has long been a key part of user design, which can support designers in designing interactive and usable products. Consequently, this step in the adaptive framework directly benefits from including the advantages of user testing (UT), and helps to develop new knowledge if there is limited literature available on a new product (see Figure3.2). The results of this step can be used as the starting point for next step.

- Development Step Three (D3: Expert Input): This step of focus group aims to consider what resources are available for addressing the various needs. These resources, such as any issues arising from the mini-UT results and the literature review, require a discussion amongst experts (in the domain 'double' and/or usability 'single'), who have a good knowledge of the product, in order to obtain a broader understanding of the specifics of the prospective domain. Conducting a focus group session with 5 to 10 people is considered a valid sample size for this method (Rogers et al., 2011). This step offers the participants (who were involved in the previous step) the opportunity to talk about the experiment, and it allows various or sensitive matters to be raised that might otherwise be missed. This can lead to a better design or to presenting new ideas for the new method. Furthermore, it entails garnering more information through conversations with experts to discuss the results of the literature review and mini-user testing. This allows the researcher to identify the areas of usability problems related to the selected domain from the overall results, identify DSI heuristics, and ultimately glean a coherent picture of the usability of the selected product. A reliability evaluation of these usability areas is computed until a satisfactory level is achieved. These areas provide designers and developers with insights into how interfaces can be designed to be effective, efficient and satisfying; they also support more uniform problem descriptions, can guide evaluators in finding real usability problems (classifying them by type), and can facilitate the problem-matching process, thereby facilitating the

evaluation process by judging each area and page in the target product. Also, this step helps to formulate specific heuristics for each usability problem area. Substantially, this step in the adaptive framework directly benefits from including the advantages of heuristic evaluation (HE), and can be used to develop new knowledge if there is limited literature available or feedbak from the users; it also benefits from the expertise of experts, thereby strengthening and supporting the results of the pervious mini- user testing (see Figure3.2).

Figure 3.2: The adaptive framework process for generating the DSI method



Development Step Four (D4: Draw up the DSI Method: data analysis): The aim of this step is to compile the results that have been collected from the previous three steps. Then, the identified DSI heuristics are classified according to the agreed usability problem areas. Thus, the DSI method is created to be closely focused on the targeted domain. Furthermore, the DSI checklist is established in order to address each area of the selected domain, facilitating the use of the DSI method (see Figure 3.2). In other words, the DSI checklist aims to provide guidelines to facilitate the process of evaluation by using the DSI method. It includes most elements of the chosen domain in order to provide a wide range of evaluation of websites in the chosen domain.These elements are classified under the appropriate heuristics. This checklist allows anyone to adopt any usability area with its heuristics and checklists to evaluate a specific part of the targeted domain. This checklist should be piloted befor using in a real experiment.

3.4 The chosen domain

The first step in an initial preparation phase is selecting the websites. Having proposed the adaptive framework, it was decided to evaluate its practicality by applying it to a real-life experiment. From the literature review, it was found that the evaluation of free educational websites and social network websites is a subject area that has not yet been fully explored, nor have any context-specific methods been generated for these domains based on the researcher's knowledge (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.

3.5 Conclusion

This chapter has introduced the adaptive framework that is to be used as the main tool for generating the DSI method, which is implemented later in this research. This adaptive framework focuses on the product's usability, so the steps of the adaptive framework are fixed. The next chapter provides a review of the main qualitative and quantitative research methods, illustrating the advantages and drawbacks of each method and the justification behind their selection for this research. Furthermore, it reviews the sampling techniques, highlighting the techniques employed here and justifying their use. The chapter concludes by identifying the testing methods for the adaptive framework, and the processes and procedures employed to collect the research data.

Chapter 4: Research Methodology

4.1 Introduction

This chapter aims to overview the research philosophy, to describe the research methodologies used to investigate the problem statement, and to highlight the methods employed in this comparative study in order to achieve the aims and objectives of this research; this is followed by an explanation of the research design. Also covered are the approaches adopted for the first and second experiments as well as their overall design, explaining the selected methods for testing the adaptive framework and how they are utilised, describing in detail all the factors, conditions and measures that were considered when preparing the experiments and during the procedure of the evaluation and testing sessions. Furthermore, this chapter illustrates how the data were collected, analysed and presented for each method. Finally, it discusses the validity and reliability of the research.

4.2 Research philosophy

The starting point for any researcher is to choose a topic in order to solve a specific problem; this entails careful consideration of the aim and objectives, the methodology and the philosophy to conduct the research. In this regard, the meaning of the word 'research' requires clarification. Sekaran and Bougie (2010, p.1) defined the meaning of research as "an organized, systematic, data-based, critical, objective, scientific inquiry or investigation into a specific problem, undertaken with the purpose of finding answers or solutions to a definite inquiry". Consequently, research requires specific methodologies to meet the study's objective (Bhattacharyya et al., 2010). The meaning of 'philosophy' in this context was defined by Collins and Hussey (2013, p.46) as "the progress of scientific practice based on people's philosophies and assumptions about the world and the nature of knowledge". They and Easterby-Smith et al. (2012) emphasised the advantages of understanding the philosophical issues inherent within the research study as these can assist in answering the research questions through developing or designing a framework, theories and methods.

There are two types of research philosophy, which are positivism/scientific and interpretivism. The former uses quantitative research methods. It adheres closely to the hypothetico-deductive approach, which includes systematic observation, description of phenomena contextualized within a model or theory, the presentation of research question, conducting controlled experimental study, the use of inferential statistics to test hypotheses, and, finally, the explanation of the statistical results in light of the original theory (Cacioppo et al., 2004); (Ponterotto, 2005). The latter uses qualitative research methods. It was developed in reaction to criticism of the positivism/scientific approach. It is an alternative to the "received view" or positivist philosophy, and it is based on the study of phenomena in their natural environment, distinguishing people from objects. Knowledge is achieved in interpretivist philosophy by using an inductive or empiricist approach (Walliman, 2006); (Hasan, 2009); (Bryman, 2012).

Furthermore, both quantitative and qualitative methods are often used by researchers as this complementarity can be used to support the research in achieving its goals and in analysing its data in order to answer its questions in more depth. The quantitative approach helps to make comparisons between the various research factors, identifying their impacts and finding correlations amongst them, whereas the qualitative approach helps to deepen our understanding of the factors affecting the phenomenon and to clarify the underlying reasons for any correlations. The quantitative method (e.g. survey methods, laboratory experiments, and mathematical modelling) was developed in the natural sciences to study natural phenomena. The qualitative method (e.g. action research, interview, questionnaire, case study) was developed in the social sciences for studying social and cultural phenomena (Myers and Avison, 1997). In this regard, the scientific philosophy tends to produce quantitative data, and the interpretivist philosophy tends to produce qualitative data (Howe, 1988). However, the quantitative and qualitative methods can be combined, which is known as multi-method research or mixed-method research; this approach benefits the collection and analysis of data, and assists in finding correlation relationships and in interpreting the data or variables, thereby providing a more complete set of findings, ultimately increasing the credibility and validity of the results. Also, it has been recommended that adopting several methodologies increases the confidence of the research output and clarifies the phenomena under investigation (Myers and Avison, 1997); (Ponterotto, 2005); (Punch, 2013); (Chen and Hirschheim, 2004); (Ritchie et al., 2013).

In this study, the positivism/scientific philosophy approach and mixed methods have been adopted in terms of the above explanations to satisfy the aim and objectives mentioned in Chapter 1. Specifically, this research aims to construct and evaluate a methodological framework adaptable across domains for generating the domain specific inspection (DSI) method in order to assess and improve the quality of the chosen product in relation to specific areas. The knowledge obtained from the adaptive framework has been used by the researcher to generate the new DSI methods; the framework was constructed by way of controlled experiment, observation and interpretation of the users' actions while interacting with the chosen product, learning from their errors and discovered usability problems, and also involving and understanding the comments of expert evaluators during focus group discussions. Then, the adaptive framework was evaluated experimentally by applying the DSI methods against two well-known evaluation methods to determine which usability evaluation methods are good in evaluating each of the usability problem areas in the chosen product (other measurements are also used). Finally, the results of the three methods will be analysed and compared analytically, empirically and statistically to examine the research hypothesis. Thus, a mixture of both quantitative and qualitative methods was adopted due to the nature of the present study; the next section explains in detail the research methods chosen in this study. The reason for employing a blended method is to offer a solid framework for rigorous research and to construct and test the adaptive framework, as explained in Chapter 3 and Section 4.4.8.

4.3 Research methods: review and selection

Research methods ensure that the data are collected with the most appropriate instruments. Bryman (2012, p.27) defined a research method thus, "it is simply a technique for collecting data. It can involve a specific instrument, such as a self-completion questionnaire or a structured interview schedule, or participant observation whereby the researcher listens to and watches others". In the field of Information System (IS), the scientific and interpretative paradigms have different methods, as they were divided by (Galliers, 1992) as shown in Table 4.1. The most well-known examples for the scientific paradigm are experimental research and surveys, and for the interpretative paradigm are action research, interview and focus group. The following sections offer more explanation on the methods that were adopted or rejected for this study.

Table 4.1: Research methods: overview

Scientific	Interpretative
Experimental research	Action research
Surveys	Focus group and interview
Case studies (Hasan, 2009)	Case studies (Collins and Hussey, 2013)
Questionnaire	Descriptive/Interpretive research

4.3.1 Experimental method (Exploratory Experiments)

An experimental method or laboratory experiment is the most conclusive of scientific methods. This method is designed to take place in the laboratory, and it is used to elicit variables of subjects that exist outside the laboratory. This method has certain characteristics, such as control and manipulation that are not evident in most non-experimental methods. This research is experimental research as a new method is proposed and then compared with other existing methods. This research is not driven by any pre-existing theories and hence is not "hypothesis driven". Thus, the exploratory approach is adopted in this study as it is better suited to the type of research that uses a research question, as is the case with this study (Vassar, 2012). There is set of processes that should be taken into account before starting and during the experiment; for example, preparing the experiment of materials such as an introduction script, consent and withdrawal forms, and instruments. Another example is a pre-test process, which includes a set of procedures, such as a training session for the participants, conducting a pilot for the experiment and its instructions, and improving some aspect of the experiment if required. Also, the participant recruitment process should be considered in the early stages of the experiment, and they should be chosen in terms of certain characteristics; they should also agree to be participants to avoid any ethical or legal risk, etc. Furthermore, conducting a debriefing meeting is generally considered to be the conclusion of the experiment; it may be through a focus group, interviews or a questionnaire with the participants, before finally closing the experiment. This method does not aim to judge the participants or to measure so-called intelligence, but to explore for relationships between their performance and other factors. Without a doubt, this method has advantages and drawbacks, as do other methods. The main advantage is that it generates quantitative data, which the investigator can analyse by using inferential statistical tests, for instance, to measure the significance difference or any correlation between the variables. Also, it can be easily replicated by using accurate measurements such as observation and user behaviours,

and the researcher can isolate and control a small group of variables. However, it is criticised for recruiting 'unreal' participants of a targeted product, for not having enough control over experimenter bias, and for the 'environment effect', which might produce results that cannot be transferred to real usage (Reips, 2000); (Harrison et al., 2004); (Beynon-Davies, 2002); (Fraenkel and Wallen, 2012); (Denscombe, 2010); (Lazar et al., 2010).

In spite the above and due to the aim of the current research study, this method has been adopted. This is largely because usability studies tend to adopt this approach for measuring the usability of a product (Kirk, 1982); (Oehlert, 2000); (Lazar et al., 2010). Consequently, it is used in this research during the building of the adaptive framework, during the testing of the generated methods (DSI) on the targeted products, during the testing of the research hypothesis.

4.3.2 Observation research

Observing the behaviour of users in a controlled environment is most often used within user testing in laboratory experiments. It aims to measure how users interact with the product. Data recording techniques, such as video-recording or screen-capture photographs, are employed but the way in which these techniques are used is different. The Think Aloud technique is useful for understanding what the users are thinking, and so it is commonly used with observation and testing methods. There are two kinds of observation for tracking users' activities: direct observation and indirect observation. The former occurs when the observer is seated with the users in the same environment, where s/he can conduct interviews or questionnaires with the users after the experiment, whereas the latter occurs when the observer cannot be present for the duration of the study. There are two methods to assist in conducting indirect observation, which are diaries and interaction logs (Rogers et al., 2011). Accordingly, direct observation is adopted in this research in order to understand every error and difficulty for every user during the experiments.

4.3.3 Case study

A case study is an empirical inquiry that helps in an intensive investigation, an in-depth examination and observation of events and phenomena within a real-life context to understand, gather information and analyse why and how these are occurring in social science settings or natural settings (Gerring, 2007); (Bryman, 2012). Crowe et al. (2011, p.1) defines case study as "a research approach that is used to generate an in-depth, multi-faceted understanding of a complex issue in its real-life context". This method has advantages and disadvantages; for example, (Benbasat et al., 1987) outlined two advantages of this method, which are that it helps to study the phenomenon of interest in its natural setting, and helps to find answers of more questions through observing actual practice. However, Meredith (1998) refuted these advantages because it requires direct observation in the actual contemporary situation, which in turn leads to an increase in access hurdles, and it entails greater costs and demands more time than other methods. Also, it is necessary to use multiple methods, tools and entities for triangulation, which can help in comprehensively grasping the nature and complexity of the complete phenomenon. Moreover, it suffers from complications of context and lack of control; there is also lack of understanding of its procedures (Meredith, 1998). Due to the nature of the current research study, this method would be appropriate to use in a narrow range when three websites were chosen from educational and social network domains.

4.3.4 Action research

This approach is a subset of the case study and field study (Antill, 1985). It is context specific and it involves direct action on the part of the users and the researcher, as a collaborative process, for the investigation and diagnosis of a problem (locally, in a specific case) when evaluating and improving a product (Fraser, 1998). Somekh (2005, p.34) defined action research as "the study of a social situation, involving the participants themselves as researchers, with a view to improving the quality of action within it". The advantages of this method are that it can be set within a specific context or phenomenon, it includes continuous evaluation and modifications that can be made as the project progresses, and it allows theory to emerge from the research, rather than always following an earlier formulated theory. However, it has limitations, such as it depends on research ability in determining the parameters of the study at the start (Koshy, 2005). Also, it has been criticised for ethical

reasons because it introduces the researcher as an extra factor during the data manipulation (Galliers, 1992). Also, Bennett (2004) pointed out another drawback, which relates to the inherent relationships that external researchers have with the local individuals that are hired to assist in the research process. For above reasons and for its similarity to the idea of case study, it too used in a narrow range when the researcher was involved in the research process to create the DSI method for each domain.

4.3.5 Descriptive/interpretive research

This method focuses on reading the available literature, past developments and actual current happenings on the topic under investigation. It can be seen as an in-depth review, which can assist a researcher in becoming familiar with previous knowledge and in obtaining new information that can contribute to his/her knowledge, thereby giving him/her the ability to further understand a research problem and to develop a theory. The advantage of this method is that it has the ability to represent reality; however, it depends on the researcher's skills and their ability to identify their biases and assumptions (Punch, 2013). In this research, this method is adopted to review the literature related to the topic being dealt with.

4.3.6 Interview and questionnaire

The interview method can be used as a quantitative research tool if its questions are designed as closed, and can be used as qualitative if its questions are designed as open. This method can be used with a single person or with small groups (e.g. a focus group). This method needs an appropriate environment and an interviewer who has good knowledge and skill in guiding the session. The questionnaire is another method that consists of set of mixed question types (e.g. closed and open). The types of questions employed in the interview and the questionnaire can be determined based on the aim of the research. In this research, both methods are employed at differing steps, including closed and open questions (e.g. context meeting interview, focus group, pre-test or post-test questionnaires).

4.3.7 Triangulation

Triangulation refers to employing more than one data-gathering approach to tackle a goal, or using more than one data analysis method to investigate the research question in order to enhance confidence in the subsequent findings. It has synonyms such as integration, blend combination or mixed method (Rogers et al., 2011). Shih (1998, p. 632) defined triangulation

as "combination of two or more theories, data sources, methods, or investigation in one study of a single phenomenon". Webb et al. (1966, p.35) suggested that "once a proposition has been confirmed by two or more independent measurement processes, the uncertainty of its interpretation is greatly reduced. The most persuasive evidence comes through a triangulation of measurement processes". Also, Elliott and Timulak (2005, p.151) describe it thus, "it can yield a richer and more balanced picture of the phenomenon, and also serves as a cross-validation method". There are many forms of triangulation (e.g. investigator triangulation, theoretical triangulation and methodological triangulation), and methodological triangulation is the one adopted here. It refers to the use of more than one method for collecting data (Bryman, 1992). The advantages of this method are that it enhances the validity and reliability of the findings, it examines and investigates the phenomenon from different perspectives, and it constructs a wider and clearer picture of the phenomenon under study (Gall et al., 2009). In this study, the researcher employs a blend of quantitative and qualitative methods (as triangulation) in order to check the validity of the DSI findings by cross-checking them with other methods.

4.3.8 Content analysis

There are many methods to analyse qualitative data. The thementic analysis method is one example. Braun and Clarke (2006, p.79) defined thematic analysis as "a method for identifying, analysing and reporting patterns (themes) within data". Braun and Clarke (2006, p.97) pointed out a disadvantage of this method when they stated "it makes developing specific guidelines for higher-phase analysis difficult, and can be potentially paralysing to the researcher trying to decide what aspects of their data to focus on" . Another method is grounded theory. Engward (2013, p.37) defines grounded theory as "a systematic research approach involving the discovery of theory through data collection and analysis". This method produces large amount of data which makes it difficult to manage, and there are no standard guidelines for the identification of categories. Also, it requires high skills on the part of the researcher using it (Olesen et al., 2007). Content analysis is another method. Downe-Wamboldt (1992, p.314) defined content analysis as "a research method that provides a systematic and objective means to make valid inferences from verbal, visual, or written data in order to describe and quantify specific phenomena". This method has been used as a quantitative and/or qualitative method. It can be used in deep interviews, focus group

interviews, one single written question, open-ended questions, questionnaires, observations, pictures, and films (Bengtsson, 2016). Bengtsson (2016, p. 10) points out that "in quantitative content analysis, facts from the text are presented in the form of frequency expressed as a percentage or actual numbers of key categories. In qualitative content analysis, data are presented in words and themes, which makes it possible to draw some interpretation of the results". This method is useful for generating new knowledge, and it is commonly used in the HCI field (Lazar et al., 2010). In this research, content analysis is adopted. Thus, the data collected from the questionnaire, focus group, mini-user testing, user testing, structured usability problem report, and literature review was analysed using the content analysis method and the result was used to produce usability areas, heuristics and checklists.

4.4 Research design and procedures

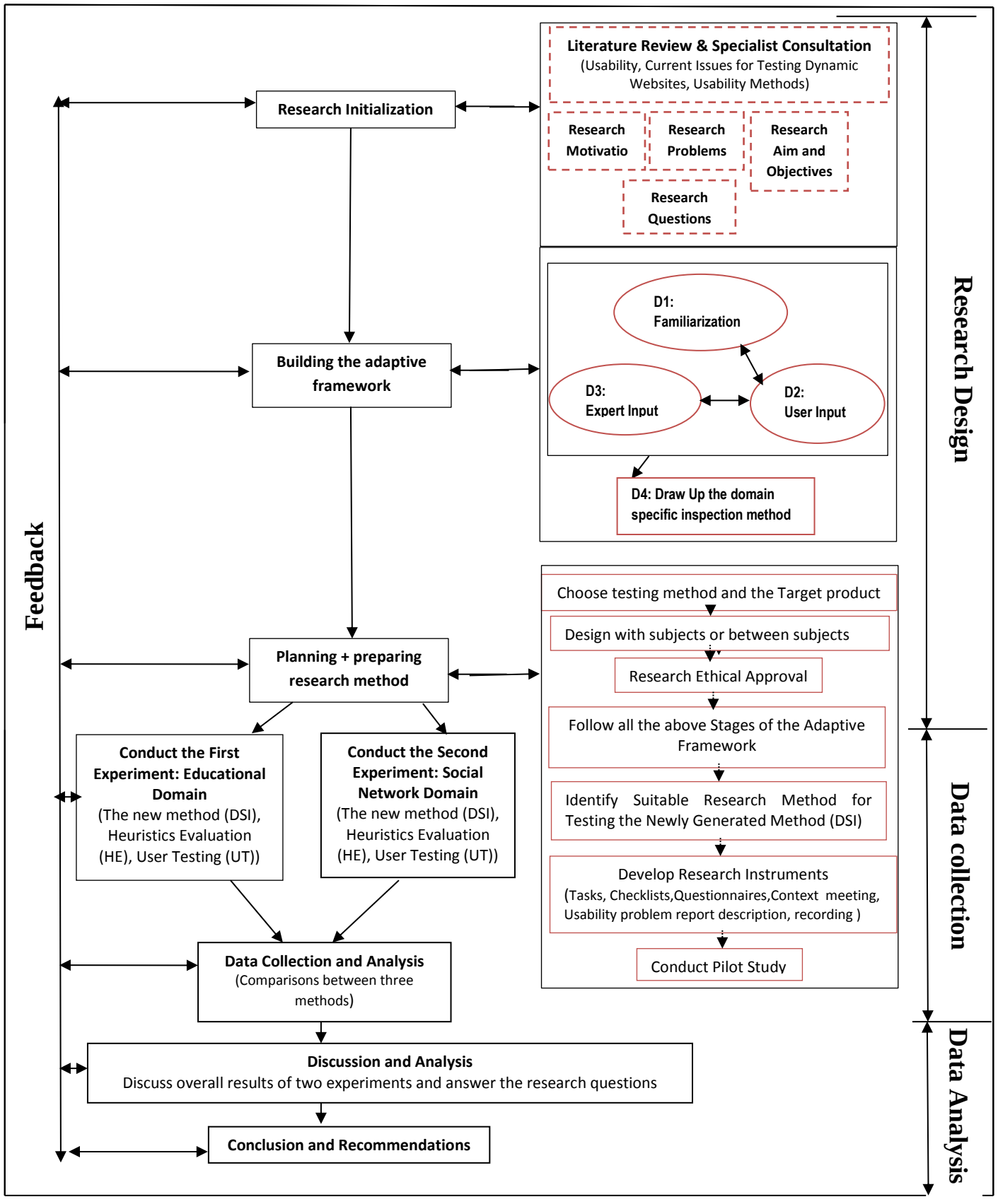
Research design is a part of research methodology; it refers to the whole strategy or structure that is chosen to integrate the various components of the research in a coherent and logical manner. It includes addressing the research problem, research question, explaining data collection, measurement methods, and the analysis of data (De Vaus and de Vaus, 2001). It is defined by Punch (2013) as "the fundamental plan for a piece of research, which contains major ideas of the research, such as the framework of the research, and presents which tools and procedures the researcher will use to collect and analyse the research data. It simply describes the research from its problem to reporting the research results". Collins and Hussey (2013) pointed out that the research philosophy should assist the researcher in choosing the correct research design. Lazar et al. (2010) highlights the research design procedures that are typically used in the lifecycle of an HCI experiment, which are; 1) Identify a research hypothesis; 2) Specify the design of the study; 3) Run a pilot study to test the design, the system and the instruments; 4) Recruit participants; 5) Run the actual data collection sessions; 6) Analyse the data; and 7) Report the results. In this regard, there are different kinds of research design, and the experimental design is adopted here based on the research aim and question. It refers to a written plan of the procedure that allows the researcher to maintain control over all the factors that may affect the results of an experiment. The above points of Lazar et al. (2010) are adopted, and the figure below (Figure 4.1) illustrates the research design of this study; each stage will be explored in detail. It is kept updated in case of any deviation or enhancement in the design. It includes all the stages from familiarisation to

conclusion, but is divided into three reliable research stages as proposed by Janesick (2000), such that the research design refers to the development stage and the application stage, and that the data collection and data analysis refer to the evaluation stage. The research progresses in the manner designated by the small dark arrows in Figure 4.1, meaning that each stage is initiated after the previous stage is completed. The double arrows show the feedback process and the possible backtracking process, for example, to compare the findings with others studies in the literature. Another example is that the researcher might need to enhance the instruments according to the pilot study findings.

4.4.1 Literature review

A literature review is fundamental to the current research as it helps to understand the previous and present contributions in the HCI field, and to identify the areas that still require further research attention, in order to fully grasp the usability of dynamic websites, to enhance and understand more deeply the topic under investigation, to define the key terms and definitions, to establish frameworks, models, instruments and theories, to select the appropriate research methodology for collecting and analysing the data, and (last but not least) to develop the research instruments. The literature review, particularly on the current usability issues that are presented for discussion on the Internet and in books, online database, conferences and journals, assisted the researcher in identifying a specific problem that is in need of more investigation and in concentrating on this in order to formulate the research questions, aim, objectives and research question, and to construct the proposed adaptive framework; the researcher was also assisted through consultation with specialists (Doctoral Consortium) (Oppenheim, 2000); (Janesick, 2000); (Lazar et al., 2010).

Figure 4.1: Research design



4.4.2 The target product

In this research, the free educational websites and social network websites have been chosen as the targeted product. These websites have been chosen as good examples of dynamic websites. The researcher sought to ensure that the selected websites represented each domain to make the validation experiments for the adaptive framework more valid. The selection process was criteria-based, based on six aspects that were determined and verified for each website, and these were: 1) representative of the chosen domain based on clear definition and classification in the literature; 2) popularity based on the number of users based on the statistical studies in the literature (for easy recruit the sampling) ; 3) free to join website; 4) the website is relevant to the subject of study and direct to the scope of the chosen domain; 5) rich functionality and different features, for example, at least four modules and four features for free educational websites, and four features for social networks websites; and 6) not familiar to the users in the testing session. In order to achieve a high level of quality in this study, the researcher selected five websites for each domain based on the above aspects and classification in Table 1.1. These websites are CosmoLearning, SchoolsWorld, Skoool, AcademicEarth and BBC KS3 Bitesize in the domain of free educational websites, and MySpace, Flickr, LinkedIn, Google+ and Ecademy in the domain of social network websites. Then, a consultation session was conducted with two experts in each domain to make sure that the chosen websites were appropriate and representative. The aim and objectives of this study, the above six aspects, the classification in Table 1.1, and the selected websites were sent to the experts before the consultation sessions. During the session, each website was checked with each aspect to make sure all aspects were met. Finally, all experts agreed on the selected websites.

CosmoLearning is a non-profit educational website and it is designed for developing the quality of homeschooling, teaching and student distinction (CosmoLearning, 2012). SchoolsWorld is a free online and it is designed to for everyone involved with or wanting to be involved in schools (SchoolsWorld, 2012). Skoool is an Intel driven initiative that delivers highly innovative and interactive learning resources via cutting-edge technologies and devices (Skoool, 2012). AcademicEarth is an organization founded with the aim of giving everyone on earth access to a world-class education (AcademicEarth, 2012). BBC KS3bitesize helps school students from 11 to 14 with their coursework, homework and test

preparation (KS3bitesize, 2012). Wikipedia (2016) defines MySpace thus "a social networking website offering an interactive, user-submitted network of friends, personal profiles, blogs, groups, photos, music, and videos''. Flickr is a social networking website where people can upload images and videos and share these images and videos around the world (Flickr, 2012). O'Murchu et al. (2004, p.4) defined LinkedIn as, "it was founded in May 2003; it focuses on professional users creating networks of co-workers and other business associates. It allows members to look for jobs, seeking out experts in a particular area, or to make contact with other professionals through trusted connections. It is probably the site with the least potential for social purposes". Magno et al. (2012, p.159) define Google+ as "a new generation of social network and includes several new features, such as "circles'' that allow users to share different content with different people and "hangouts'' that let users create a video chatting session and invite up to nine people from their circles of friends to share the environment". O'Murchu et al. (2004, p.3) defined Ecademy as "a business-networking site built up of trusted business connections for people to share contacts and business opportunities. It is free to join, however membership can be upgraded for a monthly fee. It provides a catalogue of Ecademy clubs, and calendar functionality for meetings and events. It provides a geographical list of networking regions globally for arranging meetings and events offline".

In addition, these websites were discussed with two experts in each domain to make sure that the chosen websites represents the above aspects. Thus, all of these websites have all the aspects mentioned above; this should make the research more focused and ensure that the results are representative. In this regard, CosmoLearning and SchoolsWorld in the domain of free educational websites, and MySpace and Flickr in the domain of the social networks websites, were chosen for the mini-user testing in step two (User Input) of the adaptive framework. Furthermore, Skoool, AcademicEarth and BBC KS3 Bitesize in the domain of free educational websites, and LinkedIn, Google+ and Ecademy in the domain of the social networks websites, were chosen for the validation experiments. Contact was made with each of the website owners in order to obtain their permission to subject their websites to an evaluation.

4.4.3 .Dependent and independent variables

Lazar et al. (2010, p.25) defined the dependent and the independent variables thus, "independent variables refer to the factors that researchers are interested in studying, and dependent variables refer to the outcome or effect the researchers are interested in". In the HCI field, independent variables are usually related to methods, users and the context of use (of the technologies under investigation). Also, dependent variables are frequently used to measure efficiency, accuracy, subjective satisfaction, learning and retention rate, and physical or cognitive demand (Lazar et al., 2010). In this research, the independent variables are methods with three levels which are 'HE, DSI and UT'. The dependent variables are time spent and number of problems encountered.

4.4.4 Design with subjects or between subjects

A fundamental characteristic of experimental approaches is employing three designs based on the research questions. The first one is within-subjects, the second design is a between-subjects design, and the third design is a mixed factorial. The first indicates that each participant performs under all sets of conditions. For example, a study that focuses on conducting comparisons between three websites and that allows the same participants to use all three websites would be within-subjects. However, if it allows each participant to use only one website, that would be a between-subjects design. The last one means that one independent variable is within-subjects and another between-subject (Cairns and Cox, 2008). There are advantages and disadvantages to each approach. For example, within- subjects may lead to spurious effects, also known as a 'demand effect' or 'carryover effect', such as improving or decreasing performance. An improvement in performance can happen as a result of practice (learning effect), whereas a decrease in performance can happen because of fatigue. These can occur because the participants repeat very similar procedure multiple times, and also these can affect any other independent variables. On the other hand, it has three advantages, which are that their internal validity does not depend on random assignment, they offer a substantial boost in statistical power in many frameworks, and they are more naturally aligned with most theoretical mindsets. Also, this approach requires small sample sizes, easing participant recruitment and thus reducing the cost of the experiment. Additionally, observing a small sample size helps to isolate the effect of individual differences because each participant is being compared with himself/herself in respect of a

number of experiment conditions. Furthermore, the between-subjects design is described as a cleaner one because each participant is only exposed to one condition, and thus there are minimal learning effects. It is also claimed to include more external validity and it is statistically simple to perform. Furthermore, it is not overly time-consuming and this leads to the effective control of confounding factors such as fatigue or frustration, which may affect participants after performing many tasks on different websites. Furthermore, it removes the carryover effects that can happen if two groups evaluate the same website. However, it requires a large sample size to exclude the effect of noise and to deliver significant results (Cairns and Cox, 2008); (Charness et al., 2012); (Albert and Tullis, 2013).

For user testing in both experiments in this research, the between-subjects design is used. This means that each group participated in one condition only, and this helps to avoid order effects, for example, practice or fatigue. Thus, in the first experiment there were three user groups assigned to the three educational websites. The users in each group were randomly assigned. This was the same for the second experiment in which three user groups were assigned to the three social network websites. The users in each group were randomly assigned. For comparison between heuristic evaluation (HE) and DSI methods (in both experiments), the within-subjects design was used. Accordingly, the expert evaluators in the first experiment evaluated three educational websites. This was the same for the second experiment in which three social networking websites were evaluated by other evaluators. To reduce the effects of order, the evaluators in both experiments were randomly assigned into two sequences, which were DSI, HE, DSI in the first time and HE, DSI, HE I the second time. Furthermore, extraneous variables were controlled at an optimum level to eliminate their effect on the dependent variables. For example, the situational variables were controlled by using standardised instructions, standardised procedure, and standardised environmental conditions for all of the participants.

4.4.5 Research ethics approval

The most important step in scientific research is to obtain ethical approval or to consider the ethical dimension prior to commencing a study and collecting data or recruiting participants. The relationship between the participants who provide the data and the researcher who collects those data should be clear and professional in order to clarify the nature of the study,

and this is achieved through ethics Rogers et al. (2011). McDaniel and Gates (2004) argue that the researcher must "identify the ethical principles related to personal and corporate behaviour in specific business situations and explain the potential consequences". Thus, research ethics aim to clarify the research objectives and protect the participants by respecting their rights and dignity in expressing their desire to participate (or to reject), which should result in potential benefit and in minimizing the risk of harm; the participants should also be fully informed as to who has access to their information (Denscombe, 2010). If a researcher wants to gather private or sensitive information from participants, they must be reassured that their data will be kept confidential and safeguarded from publishing; they must also be promised anonymity (Corti et al., 2000).

Blandford and Green (2008) summarized important aspects of the ethical dimension, which are vulnerable participants, informed consent, privacy, confidentiality and maintaining trust. Vulnerable participants refer to young, old and infirm groups, but also to any participant who is unable to refuse the invitation to avoid upsetting his/her friend or supervisor. So, the participants should be clearly informed that the study is going to assess the system and not them. Furthermore, the participants should be informed that they are volunteers and they have the full right to withdraw from an experiment under any circumstances and at any time without giving a reason. The participants' data should be stored securely and systematically. For the important reasons mentioned above, research ethics were included as a step in the research design (as in Figure 4.1).

4.4.6 Developing Research Instrument

Developing the test materials requires careful consideration because these materials will support the goal of the testing sessions and thus will collect the data needed to answer the research questions (Rubin and Chisnell, 2008). In this regard, the test materials were prepared based on the research aim and objective. These materials for evaluation (DSI/HE) and testing (UT) are explained in detail below

4.4.6.1 Orientation script and consent form

This step, also known as the introduction script, is very important; it includes the following points: introduces the researcher, describes the aim of the research study for the participants,

details what they will do during each test/evaluation session, describes what is expected from them, explains the testing procedures, emphasizes that the product is being tested (not them), encourages the participants by telling them that they are the right people in the right place, informs them how long each test session will take, tell them that they have the right to stop completing a task if they feel that they are unable to accomplish it, and confirms that they may withdraw from the testing session at any time. Finally, the researcher must ask the participants to read and sign the consent form prior to the testing sessions, which confirms that a participant agrees to be involved in the testing session and is aware of any risks that might be involved (Rubin and Chisnell, 2008). Accordingly, these materials were developed and used, as shown in Appendix A1, A2, A3 and A4.

4.4.6.2 Pre-test questionnaire

The pre-test questionnaire was used after the participants (evaluators and users) had agreed to participate and had signed the consent form. It was developed to provide historical information about the participants in order to better understand their behaviour and performance; it includes data on their profile and experience. Also, it asks the participants about their impression of the target product, to establish their level of experience and then to distribute them into specific groups in an equitable manner. This information can influence their results positively or negatively. Consequently, collecting and understanding that information will help to interpret the testing results based on their performance and behaviour. It was designed to include two sections, which are background experience and experience of educational/social network websites, as adopted and modified from Brinck et al. (2001) and Rubin and Chisnell (2008), as shown in Appendix B1, B2, B3 and B4.

4.4.6.3 Post-test questionnaire

The post-test questionnaire was developed to gather feedback from the evaluators on the methods (DSI and HE) and on the structured report (Sauro, 2010), as shown at Appendix B5. For UT, it was developed to gather preference information from the participants after they had finished a testing session, including identifying the problem areas on the target websites (Sauro, 2011b), as shown at Appendix B6. Also, it consists of a satisfaction scale from 1 to 7, where 1 refers to ‘highly unsatisfactory’ and 7 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). Also, some open questions were developed to encourage the participants to answer in detail using their knowledge and feelings, rather than choosing from a predetermined list, as recommended by Rubin and Chisnell (2008).

4.4.6.4 Observer and data recording

The most commonly used method with user testing is to observe people interacting with websites. It is usually conducted in a controlled environment through taking notes, and through audio- and video-recording, which can be used separately. It needs an observer who is able to manage different tasks, such as helping participants to solve a problem, knowing what they are doing and what difficulties they face (and how they can succeed), writing down all the participants' comments, monitoring their behaviours, and understanding the users' contexts and goals. There are a number of tools that can be used to facilitate these tasks, such as videotape, data logging and Google Analytics. Taking notes is adopted in this study, depending on the time, context and the sensitivity of the situation. Therefore, all the testing sessions were observed by the researcher, who used a prepared observation sheet and recording permission form, as shown in Appendix B7 and B8 (Rubin and Chisnell, 2008); (Rogers et al., 2011).

4.4.6.5 Context meeting

Thomas and Bevan (1996, p.2) stated, "analysis of context is an essential prerequisite for any work on usability". Maguire (2001a, p.458) summarized the benefits of 'context meeting' thus, "provides an understanding of the circumstances in which a product will be used, helps to identify user requirements for a product, helps address issues associated with product page usability, provides contextual validity of evaluation findings, and it also provides a system focused approach which leads to a shared view among the design team". In this regard, the context meeting aims (in this study) to glean a coherent picture of usability in the given context by collecting information on how the selected products are used, by defining the users' characteristics for selected products, identifying the usability requirements for these targeted products, and specifying what tasks are to be performed by the users on the targeted product (to create the tasks that will be used in the mini- user testing and UT experiments). Thomas and Bevan (1996) listed the people who should be involved in a context meeting:

user representatives, designers, stakeholders and human factor professionals. In terms of the time and money for this research, user representatives and designers were invited to be involved in the context meeting as shown in Table 4.2. The context meeting was conducted with five participant, all participant were willing to take the time to be involved after receiving a recruiting email (see Appendix G). Also, an email was sent to the owners of the targeted products, as in Appendix B9. After gaining their approval to take part in this meeting, the brief materials were sent to them to increase their knowledge and to reduce the time needed for the meeting, including the agenda for the context meeting, as shown in Appendix B10.

Table 4.2: Distribution of participant profiles for context meeting

Focus group	**Participant identification**	**Participant type**	**Participant characteristics**	**Level of education**	**Years of experience of work/ (using websites for user only)**
Context meeting group in the first experiment	1	Expert	Designer	Master	6
	2	Expert	Designer	Bachelor	4
	3	Expert	Designer	Master	7
	4	User	Real user	Master	5
	5	User	Real user	Master	10
Context meeting group in the second experiment	1	Expert	Designer	Master	7
	2	Expert	Designer	Bachelor	4
	3	Expert	Designer	Bachelor	6
	4	User	Real user	Master	9
	5	User	Real user	Master	8
Total	10		Mean (years)		

4.4.6.6 Task scenarios

This step is the backbone of the user testing method. The tasks should be constructed in terms of the aim of the UT sessions and must also be appropriate and realistic for the end–users; there should be a minimum of 3 to 5 task scenarios (Sauro, 2010). The task scenarios were developed as typical tasks for each of the ten studied websites as shown in Appendix C, D, K, and O. These tasks are representations of the actual work that is typically performed by participants when they are using the target websites, based on the result of the context meeting. There are different ways to develop tasks such as user observation, user story, and use case. However, the context meeting approach was used with experts and real users as mentioned in Section 4.4.6.5. The recommendations of some scholars were taken into consideration during the task design phase, as mentioned in the literature review chapter (Nielsen, 1993); (Dumas and Redish, 1999); (Snyder, 2003); (Sauro, 2010). Consequently,

the tasks were built based on the website's goals, and covered all the main functions of the target website, including searching features, interactivity and participation features, and display of records. Also, they were designed to be short and clear, and in the users' language. Sauro (2011c) emphasizes the importance of defining the task scenarios based on the context of use. So, the context meetings were conducted with user representatives and designers of the target websites in order to gather information on those websites and their intended context of use, as mentioned previously. Understanding this information can help to build realistic task scenarios based on the users' context of use (Thomas and Bevan, 1996). Furthermore, the tasks were simple and involved limited cognitive processing to avoid the individual differences in the between-group design, as recommended by (Lazar et al., 2010).

4.4.6.7 Usability problem report description

The structured report was developed to help the evaluators and the observer to report their results in a professional manner. It was designed, as shown in Appendix E, to solve the problems discussed in the literature review. Its components were adopted from a few distinguished studies that discussed this issue, such as (Lavery et al., 1997); (Cockton and Woolrych, 2001); (Cockton et al., 2004b); (Hertzum, 2006); (Hornbæk and Frøkjær, 2008). It consists of five attributes. The first is a numeric identifier of the problem; it ascribes identification numbers to the discovered problems. The second attribute is a heuristic name; it refers to the heuristic that discovered the problems that were violated by the design in each circumstance, in the judgment of the evaluator (Nielsen, 1995a). The third attribute is the problem description; it describes what is wrong and what needs repairing, justifying why it is problematic (Lavery et al., 1997). The fourth attribute is the problem context; it describes the context of the discovered problem, such as when the problem occurs and where, its impact, and its solution. Impact refers to the discovered problem being easy or difficult for the users to overcome. Hertzum (2006) assessed the impact of discovered problems based on the proportion of users who would experience them, thus "(1) No problem, (2) Minor problem, causing a brief delay, (3) Serious problem, causing a significant delay (but users eventually complete their task), and (4) Disaster, causing the users to voice strong irritation (unable to solve the task, or solve it incorrectly)". Hertzum (2006) used a ranking on a three-point scale, thus "(1) Users quickly learn to get around the problem; (2) Users only learn to get around the problem after encountering it several times; and (3) Users never learn how to

get around the problem". The fifth attribute is the problem area; it determines how the discovered problem relates to any usability problem areas, based on the DSI method (5 usability problem areas in educational domain, and 7 usability problem areas in social network domain) and the ten usability heuristics method. Also, it aims to assist in the problem matching phase (Hornbæk and Frøkjær, 2008). The sixth attribute is problem severity; it aims to classify and prioritize the severity of each discovered problem (by the evaluators) by using Nielsen's scale as mentioned in the literature review. This report was submitted by each HE and DSI evaluator to the evaluation manager, who is the researcher in this study.

Furthermore, this report was used by the expert evaluators during the evaluation sessions, but they were not allowed to fill in the sixth attribute. Also, the observer in the UT session was not allowed to fill in the second, fifth and sixth attributes. Moreover, Nielsen (1992a) recommended to use two or three raters to rate the usability problems, and he justified this by stating that ''as more evaluators are asked to judge the severity o f usability problems, the quality of the mean severity ratings increases rapidly, and ratings from three evaluators would seem to be satisfactory for many practical purposes''. In line with these recommendations, two independent evaluators were used in this research. They were involved to rate the problems reported by the evaluators, and also were involved in analysing, classifying and rating the problems reported by the observer. This technique was adopted based on recommendations to reduce the evaluator effect, to improve the reliability of the merged usability problems as well as the reliability of the ranked and matched predicted problems to the actual problems (not by the evaluators who made the predictions or the observer who reported the users problems); these steps increase the overall internal validity of the usability evaluation results (Lavery et al., 1997); (Cockton et al., 2004b); (Hertzum et al., 2014).

4.4.7 Pilot Study

The materials above need to be tested before using them, and in this regard, conducting a pilot study is crucial in any research; it is also known as a feasibility study. A pilot study is a relatively small experiment, usually no longer than an hour, and is conducted before the actual experiments with a very small number of participants (from the target population); they are given breaks at realistic intervals. It has many advantages, for instance, it improves the quality of the proposed experiment, checks the experiment of procedures and instructions

given to participants, ensures the environment and equipment are functional and safe, refines the research question, tests logistical issues and collects data, using those data in statistical tests to check for reliability and validity (Lancaster et al., 2004). Also, the suggestions obtained from the pilot study should be incorporated into the main study design, in particular if the pilot study does not lead to any modification of the materials or procedures (Cairns and Cox, 2008). The pilot study could be conducted twice to assess the revised main study (Ruxton and Colegrave, 2011). In this research, this step was adopted in the research design to verify that the new method is viable before starting on the real study, with all the materials, instruments, measures and procedures (see Figure 4.1). In this pilot, two independent evaluators and fifteen users were involved in each experiment. All the materials were checked by them to make sure that there were no spelling or grammatical errors and no ambiguous words or phrases, and that all of the sentences in the instruments (method descriptions, check-lists, time taken completing the task scenarios, questionnaires and procedures) were sufficiently clear to be used by the evaluators and users. 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 to the real test. Also, it attempted to identify what equipment the users regularly use and set it up for them before the test, for example, using the same type of machine and browser.

4.4.8 Components for testing the adaptive framework

Having constructed the adaptive framework, it tests intensively through rigorous validation methods (triangulation) to verify the extent to which it achieves the identified goals, needs and requirements that the method was originally developed to address. The testing methods for the adaptive framework were chosen here based on website usability, as the chosen products. In other words, the steps of the adaptive framework, that are stated in Chapter 3, are fixed; however, the testing methods are changeable, depending on the product being evaluated. For example, for mobile applications, the testing methods should be field studies or hands-on measurements alongside user testing. Thus, the triangulation method is used to check the validity of the DSI findings by cross-checking them with other methods (HE and UT). These methods have been chosen because they complement each other, they have been

commonly used in the evaluation of website usability, and they are able to identify usability problems from two different perspectives (Chen and Macredie, 2005). This validation process is outlined in Figure 4.2; the testing components of the adaptive framework consist of four steps, as outlined below.

- T1: 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 websites that are directed to the scope of the chosen domain; 2) Recruit expert evaluators and users; 3) Plan the sequence for conducting the evaluations (for each group) in such a way as to avoid any bias; and 4) Prepare the experimental documents (e.g. context meeting, task scenarios and questionnaires). 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.

- T2: Heuristic Validation stage (Expert Evaluation (HE)): The aim of this stage is to validate the newly generated DSI method by conducting a heuristic evaluation (HE). This method has been chosen because it is the most common inspection method used in evaluation testing. Expert evaluators need a familiarization session before the actual evaluation. The expert evaluation is then conducted using the newly generated DSI method alongside HE. The aim of this process is to collect data ready for analysis (analytically), as explained in stage 4.

- 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 (UT) on the same product. Jeng (2005) pointed out that there is a need for benchmarks in order to compare the methods, and the results of the user testing (UT) method represent the best means for comparison. Also, Nielsen (1992a) recommends conducting UT with HE because each one is complementary to the other. Hartson et al. (2003, p. 385) stated, "it is best to combine the lab test with expert review to eliminate some of the problems considered not real, thus improving the quality of the usability problem set to be used as the actual criterion". Also, they added that UT is the gold standard for comparison, as it is used overwhelmingly in evaluating studies conducted on the performance of UEMs (Hartson et al., 2003). In this regard, the performance of the newly generated DSI method is compared

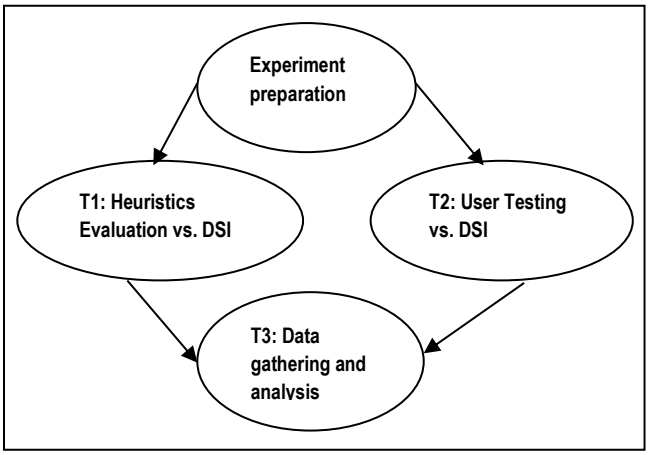
with the lab testing to identify which problems have been identified by UT and not identified by DSI and/or HE, and vice versa. The aim of this process is to collect data ready for analysis (empirically) in the final stage.

- T3: 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 vs. DSI and the other on UT vs DSI and HE. The researcher extracts 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 and two users to agree on the discovered problems, and to remove any duplicate problems or subjective problems. Then, the problems approved upon are merged into a master problem list for each of them, and any problems upon which the evaluators disagree are removed. After that the independent evaluators are involved to rank the severity of the problems derived from the HE and DSI. Ultimately, the researcher 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 experience, and the relative costs entailed in employing the two methods.

  In the second step, the researcher conduct a debriefing session with independent evaluators to rank the severity of the problems derived from the UT and to remove any duplicate problems. Following this, they establish a master list of usability problems for UT. Subsequently, a single unique master list of usability problems are consolidated from the three methods. After that, the falsification test will be conducted on the not matching predicted problems from HE and DSI to the UT problems. This test aims to investigate whether all HE and DSI problems are real problems or false problems through conducting falsification test based on these problems. If they are real problems, they will be moved to the single unique master lists of HE and DSI; otherwise, they will be removed permanently (see Section 4.4.9.2). Next, a comparison of the results of the three methods is 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, which are their relative costs, reliability, efficiency, effectiveness, validity, thoroughness. Moreover, this final step seeks to prove or refute the efficacy of conducting UT and HE with DSI, and vice versa.

Figure 4.2: Testing stages of the adaptive framework



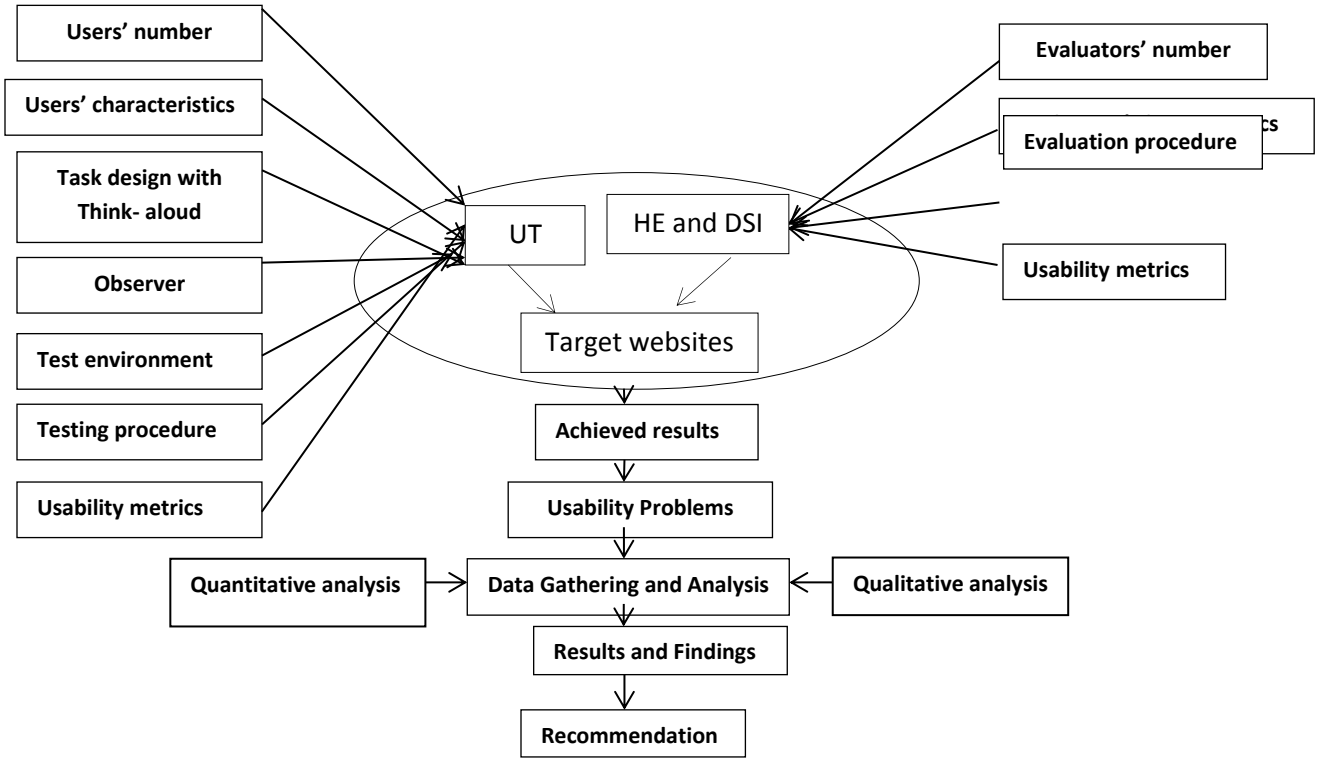
4.4.9 The first and second experiments

In order to achieve the aims of this research, a series of experiments must be conducted. The first experiment examines practically the efficiency of the adaptive framework in generating the domain specific inspection (DSI) method for evaluating the educational domain. Then, this method is evaluated analytically and empirically to measure its efficacy, effectiveness and satisfaction. This step aims to reveal whether or not the DSI method is better than UT and/or HE in discovering real usability problems as well as against different usability metrics. The second experiment is the second practical validation experiment for the adaptive framework. This experiment aims to generate a DSI method for evaluating the social network domain. Then, this method is also evaluated analytically and empirically for the same aim as the first experiment. Overall, the two experiments investigate any correlations between the number of users or evaluators on the one hand, and the time spent and number of usability problems identified (and their severity) on the other. Within the context of this experiment, correlations between usability metrics are also inspected. Chapters 5 explain these experiments in full detail.

4.4.9.1 The approach taken for both experiments

Numerous evaluation methods that can be used to validate the adaptive framework were explained in depth in the literature review chapter. However, the approach taken for both experiments is similar. For both validation experiments in this study, user testing (UT) and heuristic evaluation (HE) were chosen to compare the efficacy and effectiveness of the newly developed method (DSI) for the educational and social network domains. The reasons behind choosing UT and HE as validation methods for the adaptive framework (and its resultant DSI method) are: firstly, they are the most commonly used usability methods employed for evaluating dynamic website usability and for identifying as many usability problems as possible; secondly, they complement each other (combining their results should provide a better picture for measuring the effectiveness of both the adaptive framework and the newly developed DSI methods for the educational and social network domains); and thirdly, UT and HE identify the usability problems from two different perspectives (expert and user) (Molich and Nielsen, 1990); (Jeffries and Desurvire, 1992); (Law and Hvannberg, 2002); (Rubin and Chisnell, 2008). Consequently, this mixture of three methods (triangulation) will help to maximize the opportunities for measuring the performance of each one, and to intensively test the DSI method through rigorous validation methods in order to verify the extent to which DSI achieves the identified goals, and satisfies the needs and requirements that they were originally developed to address. Also, the effectiveness of these methods is one of the research questions, and so a solid and comparable benchmark is needed. To further support and achieve the research aim, the following framework is utilized in all experiments (Figure 4.3).

Figure 4.3: Design for both experiments



- Heuristic evaluation (HE) and Domain specific inspection (DSI)

Heuristic evaluation (HE) is a usability inspection method that has been described as fast and widely used; it was explained in detail in the literature review chapter. This method was chosen to be one of the validation methods for the adaptive framework and for its ability to generate a DSI method for both domains (educational and social networks). Here, the requirements of this method are highlighted below.

- The need for heuristic evaluation

In order to evaluate the selected domains/websites (educational and social networks) and to obtain the best possible picture for the research results, the traditional ten heuristics were employed, as were developed by Molich and Nielsen (1990). These ten heuristics have been extended to various versions such as HE-Plus and HE++ (Chattratichart and Brodie, 2004); (Chattratichart and Lindgaard, 2008). However, Nielsen's heuristics are now widely adopted as the available tool; they are good heuristics for used in the early stages of website development, they identify more usability problems than most, and they are still widely taught and practiced in different areas (Gamber and Valent, 2001); (Manzari and Trinidad-

Christensen, 2013). For these reasons, Nielsen's heuristics and their explanations are adopted as shown in Appendix F. The following is an explanation for the components needed to conduct this method.

- Number and characteristics of recruits and evaluators

In fact, that there is no agreement on how many evaluators are enough to conduct this experiment. However, in order to quantify the number of evaluators needed for HE, Molich and Nielsen (1990) pointed out that evaluator experience plays a vital role in determining their number and in influencing the evaluation results. For example, there are three kinds of expert evaluators. One is non-specialist evaluators, who need training and a set of tasks to develop their skills and to minimize their influence on the result. The second is 'single' evaluators, who can be defined as persons with general usability experience. The third is 'double' evaluators, who can be defined as persons with both general usability experience as well as experience in a specific application area being investigated (Hertzum and Jacobsen, 2001); (Tan et al., 2009). In this regard, Molich and Nielsen (1990) recommended, from previous work on heuristic evaluation, that between five and fourteen non-specialist evaluators are necessary to find between 51% and 75% of all problems. Also, 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%). However, besides the effect of the evaluators' skills and level of experience, there are other factors that can also affect the evaluation results, such as any vagueness in the evaluation procedure, problem report, analysis, and problem criteria (Hertzum and Jacobsen, 2001).

In this research, there were limitations in terms of time, money and resources in recruiting ideal evaluators. Five people were recruited for the focus group for each domain (ten people in total), eight evaluators were recruited for the first experiment, and six evaluators were recruited for the second experiment. The recommendation of Tan et al. (2009) was considered with regard to involving double evaluators with single evaluators in order to increase the discovery of usability problems. Therefore, the eight evaluators of the first experiment included four single evaluators and four double evaluators, whereas the six evaluators in the second experiment included two single evaluators and four double evaluators. For the focus

group, The five people in each focus group were three experts in usability (i.e. having a certificate in the HCI field) and in the targeted websites (i.e. having certificate in the HCI field and targeted domain), and two users who were involved in the mini-user testing (real users for the chosen domains), see Table 4.3. The expert evaluators were invited to participate based on their availability and experience (convenience sampling). They were chosen carefully after emailing them to ask them to be involved in these experiments, as shown in Appendix A1. Some of them were recommended by my supervisor and are working at the University of East Anglia, and the others are from usability groups in LinkedIn and the Aviva life insurance company. All of them have knowledge of usability evaluation through teaching or studying HCI courses, and some of them have evaluated many websites for usability. These combined facts confirmed that the experienced evaluators were chosen in the hope of maximising the benefits of using expert evaluators in an efficient manner. Once they had agreed to be involved in this study (see Appendix A3 and A4), the pre-evaluation questionnaire, as shown in Appendix B1 and B2, was sent to obtain more information about them and thus to divide them into equal and balanced groups in terms of experience.

Table 4.3: Distribution of participant profiles for focus group

Focus group	**Participant identification**	**Participant type**	**Participant characteristics**	**Level of education**	**Years of experience of evaluation/ (using websites for user only)**
Focus group in the first experiment	1	Expert	Double	PhD	5
	2	Expert	Single	PhD	3
	3	Expert	Single	Master	4
	4	User	Real user	Master	5
	5	User	Real user	Bachelor	7
Focus group in the second experiment	1	Expert	Double	PhD	8
	2	Expert	Single	PhD	5
	3	Expert	Single	Master	3
	4	User	Real user	Master	6
	5	User	Real user	Bachelor	5
Total	10		Mean (years)		5.1

- Evaluation procedure

A formal procedure should be clearly described to inform the participants what they are to do during an experiment. Also, it is important to conduct all the evaluations under similar sets of conditions, including environment, equipment and tools (Hertzum and Jacobsen, 2001); (Chen and Macredie, 2005). In this regard, there are two benefits to preparing a formal

procedure; firstly, it ensures that each evaluator has the same documents (giving them different procedures would create a confounding variable); and secondly, it allows the other evaluators to replicate the same experiment, which can give more confidence to the research findings. For this method and prior to an actual evaluation, each evaluator undertook a training session and a certain task to become familiar with the chosen system and instruments (HE and DSI) after having conducted the pilot study with them. Also, each evaluator was asked to focus their inspections in the training session on the user tasks, as recommended by Hertzum (2006). Thus, these instruments should be the same for all the evaluators. After that, the actual evaluation session should be conducted under the same set of conditions and with the same instruments (Hertzum and Jacobsen, 2001). Therefore, the evaluation procedure was carried out as Nielsen (1995a) recommended on his website (Nielsen Norman Group). Therefore, the expert evaluators conducted their evaluations independently, and they were not allowed to communicate with each other until they had finished their evaluations. These steps were justified by Nielsen (1995a) "to ensure independent and unbiased evaluations from each evaluator". After that, they may join the debriefing session to aggregate and discuss their findings, generating a list of usability problems. At the beginning of each session, the evaluation procedures and instructions were explained to each evaluator as recommended by Nielsen (1995a). Consequently, the method sheet was given and explained to the evaluators, and they were asked to visit the interfaces of the targeted websites twice. The reason behind this is that the first visit helps the evaluators to gain a feel for the flow of the interaction as well as the general scope of the targeted websites. The second visit helps the evaluators to concentrate on particular interface features whilst comprehending how they fit into the larger whole (Nielsen, 1995a). Overall, this training time (to become familiar with the targeted website or methods) is used to reduce the so-called learning effect during the actual testing sessions (Lazar et al., 2010). After the exploration, the evaluators were asked to read and sign the consent form if they were happy to continue, or to sign the withdrawal form if they were not. Then, they were divided into two groups: one group for the first experiment and another group for the second experiment. Then, each group was divided into two groups. The evaluators in each experiment employed two methods, namely DSI and HE, to evaluate the three different websites. They were assigned randomly to two prescribed sequences (DSI, HE, DSI in the first time and HE, DSI, HE in the second time). The researcher adopted this technique to avoid any bias in the results and also to avoid the risk of

any expert reproducing his/her results in the second session through over-familiarity with one method, i.e. each evaluation was conducted with a fresh frame of mind. This technique also helps to control and reduce the learning effect and order effect on the within-group design. In this regard, the evaluators were not allowed to start the second or third evaluation until they had finished their evaluation of the previous website and had submitted their report. This technique also helps to reduce the potential problem of fatigue, which can happen in the within-group design. Both techniques were recommended by Lazar et al. (2010).

During the evaluation, the evaluators use the structured report developed by the researcher to help them report their results in a professional manner, as mentioned previously. After that, the post-evaluation questionnaire is filled out by each evaluator; it includes a rating scale questionnaire for measuring their satisfaction on methods, and an open-ended questionnaire for writing down their comments and feedback on the methods used. Finally, the researcher extracts the problems from the reports and removes all duplicate problems. After that, the debriefing session is conducted to agree on the problems found in order to create a master list of unique problems, and to discuss the results of the post-evaluation questionnaire. Then, the independent evaluators are involved in ranking the severity of the problems in the master list.

- User Testing (UT)

The user testing method represents the second validation method; it is used for evaluating the output of the adaptive framework, which is the DSI method. It is another important evaluation method for ensuring system quality, in particular for websites. It needs real end-users to perform a set of tasks. It can be defined as a procedure for integrating several distinct variables (for recognizing a website's defects in terms of design and usability), such as problem numbers, time spent and user satisfaction (Lindgaard and Chattratichart, 2007). Thus, UT needs clearly structured and organized test materials to be prepared to facilitate the test (and ultimately to support the goals of the test) and to obtain the data needed to answer the research questions effectively (Rubin and Chisnell, 2008). The following is an explanation of the components needed to conduct this method

- Number of users and their characteristics

The issue of determining the number of users to perform user testing is a controversial issue, as mentioned in the literature review (Section 2.4.1). Estimating the required sample size is important, particularly when the cost of a sample is expensive. In this study, the user numbers are identified based on the recommendations of certain pioneers and on the requirement of this study. For example, Dumas and Redish (1999) suggested that to conduct UT it is necessary to recruit from 5 to 12 users. Molich and Nielsen (1990) confirmed this number and they justified it by arguing they were able to reveal 85% to 90% of the usability problems. However, Nielsen (2006) recommended recruiting 20 users for each group in quantitative studies that need benchmarking in terms of factors such as efficacy, number of problems and errors, and subjective satisfaction. Also, Sauro (2010), based on the results of 120 usability tests, mentioned using 20 users because this number typically leads to a margin of error of approximately (+/-) 20%, and this is less than using 5, 10 and 15 users. Furthermore, Rubin and Chisnell (2008) recommended using more than 5 users to obtain statistically valid results. On the other hand, difficulties in recruiting users and lack of time, resources and budget during a study should also be considered (Rubin and Chisnell, 2008).

Based on the above clarifications, this study recruited 5 users for the context meeting for each domain (ten users in total as shown in Section 4.4.6.5); 10 users were recruited for mini-user testing for each domain (20 users in total as shown in Table 4.4); 20 users for each of the three groups in the first validation experiment (60 users in total as shown in Table 5.1); and 25 users for each of the three groups in the second validation experiment (75 users in total as shown in Table 5.2). These users were chosen carefully to reflect the real users of the targeted websites in each domain. The majority of the users were students and employees, and they were mixed across the three user groups in terms of gender, age, education level and computer skills. The criteria that were considered to recruit these users were: 1) real users for the targeted websites based on the context meeting result; 2) willingness to participate; 3) having good experience in the targeted websites by using similar websites in their daily life; and 4) for the context meeting, the participants should belong to a certain group of people, as identified by Thomas and Bevan (1996). These groups are user representatives, designers, stakeholders and human factor professionals. In this regard, user representatives (real users) and designers (in the chosen domain) were invited to be involved in the context meeting in

both domains, as mentioned in Section 4.4.6.5. Moreover, Sauro (2010) recommends encouraging users to participate and then thanking them after closing the debriefing sessions through offering vouchers as incentives for taking part in the experiment; thus this was adopted in the UT sessions in this study.

In terms of the users' characteristics, this issue is critical because recruiting unrepresentative users for the target product would lead to incorrect results (Rubin and Chisnell, 2008). Consequently, a context meeting was conducted with the representative users and with designers on the target websites, and also emails were sent to the website owners, all of these to obtain information that describes the prospective users of the websites they are working on and thus to determine the users' characteristics. This method was used by Thomas and Bevan (1996) and Rubin and Chisnell (2008). In conclusion, all users were willing to take the time to be involved after receiving a recruiting email (see Appendix G). Also, an email was sent to the owners of the targeted products, as in Appendix B9. After gaining their approval to take part in this meeting, the brief materials were sent to them to increase their knowledge and to reduce the time needed for the meeting, including the agenda for the context meeting and the targeted websites, as shown in Appendix B10.

Table 4.4: Distribution of participant profiles for mini- user testing

Question	Frequency			Percentage		
Years using a computer	Less than 1 year	1 to 4 years	More than 4 years	Less than 1 year	1 to 4 years	More than 4 years
	2	*4*	*14*	*10%*	*20%*	*70%*
Daily hours on computer	Less than 1 hour	1 to 4 hours	More than 4 hours	Less than 1 hour	1 to 4 hours	More than 4 hours
	1	*3*	*16*	*5%*	*15%*	*80%*
Browser	Internet Explorer	Google Chrome	Firefox Mozilla	Internet Explorer	Google Chrome	Firefox Mozilla
	7	*11*	*2*	*35%*	*55%*	*10%*
Years using the Internet	Less than 1 year	1 to 4 years	More than 4 years	Less than 1 year	1 to 4 years	More than 4 years
	3	*8*	*9*	*15%*	*40%*	*45%*
Daily hours on the Internet	Less than 1 hour	1 to 4 hours	More than 4 hours	Less than 1 hour	1 to 4 hours	More than 4 hours
	2	*7*	*11*	*10%*	*35%*	*55%*
Daily hours visiting educational websites	Daily	Less often		Daily	Less often	
	9	*11*		*45%*	*55%*	
Total of participant	20					

- Source of recruitment of users

After gaining an understanding of the characteristics of the end-users for the target websites, recruiting the actual users who fit these characteristics is the next step. Rubin and Chisnell (2008) listed many sources for recruiting users, such as university campuses, Internet users and societies. In this regard, two sources were adopted and used, which are email broadcasting (as shown in Appendix G) and advertising on bulletin boards (as shown in Appendix H). For the first experiment, emails were sent to students in the School of Computing Sciences at UEA, and to King Abdulaziz University (to the Deanship of e-Learning and Distance Education). Secondly, advertisements were posted on the bulletin boards that are scattered around the School of Thager Intermediate Stage, at the Saudi School in Norwich, the Union of UEA Students and in the British International School of Jeddah. For the second experiment, emails were sent to students in the School of Computing Sciences and the Union of UEA Students. Advertisements were also posted on bulletin boards in various different places in Norwich (Appendix H).

After receiving an adequate number of responses, the pre-test questionnaire was designed (see Appendix B3 and B4) , based on Rubin and Chisnell (2008), for sending to the volunteers who wanted to participate. It aims to gather background information such as education level and experience in using the Internet and the targeted domain. Also, it aims to match their information to the required characteristics in this research, ultimately for selecting the most appropriate users for each website. Furthermore, a confirmation email was sent to the participants, as shown in Appendix I, which includes scheduling their testing sessions, based on their convenience in terms of date, time and place.

- Test environment

Usability testing requires a realistic and controlled environment that allows the participants to perform their tasks under the same conditions. This environment may be determined based on availability of a location as well as on the volunteers. There are two environmental factors. The first one is physical, and includes noise, temperature, lighting, vibration and humidity. The second environmental factor is social, and includes the number of persons in the surrounding test environment, and the relationships between the participants and those persons. Thus, these environmental factors should be controlled to avoid systematic errors in the observed data (Maguire, 2001b); (Lazar et al., 2010). In this research, UT is conducted

in a laboratory that is provided with all the necessary equipment; it is a clean, quiet room, and has comfortable chairs and desks, appropriate lighting, Internet access, microphones, desktop/laptop computers, cameras, and a place for a moderator/observer. The simple single-room set-up was adopted because it gives the observer an excellent sense of what is going on whilst taking notes. Also, it helps the moderator to encourage the users quickly if they are struggling in the testing sessions or if they just need assistance. In this regard, all the users conduct their testing sessions under the same aforementioned conditions to avoid any variations that might lead to failure in the ensuing comparisons. The above instructions are recommended by Rubin and Chisnell (2008), Dumas and Loring (2008) and Sauro (2010).

- Testing Procedure

Rubin and Chisnell (2008) listed eight processes for conducting UT, which are: 1) Develop the test plan; 2) Set up a testing environment; 3) Find and select participants; 4) Prepare the test materials; 5) Conduct the test sessions; 6) Debrief the participants and observers; 7) Analyse the data and observations; and 8) Report the findings and recommendations. Furthermore, Lazar et al. (2010) highlights specific procedures for experiment sessions, which are; 1) Ensure the system being evaluated and the related instruments are ready for the experiment; 2) Greet the participants; 3) Introduce the purpose of the study and procedures; 4) Get the consent of the participants; 5) Assign the participants to a specific experiment condition according to the pre-defined randomization method; 6) Participants complete training task; 7) Participants complete actual tasks; 8) Participants answer questionnaires (if any); 9) Debriefing session; and 10) Payment (if any). Also, Hertzum et al. (2014) proposed a simplified model for usability tests, which consists of four points: 1) Users interact with the system in order to solve a set of tasks prepared ahead of the test; 2) Users verbalise their thoughts while solving the tasks (and to prompt verbalisation, users are reminded to keep talking or are asked questions about their behaviour); 3) An evaluator observes the users' behaviour and listens in on their thoughts (and on this basis, the evaluator identifies and reports usability problems); and 4) The evaluation takes place in the context of an overall relationship between users and evaluator. To obtain reliable evaluation results, the users must feel at ease. In this regard, all four of the aforementioned steps are adopted and the evaluation procedure for all the UT sessions is to be conducted through following the same protocols.

1. Orientation Session: This session entails firstly the researcher/facilitator welcoming the user and the user then reading the introduction script, which includes a detailed explanation of the aims of this study. The following step entails describing the testing procedure in detail, which includes identifying the domains to be tested as well as the chosen websites for each domain, explaining the special techniques that are to be used, such as how to think aloud whilst performing the tasks (through a training session), elucidating the equipment to be used as well as the setup of the environment in which the testing is to take place (a quiet room), informing the users that they have the right to leave or take a little break from the testing, and clarifying the questionnaires and forms to be used, such as the pre-test questionnaire, the post-test questionnaire, the consent form and the withdrawal form. Finally, it is explained to the users that their behaviour will be observed by the researcher in order to better understand their results.

2. Testing Session: Before starting the actual testing, training sessions are conducted for each user, including an exploration of each targeted website for a maximum of 10 minutes. After that, a small number of questions (prepared beforehand) are put to them to ensure that they have benefited from this training session, as shown in Appendix J. Then, the pre-test questionnaire is completed by each user to gather information about his/her background and experience, and to clarify some product-related information. This is followed by giving each user group the written task scenarios (designed for each particular website), and telling them that they will not be offered any suggestions or hints, but from time to time, they may ask for clarification of what has have said or for information on what the researcher is looking for. The researcher plays the role of observer and moderator during all the test sessions. Then, the post-questionnaire is completed by each user to obtain his/her feedback after completing the tasks. This should be prior to any discussion to reduce any effects of bias.

3. Debriefing Session: The researcher, who plays the role of observer, conducts a debriefing session with each user by reading the points that were observed and written down during the test session. These points are discussed in detail through asking questions about the user's behaviour and anything that happened during the test

session, about why the problems occurred and how to fix them, and about the results of the post-test questionnaire. This session finishes with preparing the master usability problems list. Then, the independent evaluators are involved to rank the severity of the discovered problems and classifying them to the appropriate usability problem areas and under the appropriate heuristic name.

4.4.9.2 Workflow of experiment findings

Usability inspection methods such as HE and the newly developed method (DSI) can reveal potential problems that may also be predicted by evaluators. This issue was considered by Gray and Salzman (1998) with regard to improving the construct validity of these predicted problems. In this regard, Woolrych et al. (2004) carried out work for four years to investigate the reliability of inspection methods, such as HE, in terms of generating false problems. They began their investigation by verifying the extent to which researchers can be sure that an evaluator's prediction is really a false positive. In that case, they addressed the issue of false positives by employing 'falsification testing (It means that the evaluator makes a prediction of a problem and then the UT reveals whether the evaluator was right?)'. The process of falsification testing involves the accurate testing of evaluators' predictions using UT, as in Figure 4.4. It includes three steps which are collecting evaluators' predictions, translating them into tasks, and verifying tasks against predictions. Briefly, fixed UT tasks are designed; these are derived from evaluators' predictions, and they are applied to assess individual problems in order to identify any likely user difficulties that may arise in the testing session. These tasks should expose these likely difficulties. There is no need to design an individual task for each prediction; tasks can be designed to address a set of evaluators' predictions. Consequently, if a prediction is confirmed by UT, it is a real problem. If a prediction is not confirmed by UT, it can be confidently coded as a false positive. Woolrych et al. (2004, p.3) asserted, "falsification testing ensures that false positive coding of predictions is not a consequence of incomplete coverage in user testing" (Woolrych et al., 2004). After analysing the evaluators' predictions, the problems should be merged into a master problem set.

In this research, the falsification testing and a two-way mapping procedure (forward- and backward-matching) were adopted (Hvannberg et al., 2007). The experimental workflow was designed to verify the results of the newly developed method (DSI) and HE, and thus to

identify the real, false positive, false negative, missing, and hit problems as shown in the Figure 4.5.

Figure 4.4: Process of Falsification Testing (Woolrych et al., 2004)

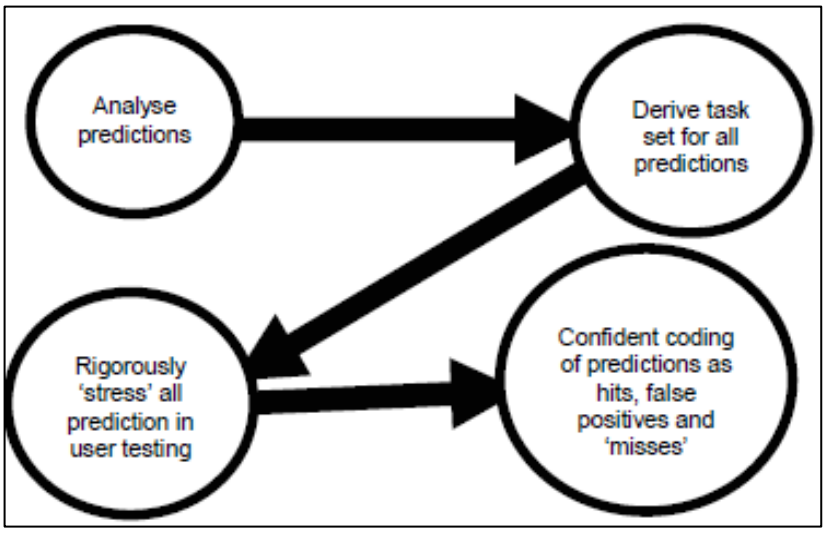
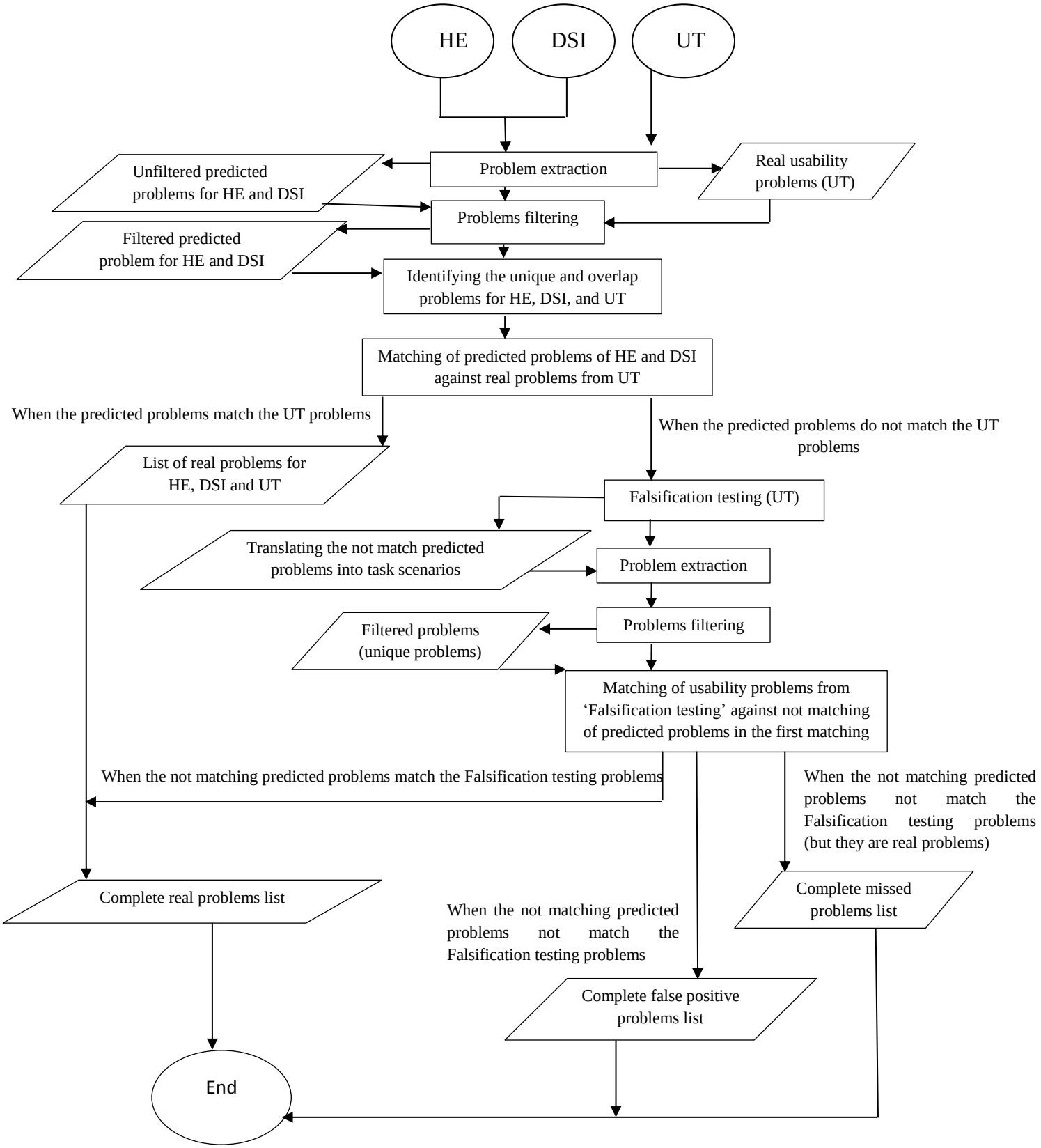


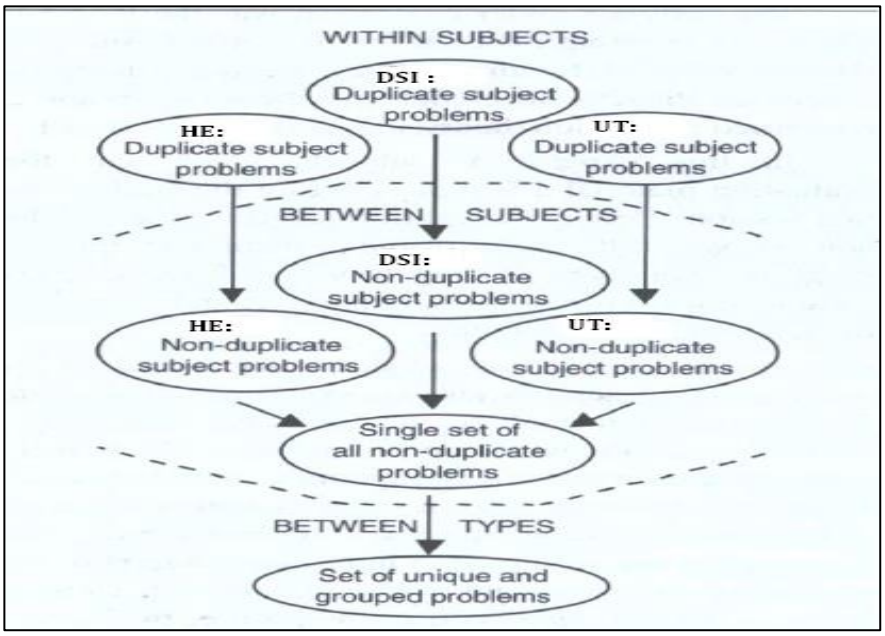
Figure 4.5: Experiment workflow



4.4.9.3 The problem reduction process

The problem reduction process technique was adopted in this research and it was proposed by Connell and Hammond (1999). It consists of two stages, which are 'within subject' and 'between subject' as shown in Figure 4.6. The ‘within subject’ stage involves the elimination of any problem duplication in each method. The ‘between subjects’ stage investigates the way in which a single problem sets an overlap with those problems of other methods. This involves identification of any problems which had been discovered by more than one method in each experiment. The result of this stage is a single set of non-duplicate problems. The single set is later used to measure the efficiency of each method to identify the number of unique problems for each method, the number of overlapping problems between methods, and to match the predicted problems of HE and DSI against real problems from.

Figure 4.6: The problem reduction process adapted from (Connell and Hammond, 1999)



4.4.9.4 Usability measures

Different types of data can be collected by HE or DSI and UT. However, there are only two types of measure in the three methods; the first type is the participants’ performance and the second is the participants’ subjective feelings. The former includes time spent and number of problems found, whereas the latter includes the participants’ satisfaction level, comments, attitude and severity rating (Nielsen, 1992a); (Dumas and Redish, 1999); (Chen and

Macredie, 2005). Moreover, efficiency, effectiveness and satisfaction are widely used metrics (Sauro and Kindlund, 2005). Also, validity, reliability and thoroughness are other metrics that should be considered (Hartson et al., 2003). Accordingly all are considered in this study, and the following is an explanation of the above measures in detail.

- Time spent

Time spent is the metric most often used to measure the efficiency attribute. In HE and DSI, it indicates the time spent by an evaluator to complete an evaluation (in minutes). However, in UT it indicates the average task completion time for the users to complete their task successfully (in second or minutes), the average time spent on failed tasks or on tasks completed incorrectly, and the average time taken per task across the users (Sauro, 2011c). There are two ways to calculate the time: use a digital stopwatch or use software or automatic tools (Law and Hvannberg, 2002); (Albert and Tullis, 2013). In this research and in terms of the available resources, digital stopwatches are used with serious consideration being paid to accuracy, especially to the start and finish of an evaluation in order to measure the efficiency attribute. The task time starts when the user has finished the reading the task scenario and it ends once they have finished all the required actions. However, an evaluation time starts when the evaluator has finished reading the developed checklist and it ends once they have finished all the necessary actions (Sauro, 2011a).

- Number of problems

Usability problem has different definitions in the literature, for example, Skov and Stage (2005, p.1) defined it thus, "it is a key element in a usability evaluation of an interactive system", whereas Nielsen and Landauer (1993, p.388) stated that "a usability problem is any aspect of a user interface that is expected (or observed) to cause users problems with respect to some salient usability measure (e.g. learnability, performance, error rate, subjective satisfaction) and that can be attributed to a single design aspect". It can also be defined as a difficulty that a user faces during an evaluation that prevents him/her from completing their task or that necessitates spending more time on the wrong page of the website (Albert and Tullis, 2013). In heuristic evaluation, an expert evaluator inspects pages in a chosen website, and finds any errors or confusions that may affect users in completing their task successfully or that affect the efficiency and effectiveness of the website (Molich and Nielsen, 1990). On

the other hand, each user in UT performs a set of tasks in the presence of an observer who notes any difficulties faced by the user and considers them as problems. Also, the observer may ask questions to ensure that no problems are missed (e.g. Why are you taking so long to perform this task? What problems are you facing? What do you think of the website after having used it?).

- Severity rating

After reporting the discovered usability problems, they need to be rated in terms of their priority for fixing. To achieve this task, the independent evaluators read the structured report to fully understand each problem, and thereby determine the severity level for each problem correctly. In HE and DSI, the severities of the discovered usability problems in the master list are classified by independent evaluators after the actual evaluation. Also, other independent evaluators are involved in UT to rank the users' problems, which are reported by the observer. In this research, these problems are classified into different groups to which a numeric scale is applied in order to measure the severity of each problem, as proposed and used by (Nielsen, 1994b) and mentioned in the literature review.

- Satisfaction ratings

The satisfaction attribute is one measure of usability, and it can be used to measure the feeling of an expert evaluator and users about a product. There are several ways to measure satisfaction attirbute, but the most commonly used is the System Usability Scale (SUS) questionnaire. It is reliable and free tool, and comprises ten items in the form of scale questions ranging from 0 to 100 (Brooke, 1996). The original SUS items refer to 'system', but some researchers have proposed minor changes to the wording of these items. For example, Finstad (2006) and Bangor et al. (2008) changed the word of 'cumbersome' to 'awkward' in Item 8. Sauro and Lewis (2012, p.198) state that " the original SUS items refer to 'system', but substituting the word 'website' or 'product' or using the actual website or product name seems to have no effect on the resulting scores. Of course, any of these types of minor substituting should be consistent across the items''.

As this research designed a DSI as tool to be used for evaluation of the websites, it is necessary to measure the satisfaction attribute for this tool against a HE tool. To the best of my knowledge, there is no research using an SUS questionnaire as a method to measure

satisfaction with evaluation methods themselves. Consequently, the rating scale questionnaire was developed as shown in Appendix B5. It consists of nine items with five response options for respondents, from 'strongly agree' to 'strongly disagree'. These items were adopted from the original SUS questionnaire, but they were modified to measure the feeling of satisfaction of evaluators with DSI and HE methods. This feeling of satisfaction was taken after the evaluator had finished his/her evaluation session but before the debriefing session. After that, the total score for each item was computed so that it would give some idea about the differences, and this score is comparable with another method. The rating scale questionnaire is used (which was adopted from the original SUS questionnaire) because it is quite simply there is no better way to expressing that in this moment but in the future might be uses another technique.

For UT, when the users have finished their tasks, a single question to measure their satisfaction should be designed, as recommended by (Hornbæk and Law, 2007). Thus, users are 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 after the testing session (Nielsen and Loranger, 2006).

- Comments, feedbacks and attitude

One of the most important methods in usability testing for gathering users' comments is the Think Aloud protocol (Albert and Tullis, 2013). This method is used during experiments, after conducting training sessions with the participants. During the sessions, the participants are observed and monitored, and reminded and asked by the observer (researcher) to express their thoughts and feelings, but this is done without bias or impacting on the participants. Also, other methods can be used after finishing the experiments to capture preference data, such as questionnaires and the debriefing session. Sauro (2010) stated that "you need a way to collect the 'why' behind the numbers".

In this research, the evaluators and users are asked to complete post-test questionnaires, writing down their comments and feedback on the methods used, and explaining any reaction that was observed during the test.

4.5 Data analysis

The data are analysed to determine which method provides optimal results with regard to the identification of comprehensive usability problems and relevant UEM metrics, with minimum input in terms of the cost and time usually spent on employing UEMs. However, conducting three evaluation methods produces a large number of data, which can be divided into two categories: performance data and preference data. The former describes what actually happened during the test sessions to measure such aspects as efficiency and effectiveness, whereas the latter describes what the participants thought during their sessions to measure aspects such as satisfaction. These data include a set of usability problems, usability evaluation metrics, participants' comments and attitudes, observer's notes, time spent, costs incurred, severity rating and satisfaction rate (Jiang, 2009). Indeed, there are two kinds of collected data, which are qualitative and quantitative data. The following is an explanation in terms of how these data were obtained from the three methods that were analysed in each phase.

4.5.1 Qualitative analysis

This study gathers the participants' comments (evaluators and users) and observer's notes. These data help to understand the users' thoughts and experiences while they perform their different tasks on the chosen websites in the different domains. The evaluators' attitudes are also analysed as qualitative data, which can be used to understand how a group of evaluators performs, how each method is performed, and the reasons behind the evaluators' comments. Moreover, comparing the results obtained from UT against HE or the new method (DSI), and vice versa, may lead to different findings, and it may be that one method outperforms the others in discovering certain types of problem.

- Usability problem report

The report will be produced at the end of this study. It consists of all the collected data, all the important findings, and the recommendations to the website owners. This report will be generated from a practical point of view, and will be organised to include the findings of the three evaluation methods (HE, DSI and UT), and it will be sent to the website owner.

4.5.2 Quantitative analysis

The quantitative data (performance data) are analysed in three ways: descriptive statistics, inferential statistics, and other usability measures.

- Descriptive statistics

This technique is used after cleansing the collected data (e.g. questionnaires). It helps to understand the nature of a dataset. The most frequently used descriptive measures are mean, median, mode, variance, standard deviation, score of frequency, and range (Lazar et al., 2010). Dumas and Redish (1999) pointed out that most usability studies require only descriptive statistics and qualitative data (e.g. participants' comments and observer's notes).

- Inferential statistics

This analysis technique helps to consider more carefully what features of the data are real and what are merely chance variations. It can confirm whether or not there are statistical differences and correlations between the groups or whether there is any influence on the part of the usability factors on the results (Cairns and Cox, 2008). Cairns (2007, p.1) stated, "inferential statistics being the usual understanding of statistics as tests producing p- values and significance results, and in doing so provide a clear picture of the quantity and quality of statistical methods as used in HCI research".

In this research, the selections of the appropriate statistical analysis tools were discussed with two experts at the University of East Anglia (UEA). The SPSS 22 software package was used as they recommended it and as I learned during the Personal and Professional Development (PPD) training courses at UEA. Consequently, descriptive statistics were used and inferential statistics were also employed for a number of specific reasons which are, first of all, they help to offer interesting findings from diagnosing usability issues. Secondly, they increase the validity and reliability of research results. Next, they are commonly used in HCI, so there is no valid reason to ignore them. Finally, this research has been designed to investigate any correlations or significant differences between the groups and methods used. In this regard, a number of statistical tests were used, including Mann Whitney U test, kruskal wallis, Cronbach's alpha, and Person correlation.

- Usability measures

There are additional measures that can be used on the gathered data (e.g. efficacy, effectiveness, validity, thoroughness, satisfaction, reliability), and that help to achieve the goals of this research; for example, participants' satisfaction, number of discovered usability problems (e.g. real, false positive, false negative, miss, hit), and time spent. All those data are analysed and compared in terms of the groups' performance based on the methods they used during the evaluation.

4.6 Reliability and validity

Conducting reliability and validity tests on what has been done in an experiment is a very important part of usability studies in order to achieve the research goals and to strengthen the findings. Chen (2006, p.24) stated, "if the research work is valid, it means that the judgment is made about the extent to which relevant evidence supports that inference as being true or correct". For instance, a valid experiment in usability studies should remove any potential biases, involve the right participants, design good tasks, and consider environmental influences (Lazar et al., 2010); (Nielsen, 1993). In this regard, Gray and Salzman (1998) mentioned three validity categories, that are the most related to HCI research, which are:

4.6.1 Internal validity

In general, there are no certain tests for measuring research validity. Consequently, a well-conducted usability study, i.e. one that controls all the potential factors that threaten the research's validity, is the main responsibility of any researcher. Seliger (1989, p.95) pointed out that "any research can be affected by different kinds of factors which, while extraneous to the concerns of the research, can invalidate the findings". Moreover, there are a number of factors which affect internal validity, such as size of sample population, subject variability, instrument, time given for performing experimental treatments, allowing the evaluators and users to rate the severity of their problems, and setting (Berg and Latin, 2008); (Gray and Salzman, 1998).

In this research, the same instruments for the three methods were used in all the experiments. Also, the measures for the usability problems were adopted from the valid and reliable severity rating that has been used in many usability studies (Nielsen, 1995a). Also, independent evaluators were involved for ranking the severity of the discovered problems

(from both evaluators and users). The researcher was involved in gathering the data and played the role of observer in the UT sessions. This means that the expert evaluators discovered the usability problems themselves during their evaluation sessions. Furthermore, the users who have been recruited, in terms of all their experiences, represent the real users of each website, and they are equal in terms of certain characteristics. Moreover, all the evaluators and users in each experiment conducted their tests under the same conditions in the same environment and followed the same procedure, and the results of each method in each experiment were analysed individually.

4.6.2 Construct validity

Construct validity ascertains whether or not the experiments have be measured and carried out as planned. Alleva and Branchi (2011, p.290) defined it as "the extent to which a procedure appears to measure a higher order, inferred theoretical construct, or trait in contrast to measuring a more limited dimension". In this research, all the experiments were conducted and measured as planned. All the methods used are clearly described in detail, and they were carefully planned before conducting the actual experiments (Gray and Salzman, 1998).

4.6.3 Statistical validity

If the experiment has internal and construct validity, then statistical validity should be examined here. As mentioned previously, statistical tests are very commonly used to assess validity in any research; for example, validity can be used to examine any different in performance between a users' group and two or more methods. Also, it can be used to examine the demographic data to determine whether or not they impact on the results achieved. In this research, statistical tests were employed at different stages, and the number of users is large enough to allow statistical validation tests to be executed and achieved. The following is explanation of how to select the correct statistical test.

- Selecting a test

There are a set of statistical tests that can be performed to complement an experiment or to find the evidence to prove a point. For this reason, it is important to choose the appropriate statistical test. There are two general categorisations of statistical procedures, which are parametric tests and nonparametric tests. The former assumes that data are normally

distributed in the population and variances are approximately equal (Albert and Tullis, 2013); (Sauro and Lewis, 2012). However, the latter is defined by Albert and Tullis (2013, p.33) where they state ‘’nonparametric tests are used for analysing nominal and ordinal data. They assume that the distribution of the data does not follow normal distributed‘’. In this regards, there are many tests for the assessment of normality such as Kolmogorov-Smirnov (K-S) test and Shapiro-Wilk test (Ghasemi and Zahediasl, 2012). Many researchers recommend to use the ShapiroWilk test for testing the normality of data. This is pointed out by Ghasemi and Zahediasl (2012) when they state that ‘’Shapiro-Wilk test provides better power than the Kolmogorov-Smirnov test. Power is the most frequent measure of the value of a test for normality—the ability to detect whether a sample comes from a non-normal distribution‘’. However, there is another point that can affect the result of the normality test. This point is the sample size. Ghasemi and Zahediasl (2012) emphasise that ‘’for small sample sizes, normality tests have little power to reject the null hypothesis and therefore small samples most often pass normality tests’’. Thus, if the sample size is small, the data should be assumed to be not normally distributed. Furthermore, each category has a set of tests based on type of the data, the number of groups, the number of independent variable (IV) and the experimental design (between and/or within subject designs). For example, a Chi-square test is applied to sets of categorical data. A t- test is used for comparing between two groups when the data are parametric and between-subject designs. If the comparason is between two groups within subject designs, then the paired- samples t-test is used. To analyse variance, one-way ANOVA and Repeated measures ANOVA are used for comparing between more than two groups when the data are parametric, but the former is between-subject designs and the latter is within-subject designs. On the other hand, if the data are not parametric, then, Wilcoxon and Mann–Whitney are used. They are like a t-test but the former is for within-subject designs and the latter is for between-subject designs. To analyse variance, the Kruskal-Wallis test is used for comparing between more than two groups when the data are not parametric and between-subject designs. Also, Friedman’s test is used for comparing between more than two groups when the data are not parametric and within-subject designs. Furthermore, all the above tests are used when there is one independent variable (IV). If there are two or more independent variables, the Factorial ANOVA is used between-subject designs, and Repeated measures ANOVA is used within-subject designs. If the study is adopted a split-plot design that involves both between and within subject designs, then Split-plot ANOVA test is used.

In addition, Bonferroni post-hoc test is used when many dependent or independent statistical tests are being performed (Napierala, 2012). Napierala (2012, p.1) defines this teas as ‘’ it is an adjustment made to P values when several dependent or independent statistical tests are being performed simultaneously on a single data set. To perform a Bonferroni correction, divide the critical P value (α) by the number of comparisons being made. Also, It is used to reduce the chances of obtaining false-positive results (type 1 errors) when multiple pair wise tests are performed on a single set of data‘’. Also, Pearson correlation is used to know whether there is a relationship between two different things (Cairns and Cox, 2008); (Lazar et al., 2010); (Sauro and Lewis, 2012). The aforementioned tests are considered before analysing the experiments' results when choosing the appropriate statistical tests in this study.

In regards to the previous literature reviews, the gathered data in this research has properties that it will be discussed in more detail in chapter of results. These properties briefly are;

1. The data of DSI and HE methods are too small in both experiments. Therefore, there is no needed to perform the normality tests due to that these tests have little power to reject the null hypothesis as it pointed out by Ghasemi and Zahediasl (2012). Consequently, the data of the both methods assumed as not normally distributed and the nonparametric tests were performed. Mann-Whitney U test was chosen at a significant level of 5% as it is the most suitable as pointed out by Ghasemi and Zahediasl (2012).

2. The data of UT method in both experiments was tested using Shapiro-Wilk test to know whether these data are normally distributed or not as the best choice for testing the normality of data. The results of Shapiro-Wilk test found that the data is not normall distributed, so the nonparametric tests were performed. Kruskal-Wallis test was chosen at a significant level of 5% as it is the most suitable as pointed out by (Lazar et al., 2010); (Sauro and Lewis, 2012).

3. For the multiple comparisons between three methods (DSI, HE, and UT), the Bonferroni post-hoc test was used to reduce the chances of obtaining false-positive results (type 1 errors) on the comparison data as recommended by Napierala (2012).

- Reliability measurement

In terms of the reliability in evaluation studies, Rogers et al. (2011, p.99) defined reliability as "how well a technique produces the same results on separate occasions under the same circumstances". Furthermore, controlled experiments lead to high levels of reliability, and consequently anyone repeating the same procedure for this experiment should in theory achieve the same results (Nielsen, 1993); (Rogers et al., 2011). In this regard, it was difficult to repeat the same experiment twice in order to examine whether it achieved the same results because of the time limitation. However, reliability measures were adopted in this study, as mentioned in the literature review; they include Cronbach's alpha and Any Two Agreement. Reliability is considered side by side with internal validity, construct validity and statistical validity, which will make this study valid and reliable.

4.7 Conclusion

This chapter has outlined the chosen research methodology, which is an integral part of study planning. The methods adopted are very commonly used in the HCI and IS fields. Also, it presented the research design based on the study objectives. It discussed both quantitative and qualitative research techniques, employing mixed methods (called triangulation), the approaches taken for all the experiments, and the techniques used for gathering and analysing the data. Furthermore, it explained the usability evaluation methods that are used in this study, and highlighted the usability evaluation metrics used, together with consideration of the reliability and validity methods (to eliminate any potential threats in this research). The next chapter will discuss the preparation and the results of the both experiment.

Chapter 5: Results

5.1 Introduction

As earlier mentioned in the introductory chapter, two experiments will be conducted on a different domain for more validation the adaptive framework and the method generated from this framework. Two domains were chosen which were educational and social networks domains. The reason behind choosing these domains is to measure the efficiency of the adaptive framework and its generated method (DSI) through applying it in two different domains in terms of aim and features. Particularly, The second experiment aims to answer this question which is "if the adaptive framework is used to generate a DSI method for another domain, will it succeed in generating a perfect method or a fail?''. This chapter details the results of the these experiments. It presents the comparison results derived from employing the three different evaluation methods used in this research. These methods are Heuristic Evaluation (HE), User Testing (UT), and Domain Specific Inspection (DSI). UT and HE were used in both experiments as validation methods for DSI in order not to affect the research validity. UT was used to find the real problems in the tested domains and to give more validity to the experiment, as its results can be used for benchmarking (Chen and Macredie, 2005). HE was used because it is a better method for making head-to-head comparisons with DSI's results; also, HE is similar to DSI, being in the same group (inspection methods). Thus, it seems that using both UT and HE may offer and confirm different results to those achieved by using the DSI method. Consequently, these methods will form a major part of this chapter. This chapter starts with a thorough exploration of the experiment's objectives. It also explains how the steps of the adaptive framework are followed to generate DSI methods for the educational and social networks domains. Then, the results achieved from HE, UT and DSI will be explored in detail. They include a number of quantitative and qualitative data analyses. For example, they include each method's performance in terms of discovering usability problems, discovering unique and overlapping

problems, discovering real and false problems, time spent, and set of performance metrics and their relationships to the results achieved. Furthermore, the sample sizes for the three methods, as well as the different rules for sample size, will be examined. Moreover, a statistical analysis will be included in order to find any significance between results obtained. Finally, the experimental results will be discussed and summarized.

5.2 The experiment's objectives

As mentioned in Chapter 4, two controlled experiments are conducted in order to achieve the research objectives. The main objectives of two experiments are;

1. To validate practically the adaptive framework by generating the DSI method for educational and social networks websites.
2. To examine the performance of the three methods (DSI, HE, and UT) in terms of discovering the number of real usability problems, time spent, and identifying a set of usability evaluation method measures (in both domains).
3. To find any relationships between the usability measures used.
4. To explore the effect of sample size and to investigate the role of the number of evaluators and users needed in a usability study.

5.3 The targeted websites

As mentioned in Section 4.4.2, the targeted websites (three websites for free educational domain and three websites for social networks domain) were initially chosen by the researcher for both experiments based on six aspects and these were: 1) representative of the chosen domain based on clear definition and classification in the literature; 2) popularity based on the number of users based on the statistical studies in the literature (for easy recruit the sampling) ; 3) free to join website; 4) the website is relevant to the subject of study and direct to the scope of the chosen domain; 5) rich functionality and different features, for example, at least four modules and four features for free educational websites, and four features for social networks websites; and 6) not familiar to the users in the testing session. Then, two experts in each domain were consulted to make sure that the chosen websites were appropriate and representative as mentioned in Section 4.4.2.

5.4 Recruiting users and experts

For evaluating the practicality of the adaptive framework for generating a DSI method, a set of users and experts were recruited for step two and three in the adaptive framework, and for the validation experiments. These users and experts are chosen carefully based on the set of aspects mentioned in Section 4.4.6.5 and Section 4.4.9.

5.5 Evaluation of the practicality of the adaptive framework

This process began from the desire of the researcher to validate practically the adaptive framework (see Figure 3.1 and 3.2) by generating the DSI methods for educational and social networks websites.

➢ Educational websites

- In the first step: the researcher conducted an extensive literature survey on the materials relating to usability of educational websites and UEMs as well as on the requirements of the free educational websites such as (Chattratichart and Lindgaard, 2008); (Chattratichart and Brodie, 2004);(Chattratichart and Brodie, 2002); (Alsumait and Al-Osaimi, 2009); (Tan et al., 2009); (ISO, 1998b); (Triacca et al., 2004); (Stracke and Hildebrandt, 2007); (Oztekin et al., 2010); (Lee, 2010); (Squires and Preece, 1996); (Magoulas et al., 2003a); (Ardito et al., 2006); (Bernérus and Zhang, 2010b); (Muir et al., 2003); (Abuzaid, 2010); (Kukulska-Hulme and Shield, 2004); (Reeves et al., 2002); (Miller, 2005); (Alkhattabi et al., 2010). From the literature, it was found that all the developed heuristics are designed for evaluating e-learning software and that they were extended from Nielsen's heuristics as shown in Appendix L1. There are therefore no focused heuristics that are designed exclusively for educational websites. These studies have used varied methods to extend Nielsen's heuristics such as extensive literature, survey, and user testing. In regard to usability areas, Squires and Preece (1996) identified three categories with their items that are concerned with educational issues, which are content, instructional quality and technical quality, as shown in Appendix L1 (Table 5). Also, Noiwan and Norcio (2000) evaluated four academic websites and they found that most usability problems fall into the categories of lack of navigational tools and a site map, old content, and inconsistency problems. Furthermore, Kostaras and Xenos (2007) evaluated the usability of the Hellenic Open University website and they found that most usability problems are related to

inconsistency problems, poor navigational support links, and the inappropriate design of the menu. Additionally, Astani and Elhindi (2008) evaluated the usability of 50 websites of different colleges and universities. They found that most of the usability problems are related to old content and inappropriate layout. Moreover, Du Toit and Bothma (2009) evaluated the usability of the website of an academic marketing department in the University of South Africa, and they found that the majority of usability problems are related to old content, lack of navigation tools, and incomplete information in some modules. However, these studies did not give details about specific types of usability problems that could be discovered on educational websites. Consequently, these results are a starting point for generating specific heuristics and checklist for educational websites, and they will be useful for combining with the user input in the next step.

- In step two, a context meeting was held with five users prior to a mini user testing. The agenda for the context meeting session was discussed with the five users, as shown in Appendix B10. As mentioned in Section 4.4.6.5, the aim of this meeting was to understand the aim of educational websites, to design the set of context tasks (see Appendix K for tasks that were designed based on the context meeting result) to be used in the mini user testing, and to identify the real users to recruit for the mini user testing. After that, the mini user testing was conducted on two websites (CosmoLearning and SchoolsWorld) with 10 users who were regular educational website users and were recruited based on the context meeting result, and the post-testing questionnaire was used to gather users' feedback. Appendix L2 (Table 1) shows the results of the context meeting. It also shows the usability problems that were discovered from the mini user testing (Table 2), and it lists the results of the post-testing questionnaire regarding the features that should be included on educational websites (Table 3). These results were analysed to develop a set of heuristics based on the results of the mini-user testing, as shown in Appendix L2 (Table 4). These heuristics were checked by one independent evaluator. These results were the second starting point to generate specific heuristics for educational websites. The results from this step will be discussed by experts in the next step in order to establish a set of DSI heuristics based on the results of the mini user testing, and are combined with the expert discussion results as shown in Appendix L4.

- In step three, a focus group discussion was conducted with three experts in usability and in the educational domain (single and double experts). The focus group also included two users who were involved in the previous step to discuss the results that were obtained from the mini user testing. Before starting the focus group, the results of the literature review and the mini user testing were analysed for content, and 35 areas were formed, as shown in Appendix L2 (Table 5). Then, this result and the results of the literature review and the mini user testing were sent to the participants before the focus group discussion with a questionnaire that used a five point Likert scale. Bertram (2007, p.1) defines this scale as "ranging from 'Strongly Disagree' on one end to 'Strongly Agree' on the other with 'Neither Agree nor Disagree' in the middle". The reasons for sending this questionnaire to the experts was to rate the discovered areas for further validation, and to elicit input from the experts by asking them to add any new areas that they thought should be included. They suggested 12 new areas, as shown in Appendix L2 (Table 5). This also enabled the researcher to identify the final usability problem areas. During the discussion, the results of the questionnaire, the results of the previous steps, and the experts' points of view were discussed. Cohen's kappa coefficient was used to enable a calculation of the reliability quotient on the result of the Likert questionnaire. The intra-observer test-retest using Cohen's kappa yielded a reliability value of 0.8, representing satisfactory agreement between the evaluators in the focus group session. Then, the usability problem areas were merged and grouped into five usability problem areas, which are user usability, motivational factors, content information and process orientation, learning process, and design and media usability. Also, a set of DSI heuristics with their explanations were identified based on the user and expert inputs. Furthermore, the results of the focus group discussion were reported (see Appendix L3) for further analysis in the next step.

- In step four, the identified heuristics were classified according to the agreed usability areas. Thus, the DSI method was created (see Appendix L4), closely focused on the free educational websites. Furthermore, the researcher analysed the results of the three steps and incorporated the findings. Therefore, the DSI checklist was created based on the results of the three steps as detailed in Appendix M2. It includes the most features of free educational websites to provide a wide range of evaluation of these websites. These elements were classified under the appropriate heuristics. This checklist aims to facilitate the process of evaluation and analysis (e.g. matching process), and to help designers and

programmers to identify the areas in their website that needed improvement, as recommended by Chen and Macredie (2005). Also, it allows anyone to adopt any usability area with its heuristics and checklists to evaluate a specific part of free educational website. Appendix M1 describes one example of how this checklist was created for one heuristic. This checklist was tested during the pilot study as mentioned previously.

- Social networks websites

- In the first step, the researcher conducted an extensive literature review on the materials relating to usability of social network websites and UEMs, as well as on the requirements of social network websites, such as (Ellison, 2007), (Estes et al., 2009), (Fox and Naidu, 2009), (Al-Badi et al., 2013), (Fu et al., 2008), (Hart et al., 2008), (Pessagno, 2010), (Preece and Maloney-Krichmar, 2003), (Bahiss et al., 2010), (Preece, 2001), (Stevenson and Liu, 2012) and (Wentz and Lazar, 2011). From the literature, it can be seen that a few studies have developed heuristics for evaluating particular areas in social network sites (SNSs). For example, Jamal and Cole (2009) used privacy heuristics to evaluate the interface of Facebook's advertising tool Beacon. These heuristics were developed from a Structured Analysis of Privacy (STRAP) framework (Jensen and Potts, 2007). Appendix Q1 (Table 1) shows the privacy heuristics that were used by Jamal and Cole (2009). Malinen and Ojala (2011, p.2) introduced a set of heuristics for evaluating sociability (see Appendix Q1 ,Table 2). They justified the introduction of the new heuristics thus: "traditional usability evaluation methods are not capturing all the important aspects of social web use, such as self-expression or social pleasure". They used an extensive literature review and user studies to establish their heuristics. Gallant et al. (2007) introduced five heuristics for increasing social interaction in web-based communities by studying users of Facebook and MySpace. They developed these heuristics after conducting content analysis with three user focus groups of Facebook and MySpace. Appendix Q1 (Table 3) shows these heuristics. In this regard, to the best of the researcher's knowledge, no specific heuristics have been developed to design and evaluate social network websites (SNSs). This was confirmed by Chinthakayala et al. (2013, p.25) when they stated "there is a lack of guidelines on developing social networking applications". Also, they stated that (p.2) ''there is no consensus on the guidelines for developing social networking sites''.

With regard to usability problem areas, Silius et al. (2011) developed a Web Service Quality (WeSQu) evaluation tool to evaluate the quality of social media in an educational context. They listed five categories which are privacy and security, information reliability, supporting navigation, accessibility and motivating the user. Also, Dwyer et al. (2007) made a comparison between Facebook and MySpace using a privacy trust model. They found that user profiles and the privacy of the shared information were the most worrying aspects for the users of these websites. Furthermore, Al-Badi et al. (2013) evaluated the LinkedIn website with user testing and a traditional heuristic evaluation. They found that the majority of usability problems were regarding sending feedback and requesting help, small difficulties with navigation, and changing the privacy settings. Additionally, Chinthakayala et al. (2013) used a user study to evaluate three social networking sites based on four criteria which were navigation, interactivity, source credibility and intelligence. These were developed from a comparative model, called NICI, which was proposed based on two major factors for evaluating the success of online communities, i.e. usability and sociability. In addition, Owens et al. (2009) evaluated the usability of Twitter with first-time users. They found that most usability issues related to sending and replying to messages, terminology and codes, and deciphering captchas. Therefore, they provided a list of recommendations on how to improve the usability of Twitter for first-time users as shown in Appendix Q1(Table 4). Similarly, Fox and Naidu (2009) evaluated the usability of three social networking sites (MySpace, Facebook, and Orkut) with first-time users. They found several usability problems, falling into several categories: terminology, colour and font use, feedback and error messages, and login and sign up. Thus, they provided a list of recommendations about how to improve their usability, shown in Appendix Q1 (Table 5). Moreover, Alam and Ali (2010) used different usability testing methods to examine their efficiency from a social network's point of view, particularly Facebook. They found usability problems in Facebook in the profile, searching, video tagging, wall, and chatting. However, these studies did not give details about specific types of usability problems that could be discovered on social network websites. Consequently, these results are a starting point for generating specific heuristics for social network websites, and they will be useful for combining with the user input in the next step.

- In step two, a context meeting was held with five users prior to a mini user testing. The agenda for the context meeting session was discussed with the five users, as shown in Appendix B10. As mentioned in Section 4.4.6.5, the aim of this meeting was to understand the aim of social network websites, to design the set of context tasks (see Appendix P for tasks that were designed based on the context meeting result) to be used in the the mini user testing, and to identify the real users to recruit them in the mini user testing. After that, the mini user testing was conducted on two websites (MySpace and Flickr) with 10 users who were regular social network website users and were recruited based on the context meeting result, and the post-testing questionnaire was used to gather users' feedback. Appendix Q2 (Table 1) shows the results of the context meeting. It also shows the usability problems that were discovered from the mini user testing (Table 2), and it lists the results of the post-testing questionnaire regarding the features that should be included on social network websites (Table 3). These results were analysed to develop a set of heuristics based on the results of the mini-user testing, as shown in Appendix Q2 (Table 4). These heuristics were checked by one independent evaluator. These results were the second starting point to generate specific heuristics for social network websites. The results from this step will be discussed by experts in the next step in order to establish a set of DSI heuristics based on the results of the mini user testing, and are combined with the expert discussion results as shown in Appendix Q4.

- In step three, a focus group discussion was conducted with three experts in usability and in the social network domain (single and double experts). The focus group also included two users who were involved in the previous step to discuss the results that were obtained from the mini user testing. Before starting the focus group, the results of the literature review and the mini user testing were analysed for content, and 33 areas were formed, as shown in Appendix Q2 (Table 5). Then, this result and the results of the literature review and the mini user testing were sent to the participants before the focus group discussion with a questionnaire that used a five point Likert scale. The reasons for sending this questionnaire to the experts was to rate the discovered areas for further validation, and to elicit input from the experts by asking them to add any new areas that they thought should be included. They suggested 28 new areas, as shown in Appendix Q2 (Table 5). This also enabled the researcher to identify the final usability problem areas in the social network websites. During the discussion, the results of the questionnaire, the results of the previous

steps, and the experts' points of view were discussed. Cohen's kappa coefficient was used to enable a calculation of the reliability quotient on the result of the Likert questionnaire. The intra-observer test-retest using Cohen's kappa yielded a reliability value of 0.9, representing satisfactory agreement between the evaluators in the focus group session. Then, the usability problem areas were merged and grouped into seven usability problem areas, which are layout and formatting, content quality, security and privacy, business support, user usability, sociability and management activities, accessibility and compatibility, and navigation site and search quality. Also, a set of DSI heuristics with their explanations were identified based on the user and expert inputs. Furthermore, the results of the focus group discussion were reported (see Appendix Q3) for further analysis in the next step.

- In step four, the identified heuristics were classified according to the agreed usability areas. Thus, the DSI method was created (see Appendix Q4), closely focused on the social network websites. Furthermore, the researcher analysed the results of the three steps and incorporated the findings. Therefore, the DSI checklist was created based on the results of the three steps as detailed in Appendix R2. It includes the most features of social network websites to provide a wide range of evaluation of these websites. These elements were classified under the appropriate heuristics. Appendix R1 describes one example of how this checklist was created for one heuristic. This checklist aims to provide guidelines to facilitate the process of evaluation and analysis (e.g. matching process), and to help designers and programmers to identify the areas in their website that needed improvement, as recommended by (Chen and Macredie, 2005). Also, it allows anyone to adopt any usability area with its heuristics and checklists to evaluate a specific part of social network website. This checklist was tested during the pilot study as mentioned previously.

5.6 The experiment data analysis

This section describes the results obtained from the validation methods by using the three method adopted in this study (UT, HE, DSI). It starts by detailing the UT results separately. Then, the results of the HE and DSI methods will be explained in detail. Finally, all the results derived from the three methods are compared in terms of usability evaluation method measures, sample sizes, and statistical tests.

5.6.1 Quantitative and qualitative UT data analysis

The results of this method were analysed first because it is considered the gold standard for comparison (Woolrych et al., 2004). This section explores the users' profile data and describes the results collected from UT in terms of each group's performance. The different usability measures are also explored. Before starting a deeper analysis, the users' profiles is explored, as follows.

5.6.1.1 Users' profiles

- Educational websites

For the first experiment, Table 5.1 shows the frequency distribution for the users' profiles. The majority of them are students, according to on the website owners. 51.7% have experience of more than 4 years in using computers, and 51.7% use a computer in their daily work more than 4 hours. Also, 60% of the users have experience of more than 4 years in using the Internet, and 63.3% of them less often visit educational websites.

Table 5.1: Distribution of user groups in terms of their profile in the first experiment

Question	Frequency			Percentage		
Years using a computer	Less than 1 year	1 to 4 years	More than 4 years	Less than 1 year	1 to 4 years	More than 4 years
	6	*23*	*31*	*10%*	*38.3%*	*51.7%*
Daily hours on computer	Less than 1 hour	1 to 4 hours	More than 4 hours	Less than 1 hour	1 to 4 hours	More than 4 hours
	14	*15*	*31*	*23.3%*	*25%*	*51.7%*
Browser	Internet Explorer	Google Chrome	Firefox Mozilla	Internet Explorer	Google Chrome	Firefox Mozilla
	15	*35*	*10*	*25%*	*58.3%*	*16.7%*
Years using the Internet	Less than 1 year	1 to 4 years	More than 4 years	Less than 1 year	1 to 4 years	More than 4 years
	11	*13*	*36*	*18.3%*	*21.7%*	*60%*
Daily hours on the Internet	Less than 1 hour	1 to 4 hours	More than 4 hours	Less than 1 hour	1 to 4 hours	More than 4 hours
	4	*15*	*41*	*6.7%*	*25%*	*68.13%*
Daily hours visiting educational websites	Daily	Less often		Daily	Less often	
	22	*38*		*36.7*	*63.3*	
Total	60			60		

- Social networks websites

For the second experiment, Table 5.2 shows the frequency distribution for the users' profiles. The majority of them are students, according to their responses with regard to participating in this experiment. 94.6% have experience of more than 4 years in using computers and the Internet, and 85.4% use a computer in their daily work. Also, 100% of them daily visit social network websites.

Table 5.2: Distribution of user groups in terms of their profile in the second experiment

Question	Frequency			Percentage		
Years using a computer	Less than 1 year	1 to 4 years	More than 4 years	Less than 1 year	1 to 4 years	More than 4 years
	0	4	71	0%	5.4%	94.6%
Daily hours on a computer	Less than 1 hour	1 to 4 hours	More than 4 hours	Less than 1 hour	1 to 4 hours	More than 4 hours
	0	11	64	0%	14.6%	85.4%
Browser	Internet Explorer	Google Chrom e	Firefox Mozilla	Internet Explorer	Google Chrome	Firefox Mozilla
	8	44	23	10.6	58.6%	30.8%
Years using the Internet	Less than 1 year	1 to 4 years	More than 4 years	Less than 1 year	1 to 4 years	More than 4 years
	0	4	71	0%	5.4%	94.6%
Daily hours on the Internet	Less than 1 hour	1 to 4 hours	More than 4 hours	Less than 1 hour	1 to 4 hours	More than 4 hours
	0	11	64	0%	14.6%	85.4%
Daily hours visiting social network websites	Daily	Less often		Daily	Less often	
	75	0		100%	0%	
Total	75			75		

5.6.1.2 Time spent

The time on task in minutes and/or seconds is an excellent method to measure the efficiency attribute (Albert and Tullis, 2013). In this regard, this section shows the time spent by users in both experiments.

- Educational websites

In terms of time measure for the first experiment, Table 5.3 shows the time spent by each user on performing the experiment. The Skoool groups spent the longest time, more than the BBC KS3bitesize and AcademicEarth groups, with 112, 96 and 88 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 Skoool website, as some tasks were abandoned because the users had doubts about how to accomplish them. Also, in the BBC KS3bitesize website, the group spent time thinking about how to perform some tasks, such as the 'registration' task and the 'post a question' task. The average time spent by each user in all three groups was more than 1.1 minutes. The efficiency formula was used for UT, in terms of number of usability problem discovered over time spent for each group, the mean score was 0.4 (Skoool = 0.1, AcademicEarth = 0.1, BBC KS3bitesize = 0.2), as shown in Table 5.4. This result will be compared later to the results of HE and DSI in the educational websites. There is one question that needs to be examined statistically: is there any correlation between time spent and number of problem discovered? The next section (5.6.1.3) answers this question.

Table 5.3: Time taken on conducting the evaluation in the first experiment

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

Table 5.4: Total efficiency score for UT in the first experiment

Method	Skoool	AcademicEarth	BBC KS3bitesize	Total
	Efficiency	Efficiency	Efficiency	
UT	0.12	0.14	0.16	0.4

- Social networks websites

In terms of time measure for the second experiment, Table 5.5 shows the time spent by users 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 again was probably due to problems in navigation, structure and function in the three websites, which caused the users to spend more time in accomplishing their tasks. This was particularly so in the Google+ website, as some tasks were abandoned because the users had doubts about how to accomplish them, such as Tasks 2 and 4 as shown at Appendix D. Also, in the LinkedIn website, the group spent time thinking about how to perform some tasks, such as Tasks 2, 3 and 4. The average time spent by each user in all three groups was more than 02.48 minutes. From the efficiency formula used for UT, in terms of number of usability problems

discovered over time spent for each group, the mean score was 0.514 (Google+ = 0.567, LinkedIn = 0.556, Ecademy = 0.419), as shown in Table 5.6. This result will be compared later to the results of HE and DSI. There is one question that needs to be examined statistically: is there any correlation between time spent and number of problems discovered? The next section (5.6.1.3) answers this question.

Table 5.5: Time taken on conducting the evaluation in the second experiment

Usability measure	Google+	LinkedIn	Ecademy
Total time spent by all users (in minutes)	429	377	372
Average time per user per task (in minutes)	2.86	2.51	2.48
Average time per user over six tasks	17.16	15.08	14.88

Table 5.6: Total efficiency score for UT in the second experiment

Method	Google+	LinkedIn	Ecademy	Mean
	Efficiency	Efficiency	Efficiency	
UT	0.567	0.556	0.419	0.514

5.6.1.3 Number of usability problems discovered

The number of usability problems and their severity are the most important measures between usability methods (Hertzum and Jacobsen, 2001). Thus;

- Educational websites

In terms of the first experiment, the analysis of the three groups' performances in this study reveals a number of interesting results, and these lead to achieving the main research objective. It can be seen that using good task design does indeed play a role in finding different usability problems with different types. Table 5.7 explains the total usability problems found by UT and their severity rating. It shows that each user group revealed different severity levels and numbers of usability problems. All the redundant problems were removed. In the Skoool website, the total number of usability problem found is 13; there are 1 catastrophic, 3 major, 2 minor and 7 cosmetic ones. In the AcademicEarth website, the total number of usability problems found is 12; there are 3 major, 2 minor and 7 cosmetic ones. In the BBC KS3bitesize website, the total number of usability problem found is 16; there are 2 major, 5 minor and 9 cosmetic ones. Overall, the total number of usability problems discovered by UT is 41; there are 1 catastrophic, 8 major, 9 minor and 23 cosmetic ones. The

usability problems detected in BBC KS3bitesize is 16, higher than in the Skoool and AcademicEarth websites (13 vs. 12). These problems are listed in Appendix N.

Table 5.7: Number of usability problems discovered in the first experiment

Problem type	Skoool	AcademicEarth	BBC KS3bitesize	Total problems without duplication
	Total usability problems	Total usability problems	Total usability problems	
Catastrophic	1	0	0	1
Major	3	3	2	8
Minor	2	2	5	9
Cosmetic	7	7	9	23
No. of problems	13 (32%)	12 (29%)	16 (39%)	41

There is one question on the above findings that need to be statistically examined. This question: is there any relationship between time spent and problems found on the result of this experiment? Pearson Correlation was used as mentioned in Section 4.6.3 and the result reveals that there is a positive relationship between time spent and problems discovered; the p- value is 0.013 (which is less than 0.05) and this means that there is a significant correlation, as shown in Table 5.8. This result reveals that the users who spent more time faced more problems, and thus they were able to discover more usability problems.

Table 5.8: Pearson Correlation test between time spent and problems found in the first experiment

Relationship	Pearson Correlation	Sig. (2-tailed)
Time spent and problems found	0.318	0.013

- Social networks websites

In terms of the second experiment, Table 5.9 explains the total usability problems found by UT and their severity rating. It shows that each user group revealed different severity levels and numbers of usability problems. All the redundant problems were removed. In the Google+ website, the total number of usability problem found is 34; there are 4 catastrophic, 9 major, 11 minor and 10 cosmetic ones. In the LinkedIn website, the total number of usability problems found is 26; there are 2 catastrophic, 5 major, 8 minor and 11 cosmetic ones. In the Ecademy website, the total number of usability problems found is 19; there are 3 major, 6 minor and 11 cosmetic ones. Overall, the total number of usability problems discovered by UT is 79; there are 6 catastrophic, 17 major, 25 minor and 32 cosmetic ones.

The usability problems detected in Google+ number 34, which is higher than in the LinkedIn and Ecademy websites (26 vs. 19). These problems are listed in Appendix S.

Table 5.9: Numbers of usability problems discovered in the second experiment

Problem type	Google+	LinkedIn	Ecademy	Total problems without duplication
	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	10	31
No. of problems	34 (43%)	26 (33%)	19 (24%)	79

There is one question on the above findings that need to be statistically examined. This question; is there any relationship between time spent and problems found in this experiment? Pearson Correlation was used as mentioned in Section 4.6.3 and the results reveal that there is a positive relationship between time spent and problems discovered, where $p < 0.05$ and this means that there is a significant correlation, as shown in Table 5.10. Again,these results reveal and confirm that the users who spent more time faced more problems, and thus they were able to discover more usability problems.

Table 5.10: Pearson Correlation test between time spent and problems found in the second experiment

Relationship	Pearson Correlation	Sig. (2-tailed)
Time spent and problems found	0.566	0.000

5.6.1.4 User Satisfaction

- Educational websites

For the first experiment and after analysing the satisfaction questionnaire, it can be clearly seen that, although BBC KS3bitesize has more problems, it delivered the highest overall score, at 7, whereas Skoool delivered the second highest score, at 5, and AcademicEarth delivered the lowest score among the three websites, at 3. In conclusion, it indicates that there were certain factors that influenced the users, which then affected the satisfaction rating for the tested website, as evidenced by the critical user comments on the design features of each website. These factors are the various activities, such as the test and revise functions that each website provided (or the games); also, the users were encouraged by simple and attractive designs.

- Social networks websites

For the second experiment and after analysing the satisfaction questionnaire, it can be clearly seen that, although LinkedIn has more problems, it delivered the highest overall score, at 5.56, whereas Google+ delivered the second highest score, at 4.52, and Ecademy delivered the lowest score among the three websites, at 4.16. These results reveal that the usability problems discovered by the users did not affect their level of satisfaction. This may explain why LinkedIn delivered the highest satisfaction score, whereas it is the second website in terms of high numbers of usability problems found. Google+ is the top website in terms of number of usability problem found, whereas it delivered the second highest satisfaction score. In conclusion, it indicates that there were certain factors that influenced the users, which then affected the satisfaction rating for the tested website, as evidenced by the critical user comments on the design features of each website. These factors are the various activities that each website provided, such as uploading CV, seeking for jobs, easy posting and connection, and simple and attractive designs, as the user comments.

5.6.2 Quantitative and qualitative HE and DSI data analysis

The evaluators' profile data were explored in order to offer a general idea and overview of the nature of the types of evaluators involved in this experiment.

5.6.2.1 Evaluators' profiles

- Educational websites

For the first experiment, Table 5.11 shows the evaluators' frequency distribution for evaluator type, level of education and years of experience. Approximately 3 years was the mean for their evaluation experience across the two groups.

Table 5.11: Distribution of evaluator profiles in the first experiment

Evaluator group	Evaluator identification	Evaluator type	Level of education	Years of experience
G1	1	Double	Master	3
	2	Double	PhD	5
	3	Single	Master	2
	4	Single	Master	3
G2	1	Single	Master	2
	2	Single	Master	4
	3	Double	Master	4
	4	Double	PhD	3
Total	8	Mean (years)		3.25

- Social networks websites

For the second experiment, Table 5.12 shows the evaluators' frequency distribution for evaluator type, level of education and years of experience. Approximately 4 years was the mean for their evaluation experience across the two groups.

Table 5.12: Distribution of evaluator profiles in the second experiment

Evaluator group	Evaluator identification	Evaluator type	Level of education	Years of experience
G1	1	Double	Master	3
	2	Double	PhD	7
	3	Single	Master	2
G2	1	Double	PhD	5
	2	Double	Master	4
	3	Single	Master	3
Total	6	Mean (years)		4

5.6.2.2 Time spent

- Educational websites

For the first experiment, time spent was calculated as the time taken by an evaluator to complete an evaluation, as shown in Tables 5.13 and 5.14. Table 5.15 shows the average time taken for doing the three evaluations using HE was 24.25 minutes with a standard deviation of 6.7, whereas the DSI average was 42.58 minutes with a standard deviation of 7.19. In this regard, there are two questions. The first one: is there a significant difference between the times spent by HE and DSI? To answer this question, the normality test should be conducted to choose the correct test (Cairns and Cox, 2008). As mentioned in Section 4.6.3, Ghasemi and Zahediasl (2012, p.487) pointed out to '' for small sample sizes, normality tests have little power to reject the null hypothesis and therefore small samples most often pass normality tests''. Based on this it can be assumed that the data is not normally distributed. Thus, Mann-Whitney was used at a significant level of 5% (0.05) as mentioned in Section 4.6.3, and it revealed that there is a strong likelihood for a significant difference between DSI and HE in terms of time spent because of the p-value = 0.000 ($P < 0.05$) as shown in Table 5.15. Then, Bonferroni test was used as mentioned in Section 4.6.3 by multiplying the p-value by 2, thus the corrected p-value = 0.000 ($P < 0.05$). The second question: is there any impact of the time spent by HE and DSI on the results of number of usability problem found? To answer this question, Mann-Whitney was used, and it revealed that there is a strong likelihood for a significant difference when p-value = 0.001 ($P< 0.05$) as shown in Table 5.16. Then, Bonferroni was used by multiplying the p-value by 2, so the corrected p-value =

0.002 ($P < 0.05$). Thus, the number of usability problem found was impacted by time spent. This means that if an evaluator spends more time, the number of problems discovered will increase. This assumption was proved; for example, the group who used HE managed to evaluate the website more quickly than the other group but discovered fewer usability problems. On the other hand, the group that used DSI spent almost double the time evaluating the website, but discovered almost three times as many usability problems. On this point, the question to confirm this result is: is there any relationship between the time spent by HE and DSI and the number of problems found? The Pearson correlation test was employed. The result shows that there is a significant positive relationship between time spent and problems discovered where the p-value is 0.020 at the 0.05 level, as shown at Table 5.16. This result reveals statistically that an evaluator who spent more time was able to discover more usability problems. An explanation for the differences in time spent and number of problems located is gleaned from the evaluators' feedback. They said that HE was not particularly helpful, understandable or memorable for them. However, DSI helped them to develop their skills in discovering usability problems in this application area; also, this set was more understandable and memorable during their evaluation. To further analyse these factors of time spent and number of problems discovered, efficiency metrics were applied. DSI proved to be more efficient than HE in discovering usability problems (DSI = 0.6 vs. HE = 0.4) as Table 5.17 shows. This result will be compared later against the UT results.

Table 5.13: Average time taken and number of problems by Group 1

Website	Skoool	AcademicEarth	BBC KS3bitesize
Evaluator	Time	Time	Time
1	25	45	24
2	30	50	40
3	25	55	22
4	20	29	23
Heuristics	HE	DSI	HE
# of problems	10	29	2
Mean time taken	25	45	27

Table 5.14: Average time taken and number of problems by Group 2

Website	Skoool	AcademicEarth	BBC KS3bitesize
Evaluator	Time	Time	Time
1	42	30	50
2	40	17	38
3	38	15	35
4	45	20	44
Heuristics	DSI	HE	DSI
# of problems	33	13	12
Mean time taken	41	21	42

Table 5.15: Mann Whitney U test on time spent for both methods in the first experiment

Variable	Methods	N	Mean	Std. Deviation	Std. Error Mean	T-test (Sig. (2-tailed))
Time spent	HE	12	24.25	6.717	1.939	0.000
	DSI	12	42.58	7.192	2.076	

Table 5.16: Mann Whitney U test and correlations between time spent and problems found in the first experiment

Test	Time Spent	No. of problems
Mann-Whitney U	6.500	12.500
Z	-3.787	-3.455
Asymp. Sig. (2-tailed)	0.000	0.001
Exact Sig. [2*(1-tailed Sig.)]	0.000	0.000
Pearson Correlation	0.473	
Sig. (2-tailed)	0.020	

Table 5.17: Mann Whitney U of efficiency attribute for two methods in the first experiment

Method	Skoool	AcademicEarth	BBC KS3bitesize	Mean
	Efficiency	Efficiency	Efficiency	
HE	0.4	0.6	0.1	0.4
DSI	0.8	0.7	0.3	0.6

- Social networks websites

For the second experiment, time spent was calculated as the time taken by an evaluator to complete an evaluation, as shown in Tables 5.18 and 5.19. Table 5.20 shows that the average time taken for doing the three evaluations using HE was 56 minutes, with a standard deviation of 8.81, whereas the DSI average was 72 minutes, with a standard deviation of 10.03. In this regard, there are two questions. The first one: is there a significant difference between the times spent by HE and DSI? To answer this question, the normality test should be conducted by using Shapiro-Wilk test. As mentioned in Section 4.6.3, Ghasemi and Zahediasl (2012, p.487) pointed out to ‘’for small sample sizes, normality tests have little power to reject the null hypothesis and therefore small samples most often pass normality tests’’. Based on this it can be assumed that the data is not normally distributed. Thus, Mann-Whitney was used at a significant level of 5% (0.05), and it revealed that there is a strong likelihood for a significant difference (P= 0.003), as shown in Table 5.18. Then, Bonferroni was used by multiplying the p-value by 2, so the corrected p-value = 0.006 ($P < 0.05$). The second question: is there any impact of the time spent by evaluators using the two methods on the results of number of usability problems found? To answer this question, Mann-Whitney was again employed. The results show that there is a strong likelihood for a significant difference

when p= 0.001 (P< 0.05), as shown in Table 5.21. Then, Bonferroni was used by multiplying the p-value by 2, so the corrected p-value = 0.002 (P < 0.05). Thus, the number of usability problems found was impacted by time spent. This means that if an evaluator spends more time, the number of problems discovered will increase. This assumption was proved; for example, the group who used HE managed to evaluate the website more quickly than the other group but discovered fewer usability problems. On the other hand, the group that used DSI spent almost double the time evaluating the website, but discovered many usability problems. On this point, the question is: is there any relationship between the time spent by the evaluators and the number of problems found? The Pearson Correlation test was employed. The results show that there is a significant positive relationship between time spent and problems discovered, where the p-value is 0.041 at the 0.05 level, as shown at Table 5.21. This result reveals statistically that an evaluator who spent more time was able to discover more usability problems. The 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). This result will be compared later against the UT results.

Table 5.18: Average time taken and number of problems by Group 1

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

Table 5.19: Average time taken and number of problems by Group 2

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

Table 5.20: Mann Whitney U test on time spent for both methods in the second experiment

Variable	Method	N	Mean	Std. Deviation	Std. Error Mean	Mann Whitney U (Sig. (2-tailed))
Time spent	HE	9	56	8.81	2.93	0.003
	DSI	9	72	10.03	3.34	

Table 5.21: Mann Whitney U test and correlations between time spent and problems found in the second experiment

Test	Time Spent	No. of problems
Mann-Whitney U	8.000	4.500
Z	-2.937	-3.189
Asymp. Sig. (2-tailed)	0.003	0.001
Exact Sig. [2*(1-tailed Sig.)]	0.003	0.000
Pearson Correlation	-0.541	
Sig. (2-tailed)	0.025	

Table 5.22: Mann Whitney of efficiency attribute for two methods in the second experiment

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

5.6.2.3 Number of usability problems discovered

The number of usability problems and their severity are the most important measures between usability methods (Hertzum and Jacobsen, 2001). Thus,

- Educational websites

For the first experiment, the analysis of the two methods' performances in this study reveals a number of interesting results, and this leads to achieving the main research objective. The usability problems discovered were extracted from the structured usability report and their figures are presented in Table 5.23. It can be seen that using different inspection methods does indeed play a role in finding different usability problems with different types. Thus, HE was able to uncover 25% of the total number of usability problems in the three websites, and this result is in line with previous findings which claimed that HE found around 29% of usability problems (Nielsen and Landauer, 1993). However, DSI was able to uncover 75% of the total number of usability problems in the three websites. In this regard, the main statistical question is: to what extent are there statistical differences among the methods' performances, in particular, with regard to the problems found? To answer this question, Mann-Whitney was used at a significant level of 5% (0.05) because of using small sample

sizes (Zahediasl, 2012). Thus, Mann-Whitney shows that there is a strong likelihood for a significant difference in terms of number of usability problems found, where $P< 0.05$ ($p= 0.001$) as shown in Table 5.23. Then, Bonferroni correction was used by multiplying the p-value by 2, so the corrected p-value = 0.002 ($P < 0.05$). In detail, HE revealed 10 problems out of 43 in the Skoool website, representing 23%. However, DSI revealed 33 problems out of 43, representing 77% on the same website. In the second website (AcademicEarth), the total number of problems was 42, and the biggest number found was for DSI, which revealed 29 problems (69%), whereas HE revealed only 13 problems (31%). In the third website (BBC KS3bitesize), HE revealed 2 unique problems (representing 14%) and DSI revealed 12 problems out of the 14 unique problems (representing 86%). Overall, the above results confirm that there is difference between the groups' performance in discovering usability problems in the same website by using different methods. This is known as the method effect. Consequently, DSI method affects positively the the groups' performance by discovering more usability problems.

Table 5.23: Summary (numbers and percentages) of usability problems uncovered on each website, by each group, each evaluator and each method in the first experiment

Website	Group	Expert and type	Method	# of problems found by each evaluator	# of problems with repetition	# of problems without repetition	# of problems with repetition between groups	% of problems found by each evaluator	% # of problems found by each group
Skoool	G 1	Ev.1+	HE	8	19	10	43	19%	23%
		Ev.2^	HE	5				12%	
		Ev.3^	HE	1				2%	
		Ev. 4+	HE	5				12%	
	G 2	Ev.1+	DSI	21	64	33		49%	77%
		Ev. 2^	DSI	15				35%	
		Ev. 3+	DSI	13				30%	
		Ev.4^	DSI	15				35%	
Academic Earth	G 1	Ev. 1+	DSI	10	42	29	42	24%	69%
		Ev. 2^	DSI	11				26%	
		Ev. 3^	DSI	9				21%	
		Ev.4+	DSI	12				29%	
	G 2	Ev. 1+	HE	6	23	13		14%	31%
		Ev. 2^	HE	6				14%	
		Ev. 3+	HE	7				17%	
		Ev. 4^	HE	4				10%	
BBC KS3bitesize	G 1	Ev. 1+	HE	1	5	2	14	7%	14%
		Ev. 2^	HE	1				7%	
		Ev. 3^	HE	1				7%	
		Ev. 4+	HE	2				14%	
	G 2	Ev. 1+	DSI	6	24	12		43%	86%
		Ev. 2^	DSI	6				43%	
		Ev. 3+	DSI	7				50%	
		Ev. 4^	DSI	5				36%	

Total number of Usability Problems Discovered by each Heuristics	Heuristics	Total number	Approx. %
	HE	25	25%
	DSI	74	75%
Mann-Whitney U	Mann-Whitney U	Z	P-value
	12.500	-3.455	0.001

+ Double Expert ^ Single Expert Ev. = Evaluator

One striking result is that the number of problems identified by each evaluator who used HE is always less than the number of problems identified by any evaluator using DSI for the same site. An explanation of this was found in the evaluator answers in the questionnaire. One of evaluators said that ‘’the HE set was difficult to use, did not remind me of aspects they might have forgotten about, and I did not believe that this set encouraged me to be

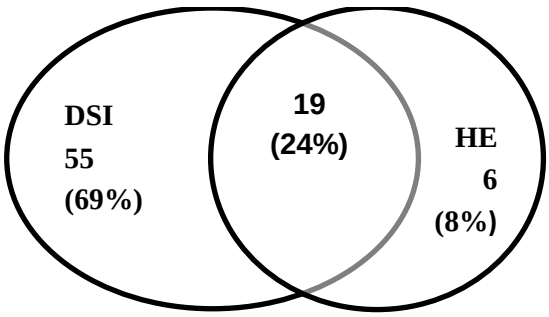
thorough in their evaluation''. In contrast, he said that ''the DSI set was easy to use; it did indeed help me to remember all the functions that needed to be tested, it is specific and was designed to cover all the aspects needed for educational websites''. In addition, the reliability of DSI and HE was measured and the result reveals that there is a slight difference between DSI and HE (0.5 vs. 0.4). Table 5.24 shows a comparison of these results to other published results. However, a high reliability score does not guarantee that a method discovers all the usability problems in a user interface, as claimed by (Hertzum and Jacobsen, 2001).

Table 5.24: Experiment results for reliability compared with some published results

	(Law and Hvannberg, 2004)	(Nielsen, 1992a)	(Nielsen and Molich, 1990)	This experiment	
Reliability of HE	0.32	0.45	0.26	HE	0.4
				DSI	0.5

In terms of the performance of each method in discovering unique and overlapping problems, Table 5.23 illustrates the total number of problems discovered, which was 99 on the three websites, out of which 25 were identified using HE and 74 using DSI. All the duplicated problems were removed and compared by two independent evaluators, in order to identify the unique and overlapping problems. When the problems from the two evaluation groups were consolidated, there were 19 duplicates; thus, we identified a total of 80 problems in all websites. The total for uniquely identified problems in all websites was 61 problems. DSI identified 55 problems (69% of the 80 problems) that were not identified by HE, and there were 6 problems (8% out of 80) identified by HE that were not identified by DSI. 19 problems (24%) out of 80 problems were discovered by both methods (as depicted in Figure 5.1).

Figure 5.1: Overlap between both methods (HE and DSI) for the first experiment



In regarding to the severity of problems discovered, the two independent evaluators were involved to rank the usability problems discovered. Table 5.25 shows the severity rating of the problems discovered (cosmetic, minor, major and catastrophic). A great many usability problems were discovered with differing levels of severity but the most important results were obtained from using the DSI method.

Table 5.25: Total number of usability problems with severity ratings and averages

Website	Problem Severity	Type of Method			
		HE		DSI	
Skoool	Cosmetic	Group 1	3	Group 2	12
	Minor		3		11
	Major		2		6
	Catastrophic		2		4
Severity (average)		2.3		2.1	
AcademicEarth	Cosmetic	Group 2	2	Group 1	11
	Minor		7		11
	Major		3		4
	Catastrophic		1		3
Severity (average)		2.2		2	
BBC KS3bitesize	Cosmetic	Group 1	2	Group 2	2
	Minor		0		6
	Major		0		3
	Catastrophic		0		1
Severity (average)		1		1.9	
Overall Severity (average)		1.8		2.1	
No. of discovered problems		25		74	

- Social networks websites

For the second experiment, the usability problems discovered were extracted from the structured usability report, and their figures are presented in Table 5.26. It can also be seen that using different usability methods does indeed play a role in finding different usability problems with different severity rating. Thus, HE was able to uncover 30% of the total number of usability problems in the three websites, and this result is in line with previous findings, which claimed that HE found around 29% of usability problems (Nielsen and Landauer, 1993). However, DSI was able to uncover 70% of the total number of usability problems in the three websites. The main statistical question is: to what extent are there statistical differences among the methods' performances, in particular with regard to the problems found? To answer this question Mann-Whitney again employed at a significant level of 5% (0.05). It shows that there is a strong likelihood for a significant difference in

terms of number of usability problems found, where P< 0.05 (p= 0.001), as shown in Table 5.26. Then, Bonferroni correction was used by multiplying the p-value by 2, so the corrected p-value = 0.002 (P < 0.05).

Table 5.26: Summary (numbers and percentages) of usability problems uncovered on each website, by each group, each evaluator and each method in the second experiment

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	48	69	28%	69%
		Ev.2+	DSI	33				48%	
		Ev.3+	DSI	17				25%	
	G 2	Ev.1+	HE	6	22	22		9%	31%
		Ev. 2^	HE	5				7%	
		Ev. 3+	HE	11				16%	
LinkedIn	G1	Ev. 1^	HE	6	21	21	67	9%	31%
		Ev.2+	HE	8				12%	
		Ev.3+	HE	10				15%	
	G 2	Ev.1+	DSI	24	59	46		35%	69%
		Ev.2^	DSI	8				12%	
		Ev.3+	DSI	27				40%	
Ecademy	G 1	Ev.1^	DSI	6	57	25	37	16%	68%
		Ev.2+	DSI	28				75%	
		Ev.3+	DSI	23				62%	
	G 2	Ev.1+	HE	5	12	12		14%	32%
		Ev.2^	HE	3				8%	
		Ev.3+	HE	4				10%	

Total number of usability problems discovered by each method	Methods	Total number	%
	HE	55	32%
	DSI	119	68%

Mann-Whitney U	Mann-Whitney U	Z	P-value
	-4.500	-3.189	0.001

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

The notable result in this experiment is that the number of problems identified by each evaluator who used HE is always less than the number of problems identified by any evaluator using DSI for the same site. An explanation of this was found in the evaluator answers in the questionnaire. One of evaluators said, "HE is general and each heuristic can be understood in various ways, which makes the evaluation so difficult and leads to discovering fewer problems". In contrast, he also said, "DSI is designed to be used within

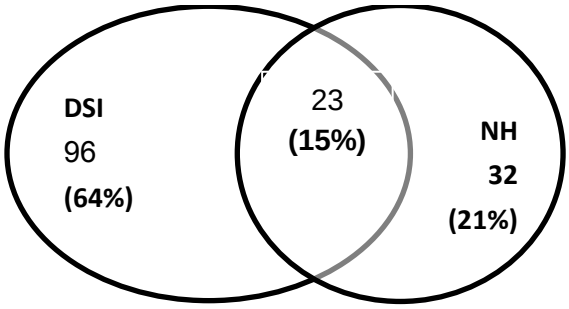
the context of social network websites, and their principles were categorised for each feature, covering all the aspects needed". In addition, the reliability of DSI and HE was measured and the results reveal that there is a slight difference between DSI and HE (0.38 vs. 0.27). Table 5.27 shows a comparison of these results to other published results.

Table 5.27: Experiment results for reliability compared with some published results

	(Law and Hvannberg, 2004)	(Nielsen, 1992a)	(Nielsen and Molich, 1990)	This experiment	
Reliability of HE	0.32	0.45	0.26	HE	0.27
				DSI	0.38

In terms of the performance of each method in discovering unique and overlapping problems, Table 5.26 illustrates the total number of problems discovered, which was 174 on the three websites, out of which 55 were identified using HE and 119 using DSI. All the duplicated problems were removed and compared by two independent evaluators, in order to identify the unique and overlapping problems. When the problems from the two evaluation groups were consolidated, there were 23 duplicates; thus, we identified a total of 151 problems in all websites. The total for uniquely identified problems in all websites was 128 problems. DSI identified 96 problems (64% of the 151 problems) that were not identified by HE, and there were 32 problems (21% out of 151) identified by HE that were not identified by DSI. 23 problems (15%) out of 151 problems were discovered by both methods (as depicted in Figure 5.2).

Figure 5.2: Overlap between both methods (DSI and HE) for the second experiment



In regarding to the severity of problems discovered, the two independent evaluators were involved to rank the usability problems discovered. Table 5.28 shows the severity rating of the problems discovered (cosmetic, minor, major and catastrophic). A great many usability

problems were discovered with differing levels of severity but the most notable results were obtained again from using the DSI method.

Table 5.28: Total number of usability problems with severity ratings and averages

Website	Severity Problems	Types of Method			
		DSI		HE	
Google+	Cosmetic	Group 1	16	Group 2	6
	Minor		28		13
	Major		11		3
	Catastrophic		0		0
	Severity (average)	1.9		1.8	
LinkedIn	Cosmetic	Group 2	11	Group 1	8
	Minor		17		8
	Major		6		5
	Catastrophic		5		0
	Severity (average)	2.8		1.3	
Ecademy	Cosmetic	Group 1	12	Group 2	4
	Minor		7		8
	Major		6		0
	Catastrophic		0		0
	Severity (average)	1.7		1.6	
Overall Severity (average)		2		2	
No. of discovered problems		127		55	

5.6.2.4 Areas of the usability problems found

- Educational websites

A qualitative assessment was conducted to compare the two methods, in particular, in terms of the areas of the usability problems found in this experiment. These areas assisted in identifying how each method performed (in each usability problem area or category of heuristics). This can be done by matching the discovered problems to their categories in the HE list of heuristics or in the DSI checklist, as recommended by (Nielsen, 1995a). This can help to identify how each method performs in each category of the guidelines. Also, it can be used to compare the results found here with current published work. The eight expert evaluators discussed and agreed upon the problem list during the debriefing session. Then, the independent evaluators decided on the categories to which the problems should belong (in both methods), as Tables 5.29 and 5.30 illustrate. The overall results from both tables show that the two groups (and the three websites) revealed more usability problems by using DSI in all areas than HE, particularly in 'Learning process', 'Design and Media usability', 'Motivational factors', 'User usability', and 'Content information and Process orientation'.

However, three out of the ten heuristics in HE worked more efficiently than four heuristics, and the remaining three failed to expose enough usability problems. This suggests that the HE list is rather general, and is unlikely to encompass all the usability attributes of user experience and design in interactive learning systems. The above results are in line with another study by Alsumait and Al-Osaimi (2009). Furthermore, Table 5.29 illustrates that 'Match between the system and the real world' and 'User control and freedom' are common weaknesses in dynamic websites as it was clear in one website out of three websites, and this in line with Chen and Macredie (2005).

Table 5.29: Usability problems found by category through HE in the first experiment

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

Table 5.30: Usability problems found by category through DSI in the first experiment

Usability problem areas	Skoool	AcademicEarth	BBC KS3bitesize
User usability	4	5	2
Motivational factors	5	6	1
Content information and Process orientation	4	3	0
Learning process	11	7	6
Design and Media usability	9	8	3
Total problems	33	29	12

- Social networks websites

A qualitative assessment was conducted to compare the two methods, in particular in terms of the areas of the usability problems found in this experiment. The six expert evaluators discussed and agreed upon the problem list during the debriefing session. Then, the two independent evaluators decided on the categories to which the problems should belong (in both methods), as Table 5.31 and 5.32 illustrate. The overall results from both tables show

that the two groups (and the three websites) revealed more usability problems by using DSI in all areas than HE, particularly in User usability, Sociability and management activities, Content quality, Navigation system and search quality, and Layout and formatting. Both methods found the most problems in the area entitled User usability, sociability and management activities (DSI) or User control and freedom (HE). These were followed in DSI by Content quality, Navigation system and search quality, Layout and formatting, and Security and privacy; these areas were not discovered efficiently or sufficiently by HE. This suggests that HE is rather general, and is unlikely to encompass all the usability attributes of user experience and design. The above results are in line with other studies that find that the layout, organization and structure of the buttons and contents are the main problems in social websites such as LinkedIn and Google+ (Al-Badi et al., 2013); (Mao et al., 2011). Also, it is in line with Chen and Macredie (2005)'s study, which reported that user control and freedom is a common weakness in dynamic websites.

Table 5.31: Usability problems found by category through HE in the second experiment

Nielsen's Heuristics	Google+	LinkedIn	Ecademy
Visibility of system status	2	4	5
Match between the system and the real world	4	3	1
User control and freedom	5	5	1
Consistency and standards	3	2	0
Error prevention	0	1	0
Recognition rather than recall	2	4	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	21	12

Table 5.32: Usability problems found by category through DSI in the second experiment

Usability problem area	Google+	LinkedIn	Ecademy
Layout and formatting	6	2	4
Content quality	9	11	3
Security and privacy	1	3	2
Business support	0	1	3
User usability, sociability and management activities	21	20	9
Accessibility and compatibility	1	1	2
Navigation system and search quality	10	8	2
Total problems	48	46	25

5.6.2.5 Rating scale questionnaire

- Educational websites

For the first experiment, Table 5.33 shows the numbers of evaluators who answered each item for each method. It can be seen that most of the evaluators responded as 'strongly disagree' and 'probably disagree' on HE items (numbers 1, 3, 4, 5, 6 and 7), whereas most of the evaluators responded as 'strongly agree' and 'probably agree' on the DSI items (numbers 1, 2, 3, 4, 5, 7 and 8). For further explanations on the above results and Table 5.34, the evaluators stated in the post-test questionnaire that they (100%) would now prefer to use DSI in future evaluations (only for educational websites) rather than HE. Also, 75% of the evaluators said that ''the DSI are easier to understand and use than HE. They explained the reasons for this: HE did not help them to remember all the functions that they needed to test''. Although the HE set was much faster to use, it did not encourage them to be thorough in their evaluations, as ''it does not offer any hints''. On the other hand, ''DSI reminded us to the most important aspects, and this was useful in building a check-list''. They (62.5%) also liked the way that DSI encourages them to be thorough in an evaluation, as this offers them more opportunities for finding usability problems. Furthermore, they (87.5%) concluded that ''a specific method, such as DSI, is better than any general method, such as HE, because specific ones can help everyone to focus on those criteria that are important to a particular kind of website over any other kind of website, and they integrate all the most suitable usability considerations, just as DSI does''. Table 5.35 shows the average percentages of overall responses of the results for HE and DSI for positive items, which are 1, 2, 3, 4, 5, 7 and 8. It can be clearly seen that 'HE' achieved a lower overall percentage of 'agree' responses, at 23%, whereas DSI achieved a higher overall percentage of 'agree' responses, at 76%. This means that the DSI method has a higher satisfaction percentage than the HE method.

Table 5.33: Number of evaluators who answered on each item in the first experiment

Statements	HE				
	Strongly disagree	**Probably disagree**	**Not sure**	**Probably agree**	**Strongly agree**
I think that I would like to use HE method frequently when I evaluate these websites.	3	3	1	1	0
I found the heuristic evaluation method unnecessarily complex	0	1	2	2	3
I think HE method was easy to use	2	4	1	1	0
The evaluation of these websites can be performed in a straightforward manner by using HE	2	4	0	1	1
I found the various principles in HE method to be well integrated and specific for these websites	6	2	0	0	0
I think there was too much inconsistency in HE method	3	3	1	1	0
I would imagine that most people would learn to use HE method very quickly	4	1	2	1	0
I felt very confident using HE method	0	1	4	3	0
I needed to learn a lot of things before I could get going with the HE method	0	1	2	0	5
Statements	**DSI**				
	Strongly disagree	**Probably disagree**	**Not sure**	**Probably agree**	**Strongly agree**
I think that I would like to use DSI method frequently when I evaluate these websites.	0	0	0	2	6
I found the DSI method unnecessarily complex	0	1	1	2	4
I think DSI method was easy to use	0	2	1	2	3
The evaluation of these websites can be performed in a straightforward manner by using DSI	0	1	2	2	3
I found the various principles in DSI method to be well integrated and specific for these websites	0	0	1	3	4
I think there was too much inconsistency in DSI method	3	4	0	1	0
I would imagine that most people would learn to use DSI method very quickly	0	1	1	5	1
I felt very confident using DSI method	0	1	1	3	3
I needed to learn a lot of things before I could get going with the DSI method	1	1	5	1	0

Table 5.34: Percentage representation of the results for HE and DSI per item in the first experiment

Statements	HE				
	Strongly disagree	**Probably disagree**	**Not sure**	**Probably agree**	**Strongly agree**
I think that I would like to use HE method frequently when I evaluate these websites.	37.5%	37.5%	12.5%	12.5%	0%
I found the heuristic evaluation method unnecessarily complex	0%	12.5%	25%	25%	37.5%
I think HE method was easy to use	25%	50%	12.5%	12.5%	0%
The evaluation of these websites can be performed in a straightforward manner by using HE	25%	50%	0%	12.5%	12.5%
I found the various principles in HE method to be well integrated and specific for these websites	75%	25%	0%	0%	0%
I think there was too much inconsistency in HE method	37.5%	37.5%	12.5%	12.5%	0%

I would imagine that most people would learn to use HE method very quickly	50%	12.5%	25%	12.5%	0%
I felt very confident using HE method	0%	12.5%	50%	37.5%	0%
I needed to learn a lot of things before I could get going with the HE method	0%	12.5%	25%	0%	62.5%
Statements	**DSI**				
	Strongly disagree	**Probably disagree**	**Not sure**	**Probably agree**	**Strongly agree**
I think that I would like to use DSI method frequently when I evaluate these websites.	0%	0%	0%	25%	75%
I found DSI method unnecessarily complex	0%	12.5%	12.5%	25%	50%
I think DSI method was easy to use	0%	25%	12.5%	25%	37.5%
The evaluation of these websites can be performed in a straightforward manner by using DSI	0%	12.5%	25%	25%	37.5%
I found the various principles in DSI method to be well integrated and specific for these websites	0%	0%	12.5%	37.5%	50%
I think there was too much inconsistency in DSI method	37.5%	50%	0%	12.5%	0%
I would imagine that most people would learn to use DSI method very quickly	0%	12.5%	12.5%	62.5%	12.5%
I felt very confident using DSI method	0%	12.5%	12.5%	37.5%	37.5%
I needed to learn a lot of things before I could get going with the DSI method	12.5%	12.5%	62.5%	12.5%	0%

Table 5.35: The average percentages of overall responses of the results of the positive items for HE and DSI in the first experiment

Response / Method	**Strongly disagree**	**Probably disagree**	**Not sure**	**Probably agree**	**Strongly agree**
HE	30%	29%	18%	16%	7%
DSI	0%	11%	13%	34%	42%

- Social networks websites

For the second experiment, Table 5.36 shows the numbers of evaluators who answered on each item for each method. It can be seen that most of evaluators responsed as 'strongly disagree' and 'probably disagree' on HE items (number 1,3,4,5,6 and 8), whereas most of evaluators responsed as 'strongly agree' and 'probably agree' on DSI items (number 1,2,3,4,5,7,8 and 9). The evaluators delivered this result because the process of evaluation was smoother by using DSI (because it was generated to cover all social network aspects), as they mentioned in their comments. For further explanations on the above results and Table 5.37, the evaluators stated in the post-test questionnaire that they (80.3% of them) would now prefer to use DSI in future evaluations (only for social websites) rather than HE. Also, 75% of the evaluators said, "DSI covers all functions in these websites, and it can be easy to update it, just by following the adaptive framework, to evaluate any new feature in the future".

Although the HE set was much faster to use, it did not encourage them to be thorough in their evaluations, as "it was difficult to evaluate some features". They (66.6%) also liked the way that DSI encourages them to be thorough in an evaluation, as this offers them more opportunities for finding usability problems. Furthermore, they (100%) concluded, "the DSI method encouraged us to evaluate more features than HE, and thus DSI helped us to discover more problems; that is, those problems not discovered by HE".

Table 5.38 shows the average percentages of overall responses of the results for HE and DSI for positive items, which are 1, 2, 3, 4, 5, 7 and 8. It can be clearly seen that 'HE' achieved again a lower overall percentage of 'agree' responses, at 14%, whereas DSI achieved a higher overall percentage of 'agree' responses, at 88%. This means that the DSI method has a higher satisfaction percentage than the HE method. Also, the percentage for satisfaction achieved by HE was because of the short period of time spent on it when it was used by the evaluators during the evaluation, not because it was helpful and efficient. On the other hand, the satisfaction percentage achieved by DSI was due to its efficiency in discovering more problems when it was used by the evaluators during the evaluation, as they mentioned in their comments.

Table 5.36: Number of evaluators who answered on each item in the second experiment

Statements	HE				
	Strongly disagree	**Probably disagree**	**Not sure**	**Probably agree**	**Strongly agree**
I think that I would like to use HE method frequently when I evaluate these websites.	3	2	1	0	0
I found the heuristic evaluation method unnecessarily complex	0	2	3	1	0
I think HE method was easy to use	2	2	2	0	0
The evaluation of these websites can be performed in a straightforward manner by using HE	0	3	2	1	0
I found the various principles in HE method to be well integrated and specific for these websites	6	0	0	0	0
I think there was too much inconsistency in HE method	1	3	1	1	0
I would imagine that most people would learn to use HE method very quickly	0	2	3	1	0
I felt very confident using HE method	1	2	2	1	0
I needed to learn a lot of things before I could get going with the HE method	0	2	1	3	0
Statements	**DSI**				
	Strongly disagree	**Probably disagree**	**Not sure**	**Probably agree**	**Strongly agree**
I think that I would like to use DSI method frequently when I evaluate these websites.	0	0	1	2	3
I found DSI method unnecessarily complex	0	1	1	2	2
I think DSI method was easy to use	0	1	1	3	1

The evaluation of these websites can be performed in a straightforward manner by using DSI	0	0	0	1	5
I found the various principles in DSI method to be well integrated and specific for these websites	0	0	0	2	4
I think there was too much inconsistency in DSI method	4	2	0	0	0
I would imagine that most people would learn to use DSI method very quickly	0	0	0	1	5
I felt very confident using DSI method	0	0	0	2	4
I needed to learn a lot of things before I could get going with the DSI method	0	0	2	3	1

Table 5.37: Percentage representation of the results for HE and DSI per item in the second experiment

Statements	**HE**				
	Strongly disagree	**Probably disagree**	**Not sure**	**Probably agree**	**Strongly agree**
I think that I would like to use HE method frequently when I evaluate these websites.	50%	33.3%	16.6%	0%	0%
I found the heuristic evaluation method unnecessarily complex	0%	33.3%	50%	16.6%	0%
I think HE method was easy to use	33.3%	33.3%	33.3%	0%	0%
The evaluation of these websites can be performed in a straightforward manner by using HE	0%	50%	33.3%	16.6%	12.5%
I found the various principles in HE method to be well integrated and specific for these websites	100%	0%	0%	0%	0%
I think there was too much inconsistency in HE method	16.6%	50%	16.6%	16.6%	0%
I would imagine that most people would learn to use HE method very quickly	0%	33.3%	50%	16.6%	0%
I felt very confident using HE method	16.6%	33.3%	33.3%	16.6%	0%
I needed to learn a lot of things before I could get going with the HE method	0%	33.3%	16.6%	50%	0%
Statements	**DSI**				
	Strongly disagree	**Probably disagree**	**Not sure**	**Probably agree**	**Strongly agree**
I think that I would like to use DSI method frequently when I evaluate these websites.	0%	0%	16.6%	33.3%	50%
I found DSI method unnecessarily complex	0%	16.6%	16.6%	33.3%	33.3%
I think DSI method was easy to use	0%	16.6%	16.6%	50%	16.6%
The evaluation of these websites can be performed in a straightforward manner by using DSI	0%	0%	0%	16.6%	83.3%
I found the various principles in DSI method to be well integrated and specific for these websites	0%	0%	0%	33.3%	66.6%
I think there was too much inconsistency in DSI method	66.6%	33.3%	0%	0%	0%
I would imagine that most people would learn to use DSI method very quickly	0%	0%	0%	16.6%	83.3%
I felt very confident using DSI method	0%	0%	0%	33.3%	66.6%
I needed to learn a lot of things before I could get going with the DSI method	0%	0%	33.3%	50%	16.6%

Table 5.38: The mean percentages of overall responses of the results of the positive items for HE and DSI in the second experiment

Response / Method	Strongly disagree	Probably disagree	Not sure	Probably agree	Strongly agree
HE	26%	31%	29%	12%	2%
DSI	0%	5%	7%	31%	57%

5.6.3 Comparative analysis to evaluate the adaptive framework

This section presents a comparative between and comprehensive analysis of three methods.

5.6.3.1 Determining the realism of usability problems

As mentioned in the literature review chapter, the discovered problems should be compared with UT problems, and falsification testing could be used if needed.

- Educational websites

For the first experiment, when the problems were compared between three methods, 6 problems were found to have been revealed by HE but not revealed by UT. Also, 55 problems were revealed by DSI but not revealed by UT. Furthermore, 22 problems were uniquely revealed by UT. Finally, 19 problems were found as overlapping between these methods. Subsequently, falsification testing was employed. Thus, the fixed UT tasks were designed (see Appendix O); these were derived from the unique problems of HE and DSI. The same procedures that were applied for UT were used. Five users were recruited for each website as this number of users can discover 80% of usability problems (Nielsen, 2000b). One independent evaluator was involved for mapping each of the unique predicted problems against the falsification testing result. The testing found that the 6 HE problems were confirmed by UT, which means that they are real problems. On the other hand, 42 DSI problems were confirmed by UT, which means that they too are real problems. Also, 13 problems were found by DSI but they are not confirmed by UT, which means that they are false positives. The majority of these problems were cosmetic and minor problems. When these results were discussed with the evaluators who had discovered these problems, they said that these seem to be problems for novice users who have low web-user experience. With regard to the unique usability problems found by UT, they were discussed with the

independent evaluators; they said that they were missed problems for DSI, and that they could be classified under the DSI categories; they were also miss problems for HE but some of them could not be classified to the HE heuristics (as will be discussed in section 5.6.3.4). Moreover, the unique problems of DSI and HE were miss problems for UT. The reason that UT missed these problems is that the expert evaluators visited pages that were not visited by the users.

Table 5.39 shows the performance of the three methods on a unique performance basis for the three websites. DSI was able to discover 6 catastrophic, 7 major, 18 minor and 11 cosmetic problems that were not revealed by the other methods. HE was not able to identify any catastrophic problems alone; however, it was able to identify 1 major, 2 minor and 3 cosmetic problems. UT was not able to discover any major problems; however, it discovered 1 catastrophic, 5 minor and 16 cosmetic problems. In comparing the results of HE and UT, this results are in line with (Virzi, 1992) who found that severe problems are more likely to be discovered by UT than HE. The main reason for this result may be because of the generality of the HE guidelines. Therefore, these results confirm that both methods (UT and HE) should be used together in any evaluation as recommended by Nielsen (1992a). Also, the DSI results are in line with (Chattratichart and Lindgaard, 2008) and (Chen and Macredie, 2005) who claimed that developing a set of detailed guidelines will help expert evaluators overcome this inherent flaw within HE. Overall, the three methods were able to discover 8 catastrophic, 13 major problems, 33 minor and 35 cosmetic problems.

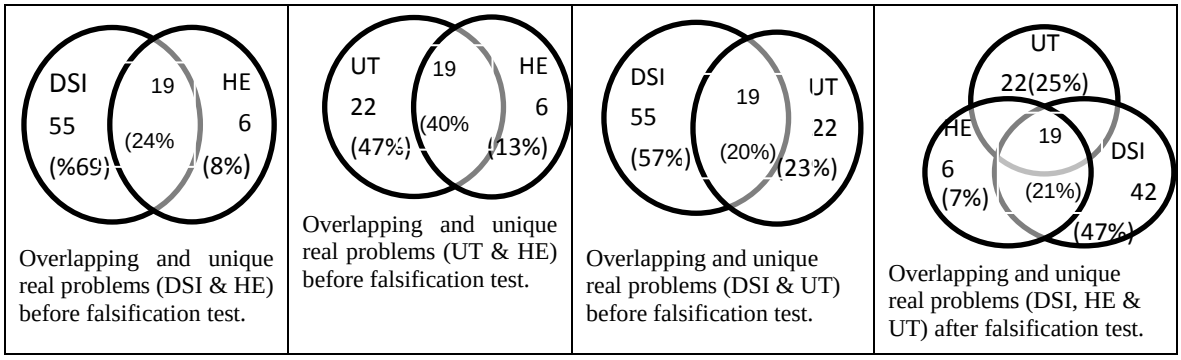
Table 5.39: Each method's performance with severity rating in the first experiment

Problem Types	HE (unique)	DSI (unique)	UT (unique)	DSI & HE&UT (overlapping)	Total number of problems in three websites (unique)
Catastrophic	0	6	1	1	8
Major	1	7	0	5	13
Minor	2	18	5	8	33
Cosmetic	3	11	16	5	35
Total	6	42	22	19	89

Figure 5.3 also shows the overlapping usability problems discovered by the three methods before and after the falsification test. 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 they also appeared to work fruitfully for the expert evaluators, who then revealed more real problems, both unique and

overlapping. For example, DSI found 47% uniquely of the total number of real usability problems (n = 42 out of 89). HE found only 7% uniquely of the total number of real usability problems (n = 6 out of 89), and UT identified 25% uniquely of the total number of real usability problems (n = 22 out of 89). 19 (21%) real problems out of 89 were discovered to be overlapping by the three methods. The clear outperformance of DSI was due in large part to incorporating user inputs whilst drawing up the method in one of the steps of the proposed framework, and due to the appropriacy of the DSI method to the characteristics of the educational domain as found in evaluators comments.

Figure 5.3: Each method's performance, uniquely and working in pairs in the first experiment



Overlapping and unique real problems (DSI & HE) before falsification test.

Overlapping and unique real problems (UT & HE) before falsification test.

Overlapping and unique real problems (DSI & UT) before falsification test.

Overlapping and unique real problems (DSI, HE & UT) after falsification test.

- Social networks websites

For the second experiment, when the problems were compared between three methods, 32 problems were found to have been revealed by HE but not revealed by UT. Also, 96 problems were revealed by DSI but not revealed by UT. Furthermore, 56 problems were uniquely revealed by UT. Finally, 23 problems were found as overlapping between these methods. Subsequently, falsification testing was employed. Thus, the fixed UT tasks were designed (see Appendix T); these were derived from the unique problems of HE and DSI. The same procedures that were applied for UT were used. Five users were recruited for each website. One independent evaluator was involved for mapping each of the unique predicted problems against the falsification testing result. The testing found that the 24 HE problems were confirmed by UT, which means that they are real problems. Also, 8 problems were found by HE but they are not confirmed by UT. On the other hand, 93 DSI problems were confirmed by UT, which means that they too are real problems. Also, 3 problems were found by DSI

but they are not confirmed by UT, which means that they are false positives. The majority of these problems were cosmetic problems in HE and DSI. When these results were discussed with the evaluators who had discovered these problems, they said that these seem to be problems for inexperienced users who have low web-user experience. With regard to the unique usability problems found by UT, these problems were discussed with the independent evaluators, and they said that they were miss problems for DSI, and that these problems could be classified under the DSI categories; these problems were also miss problems for HE but some of them could not be classified to the HE heuristics (as will be discussed in Section 5.6.3.4). Moreover, the unique problems of DSI and HE were miss problems for UT.

Table 5.40 shows the performance of the three methods on a unique performance basis for the three websites. DSI was able to discover 6 catastrophic, 24 major, 34 minor and 29 cosmetic problems that were not revealed by the other methods. HE was able to identify 1 catastrophic problem, 4 major, 11 minor and 8 cosmetic problems. UT was able to discover 7 catastrophic, 11 major problems, 17 minor and 21 cosmetic problems. In comparing the results of HE and UT, these results are in line with (Virzi, 1992) who found that severe problems are more likely to be discovered by UT than HE. Therefore, these results confirm that both methods (UT and HE) should be used together in any evaluation, as recommended by Nielsen (1992a). Also, the DSI results are in line with Chattratichart and Lindgaard (2008) and with Chen and Macredie (2005) who claimed that developing a set of detailed guidelines will help expert evaluators overcome this inherent flaw within HE. Overall, the three methods were able to discover 14 catastrophic, 43 major problems, 79 minor and 68 cosmetic problems.

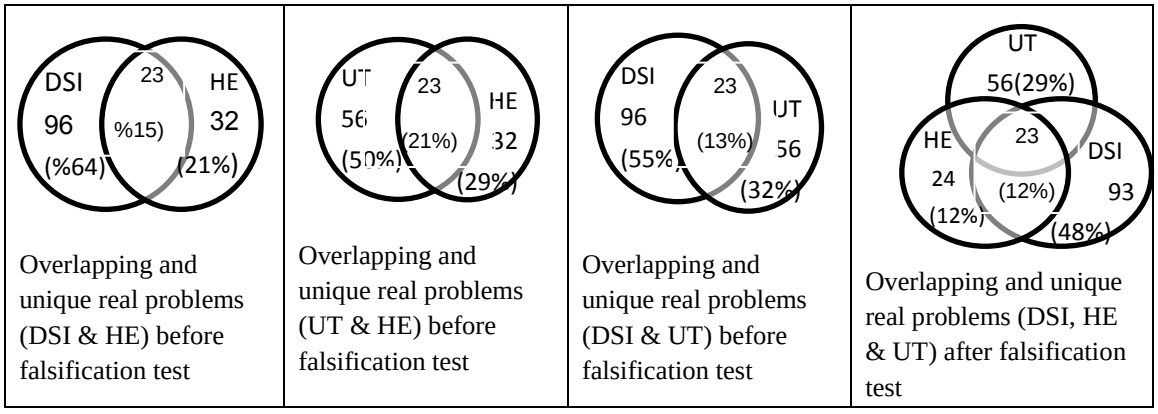
Table 5.40: Each method's performance with severity rating in the second experiment

Problem Types	HE (unique)	DSI (unique)	UT (unique)	DSI & HE & UT (overlapping)	Total number of problems in three websites (unique)
Catastrophic	1	6	7	0	14
Major	4	24	11	4	43
Minor	11	34	17	10	72
Cosmetic	8	29	21	9	67
Total	24	93	56	23	196

Figure 5.4 also shows the overlapping usability problems discovered by the three methods before and after the falsification test. 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 they also appeared to work fruitfully for the expert evaluators, who then revealed more real problems, both unique and overlapping. For example, DSI found 48% uniquely of the total number of real usability problems (n = 93 out of 196). HE found only 12% uniquely of the total number of real usability problems (n = 24 out of 196), and UT identified 29% uniquely of the total number of real usability problems (n = 56 out of 196). 23 (12%) real problems out of 196 were discovered to be overlapping by the three methods. The clear outperformance on the part of DSI was due in large part to incorporating user inputs whilst drawing up the method in one of the steps of the proposed framework, and due to the appropriacy of the DSI method to the characteristics of the social network domain, as found in evaluators comments.

Figure 5.4: Each method's performance, uniquely and working in pairs in the second experiment



Overlapping and unique real problems (DSI & HE) before falsification test

Overlapping and unique real problems (UT & HE) before falsification test

Overlapping and unique real problems (DSI & UT) before falsification test

Overlapping and unique real problems (DSI, HE & UT) after falsification test

5.6.3.2 Comparison of UT's and HE's performance to published researches

The above results were compared to previously published researches, as shown in Table 5.41, and the results reveal that the discovered usability problems range from 40% to 68% and 15% to 61% for HE and UT, respectively. However, in the first experiment, the discovered usability problems range from 14% to 31% and 29% to 39% for HE and UT, respectively. Also, in the second experiment, the discovered usability problems range from 29% to 36% and 24% to 43% for HE and UT, respectively. The differences between these studies are related to the number of evaluators and users used as well as their types.

Table 5.41: Comparison of UT's and HE's performance

Method		HE	UT	Comments
(Jeffries et al., 1991)		51%	15%	3 evaluators, and 6 users involved
(Doubleday et al., 1997)		40%	39%	Did not report user numbers or evaluator number
(Law and Hvannberg, 2002)		68%	61%	2 evaluators, 10 users, and 20 tasks
First experiment	Skoool	23%	32%	4 evaluators for each website, 20 users for each website, 4 tasks for each website.
	AcademicEarth	31%	29%	
	BBC KS3bitesize	14%	39%	
Second experiment	Google+	22 (31%)	34 (43%)	3 evaluators for each website, 25 users for each website, 6 tasks for each website
	LinkedIn	21 (31%)	26 (33%)	
	Ecademy	12 (32%)	19 (24%)	

In addition, many researchers recommend conducting UT together with HE because they have found that each method discovers unique problems (Nielsen, 1992a); (Law and Hvannberg, 2002), so when they are conducted together, they can reveal all the problems in the targeted website. Again, this experiment may confirm or deny this recommendation, depending on the above results. It can also be seen that combining the results of HE with UT offers quite good results in terms of cosmetic problems, whereas combining UT results with HE offers better results in terms of major, minor and cosmetic problems. Thus, the results of the 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, as argued by (Nielsen, 1992a). On the other hand, combining the UT results with DSI offers better results in terms of cosmetic problems. Thus, DSI, as created from the proposed adaptive framework, refutes that recommendation.

5.6.3.3 Usability problem report

- Educational websites

For the first experiment, the recommendation report was prepared and sent to the website owners. It includes the usability problems that were found using the three different methods (HE, DSI and UT), as in Appendix N. This report aims to share these problems with the website owners to help them to improve the usability of their websites. Two independent evaluators were recruited and they worked hard to compare these problems and to identify the overlapping and unique problems with their severity ratings. After sending the usability

problem report to the owners of the websites, these websites were checked whilst writing this section. Satisfyingly, all of these websites have been changed and their usability has been improved. One of the independent evaluators was asked to check the usability report against this improvement to find any remaining usability problems that have not yet been fixed. In the Skoool website, the number of the remaining problems is 18 (out of 35 problems). These problems are number 2, 3, 5, 8, 9, 10, 26, 29, 32, 33, 36, 37, 68, 69, 70, 71, 72, 73 and 75. Some of these problems have been fixed in some pages but ignored in others. For example, the 'Contact us' link is stuck at the bottom of the page of the Egypt site, whereas it is not in the Yemen site (problem number 8). Also, the website has been made compatible with mobile devices, except the videos (which are not working). In the AcademicEarth website, the design has changed and its usability has been improved. Thus, the number of remaining problems is 7 (out of 29 problems). These problems are number 19, 43, 45, 48, 51, 56 and 57. For example, the website has removed the registration and login process, and so it becomes (without these features) the same as the Skoool website. Also, the website uses YouTube for showing all its videos except some video electives. In the BBC KS3bitesize, the website has changed and they advertise these changes on the homepage by stating "Bitesize has changed!". Thus, the number of remaining problems is 9 (out of 25 problems). These problems are number 25, 58, 59, 60, 63, 67, 87, 88 and 89.

- Social networks websites

For the second experiment, the recommendation report was prepared and sent to the website owners. It includes the usability problems that were found using the three different methods (HE, DSI and UT), as in Appendix S. Two independent evaluators were recruited and they worked hard to compare these problems and to identify the overlapping and unique problems with their severity ratings. These websites were checked whilst writing this section. Happily, all of these websites have been changed and their usability has been improved. One of the independent evaluators was asked to check the usability report against this improvement to find any remaining usability problems that have not yet been fixed. In the Google+ website, the number of remaining problems is 22 (out of 85 problems). These problems are number 1, 3, 4, 5, 29, 49, 56, 67, 70, 106, 107, 108, 110, 114, 115, 117, 123, 124, 125, 128, 137 and 139. Some of these problems have been fixed in some pages but ignored in others. In the LinkedIn website, the design has been changed and its usability has been improved. Thus,

the number of remaining problems is 19 (out of 96 problems). These problems are number 38, 79, 84, 87, 141, 144, 146, 147, 159, 150, 154, 155, 157, 158, 161, 163, 164, 166 and 169. In the Ecademy, the website has changed. Thus, the number of remaining problems is 15 (out of 46 problems). These problems are number 45, 93, 94, 100, 103, 171, 176, 180, 1 83, 184, 185, 186, 191,194 and 196.

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

- Educational websites

For the first experiment, the qualitative assessment of the above results for UT was compared with the two other methods (DSI and HE), in particular, in terms of the areas of usability problems found in this experiment. These areas assisted in identifying how each method (DSI and HE) performed in each usability problem area or category of heuristics. This can be done by matching the problems discovered by UT to their categories in the HE heuristics or the DSI checklist, as recommended by (Nielsen, 1995a). This can help to identify how each method (DSI and HE) performs in each category of guidelines. In other word, this can help to identify how UT performs in each category of DSI and in each heuristic of HE. Also, it can be used to compare the results found here with current published work. 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 5.42 and 5.43 illustrate. The overall results from both tables show that 11 problems out of 16 in the BBC KS3bitesize were classified into HE. In the Skoool website, 12 UT problems out of 13 were classified into HE. Also, 11 problems out of 13 in the AcademicEarth were classified into HE. On the other hand, all the UT problems were successfully classified into DSI. This proves that HE is rather general, and is unlikely to encompass all user problems (Thovtrup and Nielsen, 1991); (Henninger, 2000), such as usability problems in ‘Learning process’ and ‘Motivational factors’. Also, this proves that DSI was indeed able to discover users’ problems (19), and that the unique problems discovered by UT (22) were miss problems for DSI.

In terms of Table 5.42 for HE, the tasks given to the users during the usability testing seem to have walked them through ‘Visibility of system status’ and ‘Help and documentation’ and this could have increased the opportunity to discover problems. Few problems were revealed in ‘Match between the system and the real world’, ‘Consistency and standards’, ‘Flexibility

and efficiency of use', 'Aesthetic and minimalist design', and 'User control and freedom'. This is in contrast to the results, proving that 'User control and freedom' is a common weakness in dynamic websites, in particular e-shops (Chen and Macredie, 2005). On the other hand, Table 5.43 for DSI shows that the tasks given to the users during the usability testing seem to have walked them through the quality of learning process and motivation factor which could have increased the opportunity to discover problems. Furthermore, the findings confirm that 'Visibility', 'Help', 'Functionality', 'Content quality', and 'Interface design and media' are common weaknesses in dynamic websites (particularly for educational ones). This finding is in line with Hasan and Abuelrub (2013). In conclusion, UT worked better than HE because seven problems were not classified in it. However, all the users' problems were classified in DSI.

Table 5.42: Usability problems found by UT compared with HE in the first experiment

Heuristic Evaluation	Skoool	AcademicEarth	BBC KS3bitesize
Visibility of system status	7	5	6
Match between the system and the real world	1	1	1
User control and freedom	0	1	0
Consistency and standards	1	1	0
Error prevention	0	0	0
Recognition rather than recall	0	0	2
Flexibility and efficiency of use	1	0	1
Aesthetic and minimalist design	0	2	0
Helps users recognize, diagnose and recover from errors	0	0	0
Help and documentation	2	1	1
Total problems	12	11	11
Number of unclassified problems	1	2	5

Table 5.43: Usability problems found by UT compared to the DSI in the first experiment

Usability problem area	Skoool	AcademicEarth	BBC KS3bitesize
User usability	7	8	8
Motivational factors	0	2	0
Content information and process orientation	1	1	2
Learning process	1	0	0
Design and media usability	5	2	4
Total problems	13	13	16

- Social networks websites

For the second experiment, the overall results from Tables 5.44 and 5.45 show that 30 problems out of 34 in the Google+ website were classified into HE. In the LinkedIn website, 19 UT problems out of 26 were classified into HE. Also, 12 problems out of 19 in the Ecademy were classified into HE. On the other hand, all the UT problems were successfully classified into DSI. This proves that HE is rather general, and is unlikely to encompass all user problems (Thovtrup and Nielsen, 1991); (Henninger, 2000), such as usability problems in the 'User usability, sociability and management activities', 'Business support', and 'Security and privacy' areas. Also, this proves that DSI was indeed able to discover users' problems (23), and that the unique problems discovered by UT (56) were miss problems for DSI.

In terms of Table 5.44 for HE, the results reveal that 'Visibility of system status', 'Match between the system and the real world', 'Aesthetic and minimalist design', and 'Helps users recognize, diagnose and recover from errors' are common weaknesses in dynamic websites (particularly for social network websites) from the perspective of HE. This is in contrast to the results, proving that 'User control and freedom' is a common weakness in dynamic websites (Chen and Macredie, 2005). On the other hand, Table 5.45 for DSI shows that the tasks given to the users during the usability testing seem to have walked them through the business and management activities, which could have increased the opportunity to discover problems. It revealed that 'Layout and formatting', 'Content quality', 'User usability, sociability and management activities', 'Navigation system and search quality' are common weaknesses in dynamic websites (particularly for social network websites) from the perspective of DSI. This finding is in line with (Lee and Kozar, 2012) and (Hart et al., 2008). In conclusion, UT worked better than HE because 18 problems were not classified in it. However, all the users' problems were classified in DSI.

Table 5.44: Usability problems found by UT compared with HE in the second experiment

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
Unclassified problems	4	7	7

Table 5.45: Usability problems found by UT compared to DSI in the second experiment

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

5.6.3.5 Performance of the three methods (UT, HE, DSI)

5.6.3.5.1 Number of usability problems

- Educational websites

For the first experiment, Tables 5.46, 5.47 and 5.48 show how UT, HE and DSI revealed different types and numbers of usability problems. The main statistical question relates to the extent to which there are statistical differences among the methods' performance in terms of problems found. Before performing this test, the normality test was conducted to select the correct test. Shapiro-Wilk test was performed, and its p-value is significant ($p<0.05$) when $p= 0.000$, then the data is not normal distributed. Therefore, Kruskal Wallis test was chosen at a significant level of 5% in order to examine whether there are significant differences amongst the method's groups in terms of usability problems found. Table 5.49 shows that there is a strong likelihood for a significant difference between the results of the three

methods in each website, where $p < 0.05$. This result can confirm that each method influences the results achieved where the method's performance results differ statistically. The UT, HE and DSI methods revealed 64%, 8% and 48% of the usability problems found in the BBC KS3bitesize website, respectively. In the Skoool website, UT, HE and DSI revealed 37%, 29% and 57% of the found usability problems, respectively. Finally, UT, HE and DSI revealed 45%, 45% and 100% of the found usability problems in the AcademicEarth, respectively. The performance of HE in discovering usability problems during the experiment ranged from 8% to 45%. UT discovered usability problems ranging from 37% to 64%, while DSI discovered usability problems ranging from 48% to 100%. Also, UT and HE performed better in discovering a few major, minor and cosmetic real usability problems, but DSI was better in discovering more catastrophic, major, minor and cosmetic real usability problems. Thus, it can be seen that DSI was better in discovering real problems; this was followed by UT, and then finally HE.

Table 5.46: Findings in BBC KS3bitesize

Method / Problem type	UT	HE	DSI	Total problems (no duplicates)
Catastrophic	1 (6%)	0 (0%)	2 (17%)	3
Major	1 (6%)	0 (0%)	0 (0%)	1
Minor	5 (31%)	2 (22%)	10 (83%)	12
Cosmetic	9 (56%)	0 (0%)	0 (0%)	9
No. of problems	16 (64%)	2 (8%)	12 (48%)	25

Table 5.47: Findings in Skoool

Method / Problem type	UT	HE	DSI	Total problems (no duplicates)
Catastrophic	1 (7%)	1 (10%)	3 (15%)	3
Major	0 (0%)	1 (10%)	3 (15%)	4
Minor	5 (39%)	5 (50%)	7 (35%)	11
Cosmetic	7 (54%)	3 (30%)	7 (35%)	17
No. of problems	13 (37%)	10 (29%)	20 (57%)	35

Table 5.48: Findings in Academic Earth

Method / Problem type	UT	HE	DSI	Total problems (no duplicates)
Catastrophic	0 (0%)	0 (0%)	2 (7%)	2
Major	3 (23%)	3 (23 %)	7 (24%)	7
Minor	5 (38%)	5 (38 %)	11 (38%)	11
Cosmetic	5 (38%)	5 (38%)	9 (31%)	9
No. of problems	13 (45%)	13 (45%)	29 (100%)	29

Table 5.49: Kruskal Wallis result for examining the differences amongst the groups in terms of usability problems found

	Chi-Square	df	Sig. value
Methods	28.906	2	0.000

- Social networks websites

For the second experiment, Tables 5.50, 5.51 and 5.52 show how UT, HE and DSI revealed different types and numbers of usability problems. Are there statistical differences among the methods' performance in terms of problems found? To answer this question, the normality test was conducted to select the correct test. Shapiro-Wilk test was performed, and its p-value is significant ($p<0.05$) when $p= 0.000$, then the data is not normal distributed. Therefore, Kruskal Wallis test was again chosen at a significant level of 5% in order to examine whether there are significant differences amongst the method's groups in terms of usability problems found. Table 5.53 shows that there is a strong likelihood for a significant difference between the results of the three methods, where $p < 0.05$. This result can confirm that each method influences the results achieved where the method's performance results differ statistically. The UT, HE and DSI methods revealed 61%, 34% and 84% of the usability problems found in the Google+ website, respectively. In the LinkedIn website, UT, HE and DSI revealed 42%, 29% and 71% of the found usability problems, respectively. Finally, UT, HE and DSI revealed 40%, 21% and 48% of the found usability problems in Ecademy, respectively. The performance of HE in discovering usability problems during the experiment ranged from 21% to 34%. UT discovered usability problems ranging from 40% to 61%, while DSI discovered usability problems ranging from 48% to 84%. Also, HE performed better in discovering sufficient major, minor and cosmetic real usability problems, but DSI and UT were good together in discovering more catastrophic, major, minor and cosmetic real usability problems. Thus, it can be seen that DSI was good in discovering real problems; this was followed by UT, and then finally HE.

Table 5.50: Findings in Google+

Method / Problem type	UT	HE	DSI	Total problems in the site from three methods (no duplicate)
Catastrophic	4 (100%)	0 (0%)	0 (0%)	4
Major	9 (43%)	3 (14%)	11 (52%)	21
Minor	11 (37%)	10 (33%)	21 (70%)	29
Cosmetic	10 (46%)	6 (27%)	15 (68%)	22
No. of problems	34 (61%)	19 (34%)	47 (84%)	56

Table 5.51: Findings in LinkedIn

Method / Problem type	UT	HE	DSI	Total problems in the site from three methods (no duplicate)
Catastrophic	2 (67%)	0 (0%)	3 (100%)	3
Major	5 (34%)	5 (34%)	9 (60%)	15
Minor	8 (27%)	8 (27%)	21 (70%)	30
Cosmetic	11 (79%)	5 (36%)	11(79%)	14
No. of problems	26 (42%)	18 (29%)	44 (71%)	62

Table 5.52: Findings in Ecademy

Method / Problem type	UT	HE	DSI	Total problems in the site from three methods (no duplicate)
Catastrophic	0 (0%)	0 (0%)	0 (0%)	0
Major	3 (50%)	0 (0%)	6 (100%)	6
Minor	6 (50%)	8 (67%)	7 (58%)	12
Cosmetic	11 (37%)	2 (7%)	10 (33%)	30
No. of problems	19 (40%)	10 (21%)	23 (48%)	48

Table 5.53: Kruskal Wallis result for examining the differences amongst the groups in terms of usability problems found

	Chi-Square	df	Sig. value
Methods	18.146	2	0.000

5.6.3.5.2 Usability problem areas

In terms of the usability problem areas that were identified from Step 3 from the adaptive framework for both experiments;

- Educational websites

For the first experiment, it can be seen in Table 5.54 that DSI helped to identify large numbers of real usability problems in all usability areas on the three websites (61). However, HE overall worked slightly better in discovering 25 real usability problems relating to three usability problem areas. However, it failed in exposing any usability problems in two main usability problems areas, which are 'Motivational factors' and 'Learning process', and it failed to identify a sufficient number of usability problems in the 'Content information and process orientation' area. Furthermore, UT worked better in discovering usability problems (41) in three usability areas, but it failed to identify a sufficient number of usability problems in 'Motivational factors' and 'Learning process'.

Table 5.54: Number of usability problem areas identified by the three methods in the first experiment

Usability Problem Areas	UT	DSI	HE
User usability	23 problems	15 problems	15 problems
Motivational factors	2 problems	3 problems	-
Content information and process orientation	4 problems	5 problems	2 problems
Learning process	2 problem	6 problems	-
Design and media usability	10 problems	32 problems	8 problems
Total number of problems	41	61	25

- Social networks websites

The second experiment, it can be seen in Table 5.55 that DSI helped to identify large numbers of real usability problems in all usability areas on the three websites (116). However, HE overall worked slightly better in discovering 47 real usability problems relating to four usability problem areas. However, it failed in exposing any usability problems in three main usability problems areas, which are 'Security and privacy' and 'Business support', and 'Accessibility and compatibility'. Furthermore, UT worked better in discovering usability problems (79 in all usability areas), but it failed to identify more usability problems in the 'Accessibility and compatibility' area

Table 5.55: Number of usability problem areas identified by the three methods in the second experiment

Usability Problem Areas	UT	DSI	HE
Layout and formatting	13	15	9
Content quality	12	25	4
Security and privacy	4	8	-
Business support	8	4	-
User usability, sociability and management activities	19	40	19
Accessibility and compatibility	2	4	-
Navigation system and search quality	21	20	15
Total number of problems	79	116	47

5.6.3.5.3 Time spent

- Educational websites

For the first experiment, Table 5.56 shows the time spent by each method; UT scored the highest time spent, more than DSI and HE, with 98.60, 42.58, and 24.25 minutes,

respectively. The main statistical question is: to what extent are there statistical differences among the methods' performance in terms of time spent? Before performing this test, the normality test was conducted to select the correct test. Shapiro-Wilk test was performed, and its p-value is significant ($p<0.05$) when $p=0.000$, then the data is not normal distributed. Therefore, Kruskal Wallis test was again chosen at a significant level of 5% in order to examine whether there are significant differences amongst the method's groups in terms of time spent. Kruskal Wallis revealed that there is a strong likelihood for a significant difference between the three methods in terms of time spent on discovering usability problems, where $p < 0.001$, as shown in Table 5.56. The other statistical question is: to what extent is there a relationship between time spent and problems found? The Pearson correlation test was used, and it reveals that there is a positive correlation between time spent and problems found in this experiment, where the p-value is less than 0.05 ($p = 0.000$), as shown in Table 5.56.

Table 5.56: Time spent by each method in the first experiment

Website Method	Skoool	AcademicEarth	BBC KS3bitesize	Mean
DSI	41	45	42	42.58
HE	25	21	27	24.25
UT	112	88	96	98.66
Kruskal Wallis	Chi-Square = 53.206, p = 0.000			
Correlations	Pearson Correlation = 0.604, P= 0.000			

- Social networks websites

For the second experiment, Table 5.57 shows that UT scored the highest time spent, more than DSI and HE, with 392.6, 72.3, and 55.6 minutes, respectively. Are there statistical differences among the methods' performance in terms of time spent? The normality test was conducted to select the correct test. Shapiro-Wilk test was performed, and its p-value is significant ($p<0.05$) when $p= 0.000$, then the data is not normal distributed. Therefore, Kruskal Wallis test was again chosen at a significant level of 5% in order to examine whether there are significant differences amongst the method's groups in terms of time spent. Kruskal Wallis reveals that there is a strong likelihood for a significant difference between the three methods in terms of time spent on discovering usability problems, where $p < 0.05$, as shown in Table 5.57. The other statistical question is: to what extent is there a relationship between time spent and problems found? The Pearson correlation test was used, and it reveals that

there is a positive correlation between time spent and problems found in this experiment, where the p-value is less than 0.05 ($p < 0.05$), as shown in Table 5.57.

Table 5.57: Time spent by each method in the second experiment

Website / Method	Google+	LinkedIn	Ecademy	Mean
DSI	73	72	72	72.3
HE	50	60	57	55.6
UT	429	377	372	392.6
Kruskal Wallis	Chi-Square = 43.442, p = 0.000			
Correlations	Pearson Correlation = 0.432, P= 0.000			

5.6.3.5.4 Usability evaluation method (UEM) performance Metrics

After applying the performance metrics that were mentioned in section 2.4.4, it can be seen that;

- Educational websites

For the first experiment, Table 5.58 that DSI is more efficient, thorough and effective than the other two methods in terms of identifying the total number of real problems relative to total time spent, and in terms of its ability to identify real usability problems relating to the user interface. UT is the second good method, but it is more reliable than DSI. HE has the worst result in identifying real problems; however, it is the cheapest to use, as shown in Table 5.59. Moreover, DSI is slightly more expensive than HE, but is cheaper than UT.

- Social networks websites

For the second experiment, Table 5.58 that DSI is more efficient than HE and slightly more efficient than UT. Also, DSI is more thorough, effective and valid than HE in terms of identifying the total number of real problems relative to total time spent, and in terms of its ability to identify real usability problems relating to the user interface. UT is the second good method in identifying usability problems, and it is more reliable than HE and slightly more reliable than DSI. HE has the worst result in identifying real problems; however, it is the cheapest to use, as shown in Table 5.59. Moreover, DSI is slightly more expensive than HE, but is cheaper than UT.

Table 5.58: Comparing the metrics between the three methods

Metric / Method		Efficiency	Thoroughness	Validity	Effectiveness	Cost
First experiment	HE	0.4	0.3	1	0.3	$ 1,017
	DSI	0.6	0.7	0.8	0.6	$ 1,206
	UT	0.4	0.5	1	0.5	$ 3420
Second experiment	HE	0.84	0.12	0.7	0.1	$706.66
	DSI	1.7	0.5	0.9	0.5	$863,33
	UT	1.67	0.4	1	0.4	$4275

Table 5.59: Costs for employing the three methods

	Method	Skoool	AcademicEarth	BBC KS3bitesize	Mean cost
The first experiment	Heuristic evaluation (HE)	$ 1,020 This includes the time spent by 4 evaluators (1.6 hours), 2.6 hours collecting data from the evaluation sessions, 6 hours analysing data.	$ 990 This includes the time spent by 4 evaluators (1.3 hours), 2.6 hours collecting data from the evaluation sessions, 6 hours analysing data.	$ 1,040 This includes the time spent by 4 evaluators (1.8 hours),2.6 hours collecting data from the evaluation sessions ,6 hours analysing data	$ 1,017
	Domain Specific Inspection (DSI)	$ 1,130 This includes the time spent by 4 evaluators (2.7 hours), 3 hours collecting data from the evaluation sessions, 6.6 hours analysing data.	$ 1,250 This includes the time spent by 4 evaluators (2.9 hours), 3 hours collecting data from the evaluation sessions, 6.6 hours analysing data.	$ 1,240 This includes the time spent by 4 evaluators (2.8 hours), 3 hours collecting data from the evaluation sessions, 6.6 hours analysing data.	$ 1,206
	Method	Google+	LinkedIn	Ecademy	Mean cost
	User Testing (UT)	$ 3420 This includes the time spent by 20 users (1.9 hours), 10 hours collecting data from the evaluation sessions, 5 hours analysing data.	$ 3420 This includes the time spent by 20 users (1.5 hours), 10 hours collecting data from the evaluation sessions, 5 hours analysing data.	$ 3420 This includes the time spent by 20 users (1.6 hours), 10 hours collecting data from the evaluation sessions, 5 hours analysing data.	$ 3420
The second	Heuristic evaluation (HE)	$680 This includes the time spent by 3 evaluators (2.5 hours), 1 hour collecting data from the evaluation sessions, 3.3 hours analysing data.	$730 This includes the time spent by 3 evaluators (3 hours), 1 hour collecting data from the evaluation	$710 This includes the time spent by 3 evaluators (2.8 hours), 1 hour collecting data from the evaluation sessions, 3.3 hours analysing data.	$706.66

			sessions, 3.3 hours analysing data.		
	Domain Specific Inspection (DSI)	$870 This includes the time spent by 3 evaluators (3.6 hours), 1.3 hours collecting data from the evaluation sessions, 3.8 hours analysing data.	$860 This includes the time spent by 3 evaluators (3.5 hours), 1.3 hours collecting data from the evaluation sessions, 3.8 hours analysing data.	$860 This includes the time spent by 3 evaluators (3.5 hours), 1.3 hours collecting data from the evaluation sessions, 3.8 hours analysing data.	$863,33
	User Testing (UT)	$4275 This includes the time spent by 25 users (7.15 hours), 12 hours collecting data from the evaluation sessions, 6.30 hours analysing data.	$4275 This includes the time spent by 25 users (6.28 hours), 12 hours collecting data from the evaluation sessions, 6.30 hours analysing data.	$4275 This includes the time spent by 25 users (6.2 hours), 12 hours collecting data from the evaluation sessions, 6.30 hours analysing data	$4275

5.6.3.6 Sample size

One of the objectives of this research is to investigate the issue of sample size. Also, it is to examine further the impact of sample size on the findings of usability tests, thus to determine an appropriate sample size for HE and UT. Moreover, another objective is to quantify the sample size required for DSI. As mentioned in section 2.4.1, the numbers of users for UT and evaluators for HE is still a debatable point, and this has led to the establishment of many rules such as the 10±2 rule (Hwang and Salvendy, 2010) or 4±1 (Nielsen and Molich, 1990). The rules of 10±2 and 4±1 will be examined in the first experiment (educational domain). In the second experiment (social network domain), the rule of 3 evaluators (including double and single) and the rule of 20 users and 5 users (the 'magic number') will be examined.

- Educational websites

As previously mentioned in the methodology chapter, 20 users were recruited for UT for each website's group (total 60 users) and 4 evaluators for each method's group (total 8 evaluators). Therefore, the UT results for each group were divided into two teams. The first team consists of the results of 8 users, whereas the second team consists of the results of 12 users for the same group. This means that there are 6 teams overall, three teams consisting of 8 users, and the other three teams consisting of 12 users for each website. The users in

each team were chosen according to pick one after every five users. Each team was given the name A, B, C, D, E and F, as shown the Table 5.60.

Table 5.60: Sample size for UT according to the probability level of problem discovery in the first experiment

Website	Team name	# Users	# Issue found	Mean found	Problem discovery rate (P)	Percentage of problems discovered
Skoool	Team A	8	46	5.75	0.23	88%
	Team B	12	92	7.66	0.31	98.9%
AcademicEarth	Team C	8	39	4.88	0.18	80%
	Team D	12	91	7.58	0.37	99.7%
BBC KS3bitesize	Team E	8	50	6.25	0.27	92%
	Team F	12	88	7.33	0.34	99.4%

For UT, the figures in Table 5.60 show that, in the Skoool website, Team A reported 46 issues, whereas Team B reported 92 issues. In AcademicEarth, Team C reported 39 issues; however, Team D reported 91 issues. Finally, Team E in BBC KS3bitesize reported 50 issues, but Team F reported 88 issues. The maximum overlap was 122 issues; it occurred between Team E, which tested 8 users and reported 50 issues, and Team F, which tested 12 users and reported 88 issues. The minimum overlap was 114 issues; between Teams A and B. The whole study identified 41 critical issues, 2 catastrophic issue, 4 major issues, 15 minor issues and 20 cosmetic issues. Moreover, when the problems of both groups were classified according to the five problem areas in DSI for this domain, the group of 8 users was more efficient in discovering problems relating to two particular areas, which were ‘User usability’ and ‘Content information and process orientation’. However, the group of 12 users was more efficient in discovering problems relating to three areas, which were ‘Design and media usability’, ‘Learning process’ and ‘Motivational factors’. Furthermore, the problem discovery rate (p) was used (as mentioned in section 2.4.1) to determine the number of users needed in order to achieve satisfactory results. It has been calculated by Lewis (2006) and Turner et al. (2006), and they reported that the p-value usually ranges from 0.16 to 0.42. In this experiment the probability value (p) ranges from 0.18 to 0.37. Table 5.60 shows that 8 users can find percentages of problems ranging between 80% and 92%. 12 users can find percentages of problems ranging between 98.8% and 99.4%. Also, Sauro (2006) proposed an online system called the Sample Size Calculator to compute the sample size needed for discovering problems in a user interface. This calculator is based on a binomial probability formula and it uses the Good-Turing and Normalization procedure as outlined by Lewis

(2001). Thus, Table 5.61 was produced after determining the calculator (p) value (from Table 5.60, the p average = 0.28) and the targeted percentages (P). This calculator suggests that 14 users are needed to discover 99%, and from 5 to 6 users are needed to discover from 80% to 85% of the total usability problems. The result in this experiment is in line with Nielsen's claim that 5 users are enough to discover from 80% to 85% of usability problems (Turner et al., 2006, Nielsen, 2000b). Furthermore, the 10±2 rule provides optimal results, and this finding in line with (Hwang and Salvendy, 2010) and it proves their rule.

Table 5.61: User number and problems discovered (percentage) for UT in the first experiment

The targeted percentage	The required number of users
99%	14
95%	10
90%	7
85%	6
80%	5

In terms of the 4±1 rule, each evaluator group was given a name (G, H, I, J, K and L) and their results were analysed individually, as shown the Tables 5.62 and 5.63. For HE, the figures in Table 5.62 show that, Teams G and J reported 10 and 13 issues, respectively, whereas Team K reported 2 issues. The probability value (p) ranges from 0.29 to 0.37. Thus, the percentage of problems is ranging between 75% and 85%. The result of this experiment proves and is in line with previous studies that claim 4 evaluators can discover 80% of usability problems (Nielsen and Molich, 1990); (Turner et al., 2006); (Hwang and Salvendy, 2010). For DSI, the figures in Table 5.63 show that Teams H and I reported 33 and 29 issues, respectively, whereas Team L reported 12 issues. The probability value (p) ranges from 0.59 to 0.71. Thus, the percentage of problems is ranging between 98% and 99%. This result is in line with Henninger (2000, p.228), which is that "contextualized guidelines are better than abstract or decontextualized ones''.

Table 5.62: The performance of the sample size of HE evaluators according to the probability level of problem discovery in the first experiment

Website	Group	Expert and type	Method	# of problems found	# of problems without repetition	Problem discovery rate (P)	% of problems discovered
Skoool	G	Ev. 1+	HE	8	10	0.29	75%
		Ev. 2^	HE	5			
		Ev. 3^	HE	1			
		Ev. 4+	HE	5			
AcademicEarth	J	Ev. 1+	HE	6	13	0.37	85%
		Ev. 2^	HE	6			
		Ev. 3+	HE	7			
		Ev. 4^	HE	4			
BBC KS3bitesize	K	Ev. 1+	HE	1	2	0.35	83%
		Ev. 2^	HE	1			
		Ev. 3^	HE	1			
		Ev. 4+	HE	2			

(+) Double expert (^) Single expert (EV) Evaluator

Table 5.63: The performance of the sample size of DSI evaluators according to the probability level of problem discovery in the first experiment

Website	Group	Expert and type	Method	# of problems found	# of problems without repetition	Problem discovery rate (P)	% of problems discovered
Skoool	H	Ev. 1+	DSI	21	33	0.59	98%
		Ev. 2^	DSI	15			
		Ev. 3+	DSI	13			
		Ev. 4^	DSI	15			
AcademicEarth	I	Ev. 1+	DSI	11	29	0.64	98%
		Ev. 2^	DSI	10			
		Ev. 3^	DSI	9			
		Ev. 4+	DSI	12			
BBC KS3bitesize	L	Ev. 1+	DSI	6	12	0.71	99%
		Ev. 2^	DSI	6			
		Ev. 3+	DSI	7			
		Ev. 4^	DSI	5			

(+) Double expert (^) Single expert (EV) Evaluator

Moreover, the effects of the evaluators' characteristics (double or single) have been confirmed in this study. The single expert evaluators were more efficient in discovering usability problems relating to design navigation and layout, whereas the double expert evaluators were more efficient in discovering usability problems relating to content quality, learning process and motivational factors; this is because they have expertise in this domain. In fact, the double evaluators discovered more problems than the single evaluators in each

method and overall. This in line with Nielsen (1992a) when he stated, "usability specialists with expertise in the specific kind of interface being evaluated did much better than regular usability specialists without such expertise, especially with regard to certain usability problems that were unique to that kind of interface". Furthermore, Table 5.64 shows that the double evaluators discovered more problems by using two methods. It can be seen that four double evaluators found proportions of usability problems of between 42% and 72%. The result of this experiment is not dissimilar to the results of previous studies that have found double evaluators discovering between 74% and 87% of usability problems (Nielsen and Molich, 1990); (Nielsen, 1992a). Moreover, it can be seen that four single evaluators found proportions of usability problems of between 26% and 66%. This result also is in line with a previous study that found five single evaluators finding 51% of the known usability problems (Nielsen, 1992a). It was concluded from Table 5.64 that the results of the single experts, who used DSI method, provided results approaching or outperforming effectiveness of the double experts, who used HE. This mean the DSI improves the evaluator's performance.

Table 5.64: Results of effecting double evaluators in the first experiment

# Evaluator		Method type	# problems found	Total average proportion
Ed. Domain	4 double	HE	29	42%
	4 single	HE	18	26%
	4 double	DSI	70	72%
	4 single	DSI	60	66%

For more examination this issue, the Sample Size Calculator was used for HE and DSI methods. Table 5.65 was produced after determining the calculator (p) value (from Tables 5.62 and 5.63, the p average = 0.33 for HE and 0.64 for DSI) and the targeted percentages (P). This calculator suggests that 11 evaluators are needed to discover 99% for HE but only 4 evaluators for DSI, and from 4 to 5 evaluators for HE and 1 to 2 evaluators for DSI are needed to discover from 80% to 85% of the total usability problems. This result agrees and is in line with Nielsen's claim that 4±1 users are enough to discover from 80% to 85% of all usability problems (Nielsen and Molich, 1990); Nielsen, 2000; (Turner et al., 2006, Nielsen, 2000b); (Hwang and Salvendy, 2010). In conclusion, the effects of the evaluators' characteristics (double or single) have been confirmed in this study. This result is in line with a previous study when Nielsen (1992a) stated that the "effectiveness of HE can be substantially improved by having usability specialists as evaluators". Also, the effect of using

different methods by evaluators (i.e. HE and DSI) has been confirmed in this study, and this result is in line with a previous study (Hertzum and Jacobsen, 2001) .

Table 5.65: The required number of evaluators for both methods based on Sample Size Calculator in the first experiment

The targeted percentage	The required number of evaluators for HE	The required number of evaluators for DSI
99%	11	4
95%	7	3
90%	6	2
85%	5	2
80%	4	1

- Social networks websites

In the second experiment, 25 users were recruited for UT for each website's group (total 75 users) and 3 evaluators for each method's group (total 6 evaluators). Consequently, the UT results for each group were divided into two teams. The first team consists of the results of 20 users, whereas the second team consists of the results of 5 users for the same group. This means that there are 6 teams overall, three teams consisting of 20 users, and the other three teams consisting of 5 users for each website. The users in each team were chosen according to pick one after every five users. Each team was given the name M, N, O, P, Q and R, as shown the Table 5.66. To investigate the 3 evaluators rule, each evaluator group was given a name (S, T, U, V, W and X) and their results were analysed individually, as shown the Tables 5.68 and 5.79.

Table 5.66: Sample size for UT according to the probability level of problem discovery in the second experiment

Website	Team name	# Users	# Issue found	Mean found	Problem discovery rate (P)	Percentage of problems discovered
Google+	Team M	5	43	7	0.03	15%
	Team N	20	198	9.9	0.1	88%
LinkedIn	Team O	5	48	9.6	0.06	26.7%
	Team P	20	159	7.95	0.21	99.2%
Ecademy	Team Q	5	35	8.6	0.09	37.6%
	Team R	20	123	6.5	0.18	98.2%

For UT, the figures in Table 5.66 show that, in the Google+ website, Team M reported 43 issues, whereas Team N reported 198 issues. In LinkedIn, Team O reported 48 issues; however, Team P reported 159 issues. Finally, Team Q in Ecademy reported 35 issues, but

Team R reported 123 issues. The highest overlap was 207 issues; it occurred between Team M, which tested 5 users and reported 43 issues, and Team N, which tested 20 users and reported 198 issues. The lowest overlap was 139 issues, between Teams Q and R. The whole study identified 79 critical issues, 6 catastrophic issues, 17 major issues, 25 minor issues and 32 cosmetic issues. Furthermore, Nielsen (2000b) proclaims that 5 users are enough to catch 80% of the problems on practically any website. However, our data in this experiment provide evidence to the contrary. When analysing samples of 5 users, in the best cases only 37.6% of the total problems are found, i.e., not near the 80% objective. Moreover, when the problems of both groups were classified according to the 7 problem areas in this domain, the group of 5 users were more efficient in discovering problems relating to three areas, which were 'Layout and formatting', 'Content quality' and 'Accessibility and compatibility'. However, the group of 20 users were more efficient in discovering problems relating to four areas, which were 'User usability, sociability and management activities', 'Navigation system and search quality', 'Security and privacy' and 'Business support'. Furthermore, the problem discovery rate (p) was used again. The probability value (p) ranges from 0.03 to 0.21. Table 5.66 shows that 5 users can find percentages of problems ranging between 15% and 37.6%. 20 users can find percentages of problems ranging between 88% and 99.2%. Also, Sauro (2006) proposed an online system called the Sample Size Calculator to compute the sample size needed for discovering problems in a user interface as outlined by Lewis (2001). Thus, Table 5.67 was produced after determining the calculator (p) value (from Table 5.66, the p average = 0.11) and the targeted percentages (P). This calculator suggests that 40 users are needed to discover 99%, and from 14 to 16 users are needed to discover from 80% to 85% of the total usability problems. This result is not in line with Nielsen's claim that 5 users are enough to discover from 80% to 85% of usability problems (Turner et al., 2006, Nielsen, 2000b). Furthermore, the rule of 20 users provides optimal results, and it was able to discover 90% of the total usability problems, and this finding in line with Faulkner (2003).

Table 5.67: User number and problems discovered (percentage) for UT in the second experiment

The targeted percentage	The required number of users
99%	40
95%	25
90%	20
85%	16
80%	14

For HE, the figures in Table 5.68 show that, Teams G1-S and G2-T reported 22 and 13 issues, respectively, whereas Team G3-U reported 12 issues. The probability value (p) ranges from 0.08 to 0.1. Thus, the percentage of problems is ranging between 23% and 27%. This result is not in line with previous studies that claim 3 to 5 evaluators can discover 80% of usability problems (Nielsen and Molich, 1990); (Turner et al., 2006); (Hwang and Salvendy, 2010). For DSI, the figures in Table 5.69 show that Teams G1-V and G2-W reported 55 and 47 issues, respectively, whereas Team G3-X reported 33 issues. The probability value (p) ranges from 0.55 to 0.97. Thus, the percentage of problems is ranging between 90% and 99%.

Table 5.68: The performance of the sample size of HE evaluators according to the probability level of problem discovery in the second experiment

Website	Group	Expert and type	Method	# of problems found	# of problems without repetition	Problem discovery rate (P)	% of problems discovered
Google+	G1-S	Ev.1+	HE	6	22	0.1	27%
		Ev. 2^	HE	5			
		Ev. 3+	HE	11			
LinkedIn	G2-T	Ev. 1^	HE	2	13	0.09	25%
		Ev.2+	HE	8			
		Ev.3+	HE	6			
Ecademy	G3-U	Ev.1+	HE	5	12	0.08	23%
		Ev.2^	HE	3			
		Ev.3+	HE	4			

(+) Double expert (^) Single expert (EV) Evaluator

Table 5.69: The performance of the sample size of DSI evaluators according to the probability level of problem discovery in the second experiment

Website	Group	Expert and type	Method	# of problems found	# of problems without repetition	Problem discovery rate (P)	% of problems discovered
Google+	G1-V	Ev. 1^	DSI	16	55	0.97	99%
		Ev.2+	DSI	33			
		Ev.3+	DSI	17			
LinkedIn	G2-W	Ev.1+	DSI	24	47	0.73	98%
		Ev.2^	DSI	8			
		Ev.3+	DSI	27			
Ecademy	G3-X	Ev.1^	DSI	6	33	0.55	90%
		Ev.2+	DSI	28			
		Ev.3+	DSI	23			

(+) Double expert (^) Single expert (EV) Evaluator

Another interesting observation is that, the single expert evaluators were more efficient in discovering usability problems relating to layout, formatting, navigation and search, and content quality, whereas the double expert evaluators were more efficient in discovering usability problems relating to business support, user usability, sociability and management activities; this is likely to be they know, based on their expertise, the factors that lead to the success of websites in this domain. In fact, the double evaluators discovered more problems than the single evaluators in each method and overall. This in line with Nielsen (1992a) when he stated, "usability specialists with expertise in the specific kind of interface being evaluated did much better than regular usability specialists without such expertise, especially with regard to certain usability problems that were unique to that kind of interface". Furthermore, Table 5.70 shows that the double evaluators discovered more problems using each method. It can be seen that two double evaluators found proportions of the usability problems of between 65% and 85%. This result is in line with previous studies that found double evaluators discovered between 74% and 87% of usability problems (Nielsen and Molich, 1990); (Nielsen, 1992a), and is stated in Nielsen (1992a) (Nielsen, 1992a) thus, "for the double specialists, it is sufficient to use between two and three evaluators to find most problems between 81% and 90%". It was concluded from Table 5.70 that the results of the single experts, who used DSI method, provided results approaching or outperforming effectiveness of the double experts, who used HE. This mean the DSI improves the evaluator's performance. This is also the case when Nielsen and Landauer (1993) found that "evaluators with usability expertise found many more problems in a heuristic evaluation than evaluators without such expertise and then evaluators with double expertise".Consequently, the effects of the evaluators' characteristics (double or single) have been also confirmed in the second experiment.

Table 5.70: Results of effects of double evaluators in the second experiment

# Evaluator		Method type	# problems found	Total average proportion
Social. Domain	2 double	HE	44	65%
	1 single	HE	16	17%
	2 double	DSI	152	85%
	1 single	DSI	30	48%

With regard to the Sample Size Calculator, Table 5.71 was produced after determining the calculator (p) value (from Tables 5.68 and 5.69, the p average = 0.09 for HE and 0.75 for

DSI) and the targeted percentages (P). This calculator suggests that 17 evaluators are needed to discover 80% for HE but only 1 evaluator for DSI, and from 20 to 24 evaluators for HE and 2 to 3 evaluators for DSI are needed to discover from 85% to 90% of the total usability problems. Again, this result is not in line with Nielsen's claim that 4±1 evaluators are enough to discover from 80% to 85% of all usability problems for HE (Nielsen and Molich, 1990); (Turner et al., 2006, Nielsen, 2000b); (Hwang and Salvendy, 2010).

Table 5.71: The required number of evaluators for both methods based on Sample Size Calculator in the second experiment

The targeted percentage	The required number of evaluators for HE	The required number of evaluators for DSI
99%	48	4
95%	31	3
90%	24	2
85%	20	2
80%	17	1

5.7 Discussion and findings

In this section the results of both experiments are explored, and the key outcomes highlighted. Also, the lessons learned will be outlined.

5.7.1 Results and outcomes

The key outcomes resulting from this experiment are as follows:

1- The second, third and fifth objectives of this research are to construct the adaptive framework and to generate the domain-specific inspection (DSI) method for two domains by using the adaptive framework. For educational websites, steps of the adaptive framework were followed, and the DSI method and its checklist were built (see Appendix L4 and M2). It consists of five usability problem areas that were developed through the results of the users and experts in steps two and three. For social network websites, the steps of the adaptive framework were followed, and the DSI method and its checklist were built (see Appendix Q4 and R2). It consists of seven usability problem areas that were developed through the results of the users and experts.

2- The forth objective of this research is to examine the performance of the DSI method in terms of discovering real usability problems and in terms of a set of UEM measures.

Another two methods, which were UT and HE, were involved in order to compare their results with DSI. The result of the first experiment has been clearly shown that DSI was able to find 76% of the real problems that were discovered by HE, and it was able to find 46% of the real problems that were discovered by UT. DSI was able individually to reveal 47% of the total number of real usability problems in this experiment, whereas UT and HE were able to reveal only 25% and 7% of the total number of real usability problems, respectively. The HE method did not perform as well as either DSI or UT, based on the number of usability problems discovered during this experiment. In terms of the second experiment, DSI was able to find 48% of the real problems that were discovered by HE and UT, and HE was able to find 12% of the real problems that were discovered by UT and DSI. UT was able to find 29% of the real problems that were discovered by DSI and HE. The HE method in the second experiment did not perform as well as either DSI or UT, based on the number of usability problems discovered. Consequently, this result between HE and UT in both experiments is in line with other studies (Jeffries and Desurvire, 1992); (Thyvalikakath et al., 2009). Also, DSI in both experiments was better at discovering catastrophic, major, minor and cosmetic real problems. The methods' performances are statistically different; the statistical tests show that there is statistical significance in terms of the number of problems discovered and time spent. Thus, this finding confirms that UT is more powerful than HE. DSI improves these methods and the effort to propose the adaptive framework and create DSI could be a valuable contribution to the field of usability evaluation.

In terms of the performance metrics in both experiments, DSI was more efficient, thorough and effective in terms of identifying real problems relative to total time spent and in its ability to identify real usability problems relating to user interfaces than the other methods. UT is the second good method. HE delivered the worst result in identifying a sufficient number of real problems; however, it is the cheapest to use. Moreover, DSI is slightly more expensive than HE, although it is cheaper than UT. One expert commented that "DSI helps me to guide my thoughts in judging the usability of the website through clear guidelines that included all aspects of the social network/educational websites' quality". As a result, it is unsurprising that the DSI method revealed a number of problems not discovered by the other two methods. The experts that used HE seemed to have their

confidence undermined whilst performing the evaluation; for example, one expert commented that " when I performed the evaluation, I found no readily applicable heuristic within HE for performing some of the main functions in these social network websites, such as sociability, management activities, business support, and security and privacy". Another said " HE is no readily applicable heuristic for evaluating the main functions in educational websites, such as Educational process and management". Consequently, HE performed poorly in discovering problems. The UT method performed modestly against DSI, and well against HE, based on the number of problems identified. Thus, the findings indicate that it is 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 the social network domain and educational domain, should be employed. Furthermore, the adaptive framework provided optimal results regarding the identification of comprehensive 'usability problem areas' on the educational and social network websites, with minimal input in terms of cost and time spent in comparison with the employment of the other two usability evaluation methods. The framework was used here to generate DSI, which helped to guide the evaluation process as well as reducing the time that it would have taken to identify these usability issues through current evaluation methods. In terms of the definition of missed problem given by Cockton and Woolrych (2002), we can consider that the problems found by any one method and not found by the others as being missed problems. From this standpoint, DSI missed discovering 28 real usability problems in the first experiment. However, HE and UT missed 77 and 56 real usability problems, respectively. In the seconed experiment, DSI missed discovering 80 real usability problems. However, HE and UT missed 149 and 117 real usability problems, respectively. The above findings facilitate decision-making with regard to which of these methods to employ, either on its own or in combination with another, in order to identify usability problems on educational websites or social network websites. The selection of the method or methods will depend on the types of problem good identified by each of them. In conclusion, DSI improves usability evaluation methods (UEMs) and the effort to propose the adaptive framework and hence to create DSI for a particular area could be a valuable contribution to the field of usability evaluation.

3- The sixth objective is to investigate the role of the number of evaluators and users needed in usability studies. In the first experiment, two rules were examined, which were the 10±2 rule for UT and the 4±1 rule for HE and DSI. For user sample size (UT), the 8 users found percentages of usability problems ranging between 80% and 92%, whereas the 12 users found percentages of usability problems ranging between 98.8% and 99.4%. The average of the probability value for UT was calculated (p = 0.28), and the Sample Size Calculator was used. The results reveal that 14 users are needed to discover 99%, and from 5 to 6 users are needed to discover from 80% to 85% of the total usability problems. This result is in line with Nielsen's claim that 5 users (the magic number) are enough to discover from 80% to 85% of all usability problems. For evaluator sample size (HE and DSI), it was difficult to recruit 5 evaluators for each group. So, 4 evaluators were recruited. The results show that the percentage of problems for four evaluators with HE ranges between 75% and 85%. However, the percentage of usability problems found by four evaluators with DSI ranges between 98% and 99%. Again, the average probability value for HE was calculated (p = 0.33), and the Sample Size Calculator was used. The results reveal that 11 evaluators are needed to discover 99% for HE, and that 4 to 5 evaluators are needed to discover from 80% to 85%, of the total usability problems. On the other hand, the average probability value for DSI was calculated (p = 0.64), and the Sample Size Calculator was used. The results reveal that 4 evaluators are needed to discover 99% for DSI and 1 to 2 evaluators are needed to discover from 80% to 85%, of the total usability problems. Thus, these results confirm Nielsen's claim that 4±1 evaluators are enough for HE to discover from 80% to 85% of usability problems (Nielsen and Molich, 1990).

In the second experiment, two rules were examined, which were the 5 and 20 rule for UT and the 3 rule for HE and DSI. For user sample size (UT), the 5 users found percentages of usability problems ranging between 15% and 37.6%, whereas the 20 users found percentages of usability problems ranging between 88% and 98.2%. For more examination of this issue, the average of the probability value for UT was calculated (p = 0.11), and the Sample Size Calculator was used. The results reveal that 20 users are needed to discover 90%, and from 14 to 16 users are needed to discover from 80% to 85% of the total usability problems. This result is not in line with Nielsen's claim that 5 users (the magic number) are enough to discover from 80% to 85% of all usability problems. For

evaluator sample size (HE and DSI), the results show that the percentage of problems for three evaluators with HE ranges between 23% and 27%. However, the percentage of usability problems found by three evaluators with DSI ranges between 98% and 99%. For more examination of this issue, the average probability value for HE was calculated (p = 0.09), and the Sample Size Calculator was used. The results reveal that 24 evaluators are needed to discover 90% for HE, and that 17 to 20 evaluators are needed to discover from 80% to 85% of the total usability problems. On the other hand, the average probability value for DSI was calculated (p = 0.75), and the Sample Size Calculator was used. The results reveal that 2 evaluators are needed to discover 99% for DSI, and 1 evaluator is needed to discover 80% of the total usability problems. Thus, the results of the Sample Size Calculator are in line with the results in this experiment. Also, the results in this experiment for the both methods refute Nielsen's claim that 4±1 evaluators are enough to discover from 80% to 85% of usability problems (Nielsen and Molich, 1990).

In conclusion, it can be seen that the sample size can impact on evaluation results and so it should be considered before starting any usability studies. However, it is difficult to identify specific sample size for finding all usability problems.

4- The seventh objective is to investigate the relationships amongst the usability measures used. This study confirms statistically that there is a relationship between number of usability found and time spent by the users or evaluators. This means that when participants spend more time, they will discover more problems.

5- In terms of evaluator effect, the types of evaluator (single and double) played a role in affecting the evaluation results. In the first experiment, the single evaluators in HE discovered 18 problems, but the double evaluators discovered 29 problems. Furthermore, the single evaluators in DSI discovered 60 problems, whereas the double evaluators discovered 70 problems. In the second experiment, the single evaluators in HE discovered 16 problems, but the double evaluators discovered 44 problems. Furthermore, the single evaluators in DSI discovered 30 problems, whereas the double evaluators discovered 152 problems. Another effect is the methods that were used by the evaluators. The results show that the performance of the evaluators in discovering problems was affected, so there were differences in terms of problem discovered between the evaluators' groups who evaluated

the same website (but they used different methods). Thus, the effect of using different methods on the part of the evaluators (or recruiting different evaluator types) has been confirmed in this study, and these results are in line with previous studies (Nielsen, 1992a); (Hertzum and Jacobsen, 2001).

5.8 Conclusion

This chapter analysed and discussed the collected data for the both experiments. Contrary to most 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. The work presented here illustrates and evaluates this process for the generation of the DSI method to assess the usability of educational and social network websites. DSI outperformed both HE and UT, even when taken together. This clearly represents a step in the right direction. The next chapter will discuss and compare the findings of the two experiments in more detail. Also, the research question will be answered.

Chapter 6: Discussion and Recommendations

6.1 Introduction

This chapter explores and discusses the results of the research presented in Chapters 5. It commences with the aims and objectives of the research referred to in Chapter 1 followed by a comparison of the results achieved in Chapters 5 with those of the Literature Review presented in Chapter 2. As far as the researcher is aware, this is the first study of its kind to develop the adaptive framework to generate the domain specific inspection (DSI) method. Throughout the thesis a combination of research methods were used to build the adaptive framework and to validate practically this framework by generating a DSI method for assessing the usability attribute. In addition, two evaluation methods were employed to validate the generated DSI method for both educational domain and social networks domain in terms of a set of measurements.

This research has achieved its aim by answering the research question and developing and testing the adaptive framework for evaluating the usability of a selected product, as presented in Chapter 3, to address a specific gap in the literature regarding the lack of a methodological framework that can be used systematically to generate a domain-specific evaluation method, which can help to improve the current usability evaluation methods and usability assessment process. Significant knowledge has been gained in this research within the context of its results and outcomes, which can be categorised into four important sections. The first area is the effectiveness of the chosen usability evaluation methods, as shown in Section 6.2. In this section, the research question will be answered. The second area is the usefulness of the adaptive framework (Section 6.3). The third area is identification the sample size required for DSI, HE, and UT based on the result of this research (Section 6.4). The fourth area is the set of recommendations (Section 6.5). These four areas will be discussed below in further detail.

6.2 The Effectiveness of the chosen usability evaluation methods

This research has employed three evaluation methods, namely, heuristic evaluation (HE), usability testing (UT) and domain specific inspection (DSI). The HE and UT methods have been employed in different research contexts with a view to measuring website usability, and has found strengths and weaknesses for each method (as mentioned in sections 2.3.2.1 and 2.3.2.2). Briefly, previous studies recommend conducting UT with HE, as each complements the other (Nielsen, 1992a); other studies emphasise the importance of developing UEMs as a matter of priority in order to increase their effectiveness and to identify the most acceptable approach for assessing such interactions (Hertzum, 2006). Also, Nielsen (1992a) pointed out that the best way to improve the HE is by using a double evaluator when he noticed that the number of usability problems uncovered is significantly greater than with a regular specialist. However, this will make using HE more difficult and expensive regarding to the struggle recruiting sufficient number of double evaluators and the cost of hourly work rate. Furthermore, other studies recommend using certain methods to reveal certain types of problems (Doubleday et al., 1997). 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 current usability methods, represents an area lacking in usability testing. Therefore, this research aims to improve UEMs through developing adaptive framework. In the context of this research, the two experiments have produced a number of interesting results in terms of the set of measurements associated with each method. In order to provide greater detail, the following section will discuss and compare the research findings with a number of findings from the current literature.

6.2.1 Time spent

Time spent is the metric most often used to measure the efficiency attribute. It indicates the time spent by evaluators or users to complete their work (in minutes or hours). In this research, two experiments were conducted; the first on three educational websites and the second on three social network websites. In terms of employing the three methods in the first experiment, the average time taken for undertaking the three evaluations were: using HE, 24.25 minutes, and, using DSI, 42.58 minutes; whereas the UT average was 99.33 minutes.

Statistically, the differences among the methods' performance in terms of time spent was examined. The kruskal wallis test was used; this revealed a significant difference between the three methods in terms of time spent on identifying usability problems where $p < 0.05$, as shown in Table 5.56. Furthermore, the cost of employing the formula was applied, as mentioned in Chapter 2, which included time spent on evaluation, testing, collecting and analysing the data, as shown in Table 5.59. The results showed that HE cost $1,017, DSI cost $1,206 and UT cost $3420. In this regard, the research question is answered when the results show that there are differences between DSI, HE and UT in terms of time spent and employment costs. Based on the above results HE less time consuming and is cheaper than DSI and UT. However, DSI consumes less time and is cheaper than UT.

In terms of the second experiment, the average time spent conducting the three evaluations using HE was 56 minutes and using DSI was 72 minutes, whereas the UT average was 392.66 minutes. Statistically, the differences among the methods' performance in terms of time spent were examined. The kruskal wallis test was used, which revealed a significant difference between the three methods in terms of time spent on discovering usability problems, where $p< 0.05$, as shown in Table 5.57. Furthermore, the cost of employing the formula was applied, as mentioned in Chapter 2, which included time spent on the evaluation, testing, collecting and analysing the data as shown in Table 5.59. The results showed that HE cost $706.66, DSI cost $863.33, and UT cost $4275. In this regard, DSI is less time consuming and cheaper than UT. However, the HE is less time consuming and is slightly cheaper than DSI. Then the research question is clearly answered in both experiments.

With regard to comparing the above findings with those of previous studies, many studies have reported that UT costs more money and time than HE, as shown in Table 6.1. It is clear that the differences in the results of HE and UT amongst these studies relate to various factors, such as the differences in the experience of the number users and experts who are involved in those studies, their characteristics, tools used, the tested products, plus time spent on setting up, designing, collecting and analysing the data. In conclusion, the results of this research are in line with these studies. Thus, the adaptive framework has proved successful in generating DSI methods in two different domains, in terms of cost and time compared to UT. The differences between DSI and HE in terms of cost of employing are slight; thus, this framework is able to generate a discount method.

Table 6.1: Cost of employing usability evaluation methods (Hasan, 2009)

Previous study		UT	HE	DSI
(Doubleday et al., 1997)		125 hours This time included 25 hours conducting 20 users' sessions, 25 hours of evaluator time supporting during users' sessions and 75 hours of statistical analysis	33.30 hours This time included 6.25 hours of five experts' time in the evaluation, 6.25 hours of evaluators' time taking notes and 21 hours transcription of the experts' comments and analysis	
(Jeffries et al., 1991)		199 hours This time was spent on analysis. Six subjects participated in this study	35 hours This time was spent on learning the method and on becoming familiar with the interface under investigation (15 hours) and on analysis (20 hours). Four usability specialists conducted this method.	
(Law and Hvannberg, 2002)		200 hours This time was spent on the design and application of this method. Ten subjects participated in this study.	9 hours This time was spent on the design and conduction of this method by two evaluators	
(Hasan, 2009)		326 hours This time included 136 hours setup and designing, 20 hours collecting data from 20 users' sessions, and 170 hours analysing the data	247 hours This time included 128 hours setup and designing, 15 hours collecting data from five web experts, and 104 hours analysing the data	
In this research	First experiment	50 hours This includes the time spent by 20 users (5 hours), 30 hours collecting data from the evaluation sessions, 15 hours analysing the data.	30.5 hours This includes the time spent by 8 evaluators (4.7 hours), 7.8 hours collecting data from the evaluation sessions, and 18 hours analysing the data.	37.2 hours This includes the time spent by 8 evaluators (8.4 hours), 9 hours collecting data from the evaluation sessions, and 19.8 hours analysing the data.
	Second experiment	74.53 hours This includes the time spent by 25 users (19.63 hours), 36 hour collecting data from the evaluation sessions, 18.9 hours analysing the data.	21.2 hours This includes the time spent by 3 evaluators (8.3 hours), 6 hours collecting data from the evaluation sessions, and 9.9 hours analysing the data.	25.9 hours This includes the time spent by 6 evaluators (10.6 hours), 3.9 hours collecting data from the evaluation sessions, 11.4 hours analysing the data.

However, it is important to take into consideration the time taken to develop the DSI method which is called fixed cost. In fact that the DSI method has two costs, which are a fixed cost and a variable cost. The fixed cost is the cost of applying the adaptive framework to develop the DSI method. The variable cost is the cost of applying the newly developed DSI method on a specific website. However, the HE and UT have just the variable cost which is the cost of applying each method on a specific website. The results in Table 6.1 did not include the time or the fixed cost of developing the DSI methods for the educational and social network domains. It shows just the variable costs for each method. Table 6.2 shows the fixed cost for DSI and the variable costs of other methods. The last column in this table shows the total time spent on applying the adaptive framework and the time spent on applying HE and UT.

It is clear that DSI is more expensive than HE and UT in terms of the fixed cost in developing DSI for the targeted domains. Thus, using HE and UT to evaluate one website is better than applying the adaptive framework for generating the DSI method and then to use the DSI in evaluating one website only. This is due to that the fixed cost of DSI plus its variable cost will be much higher (7 weeks + time spent on using DSI to evaluating a website) than the HE (1 week which is the time spent on using HE to evaluating a website) and UT (4 weeks which is the time spent on using HE to evaluating a website). However, if there are four websites and they are all evaluated, applying the adaptive framework will be much cheaper than using UT four times (4 weeks * 4 times = 16 weeks by using UT, whereas DSI = just 7 weeks + time spent of evaluation). If there are 8 websites, applying the adaptive framework will be much cheaper than using HE 8 times (1 weeks * 8 times = 8 weeks by using HE, whereas DSI = just 7 weeks + time spent of evaluation). This means that using the adaptive framework for evaluating a number of websites will be better than using traditional evaluation methods. Consequently, it is clear that while the process for constructing the DSI can take considerable time, once the DSI is constructed, it is relatively faster, cheaper and more productive at discovering usability problems.

Table 6.2: Time spent for developing DSI methods versus other methods

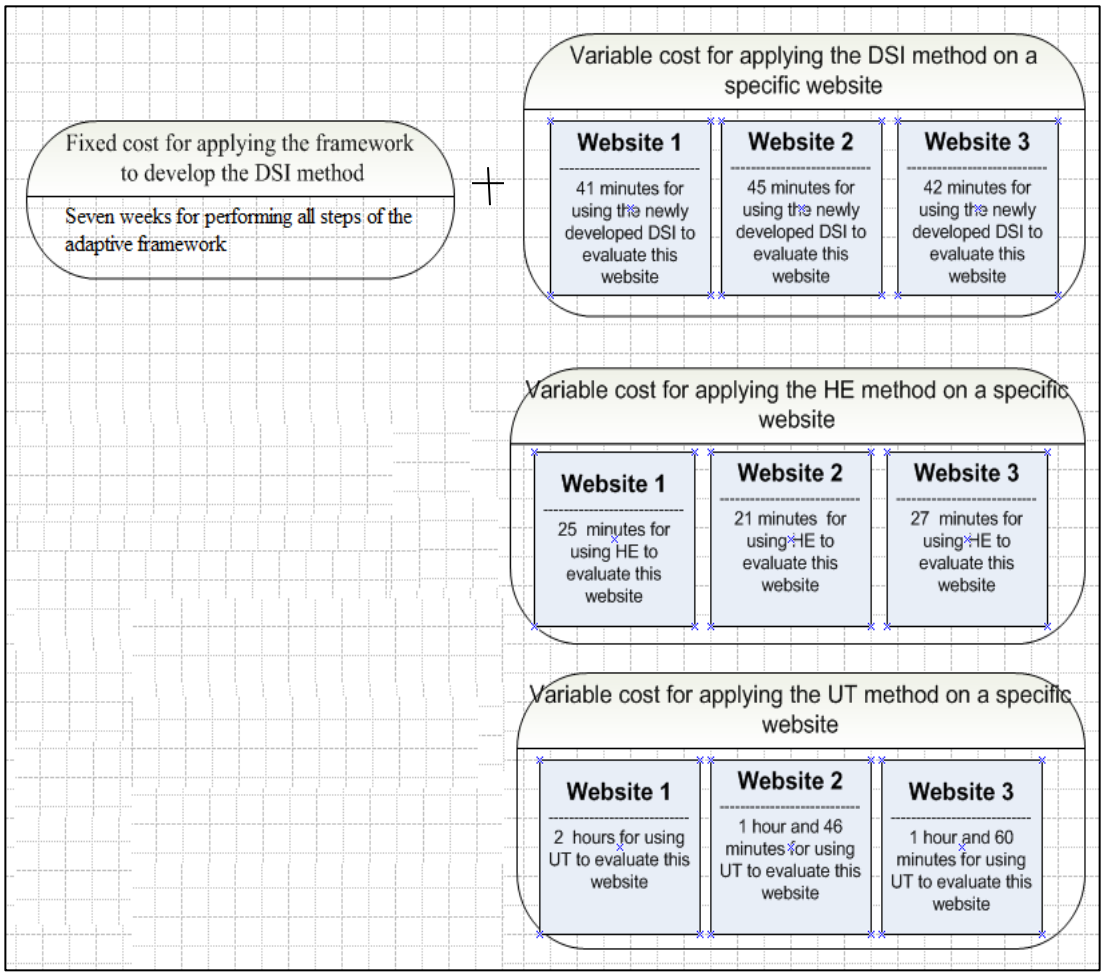
Methods / **Steps of the adaptive framework**			**Time spent**					**Total of weeks**
			Step 1	**Step 2**		**Step 3**	**Step 4**	
			literature review	**Context meeting**	**User testing (user input)**	**Focus group (expert input)**	**Draw up DSI**	
First and Second Experiments	**Variable cost**	HE	1 week	N/A	N/A	N/A	N/A	1 week
		UT	2 weeks	Two hours for meeting session with 5 users for identifying users and tasks	2 weeks for setting up + conduct big experiment with 20 users	N/A	N/A	4 weeks
	Fixed cost	DSI	3 weeks		1 week for setting up + conduct mini-testing with 10 users	1 week for recruiting 5 experts + preparing material + conducting discussion	2 week for establishing a DSI and its checklist	7 weeks

* N/A= Not applicable

Furthermore, Figure 6.1 shows the fixed cost and the variable costs for the three methods on the educational websites as real example. It is clearly seen that the fixed cost for the DSI

method is much higher (7 weeks) than others methods. So, if the DSI is used to evaluate only one website, it will be there a high fixed cost (7 weeks) and low variable cost (41 minutes) which together will be still higher than one variable cost from other methods (25 minutes for HE and 2 hours for UT). However, if there are three websites that will be evaluated by DSI method, it will be effectively saving the fixed cost due to that there is no needed to do all steps from scratch all times (7 weeks + 41 minutes+ 45 minutes+ 42 minutes) . In other side, there is big variable cost all times when other methods are used for evaluating three websites and more (25 minutes + 21 minutes + 27 minutes for HE and 2 houres + 1 hour and 46 minutes + 1 hour and 60 minutes). The above calculations will be a clear difference when they are ten or more websites evaluated.

Figure 6. 1:Comparing between the fixed cost and the variable cost for the three methods



6.2.2 Number of problems

This is the most important measure that can be used to assess the effectiveness of usability evaluation methods (UEMs). After reporting the identified usability problems, they then need to be rated in terms of their priority for fixing. However, any problem that is discovered by a non-real user (e.g. evaluator) must be examined in order to identify whether the problem is a real problem or a false positive problem. In the first experiment, HE revealed 25 problems, DSI revealed 74 problems, whereas UT revealed 41 real problems, as shown in Figure 5.3. When the falsification test was conducted, on the problems revealed by HE and DSI, HE revealed 25 real problems, whereas DSI revealed 61 real problems, as shown in Figure 5.3. In order of priority for the identified problems to be fixed, Table 5.39 shows that HE was not able to identify any catastrophic problems; however, it was able to identify 1 major, 2 minor and 3 cosmetic problems. DSI was able to identify 6 catastrophic, 7 major, 18 minor and 11 cosmetic problems that were not revealed by the other methods. UT was not able to discover any major problems; however, it discovered 1 catastrophic, 5 minor and 16 cosmetic problems. Statistically, the differences among the methods' performances, in particular with regard to the problems found, were examined. The kruskal wallis test was used, thus, the results found significant differences between the results of the three methods, where $p < 0.05$, as shown in Table 5.49. In this regard, the research question is answered when the results show that there are differences between DSI, HE and UT in terms of the number of usability problems found and their severity. Based on the above results DSI discovered more real usability problems with high severity than HE and UT.

In the second experiment, HE revealed 55 problems and DSI revealed 119 problems, whereas UT revealed 79 real problems, as shown in Figure 5.4. When the falsification test was undertaken on the problems revealed from HE and DSI, HE revealed 47 real problems, whereas DSI revealed 116 real problems, as shown in Figure 5.4. In terms of priority of the identified problems in order to fixing, Table 5.40 shows that HE was able to identify 1 catastrophic problem, 4 major, 11 minor and 8 cosmetic problems. However, DSI was able to identify 6 catastrophic, 24 major, 34 minor and 29 cosmetic problems that had not been revealed by the other methods. UT was able to identify 7 catastrophic, 11 major problems, 17 minor and 21 cosmetic problems. Statistically, the differences among the methods' performances, particularly with regard to the problems found, were examined. The kruskal

wallis test test was used; the results found significant differences between the results of the three methods, where $p < 0.001$, as shown in Table 5.53. In this regard, the DSI method outperforms HE and UT in terms of the number of usability problems found and their severity. Based on the above results, the research question is answered when the results of both experiments show that there are differences between DSI, HE and UT in terms of the number of usability problems found and their severity.

With regard to comparing the above findings to those of previous studies, it was found that HE identifies more usability problems compared to UT (Jeffries et al., 1991); (Doubleday et al., 1997); (Fu et al., 2002); (Jeffries and Desurvire, 1992); (Desurvire et al., 1992a); (Desurvire et al., 1992b); (Law and Hvannberg, 2002); (Hasan, 2009). However, our results are not in line with their findings. The reason behind their findings is that they used HE to evaluate most parts of the interfaces, whereas they used UT to perform only specific tasks when users interacted with the interfaces (Hasan, 2009). Thus, it is unsurprising that HE identified more problems. Furthermore, according to previous research, HE is more effective than UT in identifying uniquely minor problems, whereas UT was more effective than HE in uniquely identifying major problems (Law and Hvannberg 2002). Moreover, they reported that UT is more accurate and objective than HE. The results obtained in this research are in line with these findings. In this regard, many studies recommended the use of both methods, as they are complementary (Nielsen, 1992a); (Law and Hvannberg, 2002); (Jeffries and Desurvire, 1992); (Jeffries and Desurvire, 1992); (Fu et al., 2002); (Kantner and Rosenbaum, 1997); (Mack and Nielsen, 1994). In this research, the results of UT and HE in both experiments confirm conducting UT with HE in order to overcome the shortcomings of each method. However, DSI, as created from the proposed adaptive framework, refutes this recommendation. In conclusion, the adaptive framework was successful in generating the DSI methods for the two domains, as it proved superior in identifying real usability problems with high severity and without the need to conduct UT or HE in parallel. These findings answered the research question. Thus, this framework is able to generate a productive and powerful method.

6.2.3 Usability metrics

There are various usability metrics; however, efficiency, thoroughness, validity and effectiveness were the metrics used in this research, as mentioned in sections 2.4.4. In terms

of the efficiency of the three methods assessed through the three websites in the first experiment, the efficiency formula was used by dividing the numbers of usability problems detected by the total time spent, as mentioned in Chapter 2. The results showed that the efficiency average score of HE was 0.4, DSI was 0.6, and UT was 0.4, as shown in Table 5.58. Thus, DSI was more efficient than HE and UT. Furthermore, the thoroughness formula was used by dividing the number of real usability problems found by the total number of real usability problems, as mentioned in Chapter 2. The results showed that the thoroughness of HE was 0.3, DSI was 0.7, and UT was 0.5, as shown in Table 5.58. Therefore, DSI is more thorough than HE and UT in identifying real usability problems. Additionally, the validity formula was used by dividing number of real usability problems found by number of issues identified as a usability problem. The results showed that the validity of HE was 1, DSI was 0.8, and UT was 1, as shown in Table 5.58. This means that HE and UT were slightly more valid than DSI in terms of identifying usability problems accuratelyin the first experiment. Moreover, the effectiveness formula was used by multiplication thoroughness to validity. The results showed that the effectiveness of HE was 0.3, DSI was 0.6, and UT was 0.5, as shown in Table 5.58. Therefore, DSI is more effective than HE and slightly more effective than UT in terms of identifying usability problems relating to the user interface. In this regard, the research question is answered when the results show that there are differences between DSI, HE and UT in terms of UEM performance metrics. Based on the above results, the DSI is more efficient, thorough, and effective than either HE or UT in terms of identifying real usability problems, and identifying usability problems relating to the user interface. However, HE and UT were slightly more valid than DSI in terms of identifying usability problems accurately.

In the second experiment, the efficiency formula was used. The results showed that the efficiency of HE was 0.84, DSI was 1.7, and UT was 1.67, as shown in Table 5.58. Thus, DSI was more efficient than HE and slightly more efficient than UT. Furthermore, the thoroughness formula was used. The results showed that the thoroughness of HE was 0.12, DSI was 0.5, and UT was 0.4, as shown in Table 5.58. It can be seen that DSI was more thorough than HE and slightly more thorough than UT in identifying real usability problems. Additionally, the validity formula was used. The results showed that the validity of HE was 0.7, DSI was 0.9, and UT was 1, as shown in Table 5.58. This means that UT was more valid than HE and slightly more valid than DSI in terms of discovering usability problems

accurately. Moreover, the effectiveness formula was used. The results showed that the effectiveness of HE was 0.1, DSI was 0.5, and UT was 0.4 as shown in Table 5.58. Therefore, DSI was more effective than HE and slightly more effective than UT in terms of identifying usability problems relating to the user interface. In this regard, the research question is answered when the results show that there are differences between DSI, HE and UT in terms of UEM performance metrics. Based on the above results, the DSI is more efficient, thorough, and effective than HE and UT in terms of identifying real usability problems and identifying usability problems relating to the user interface. Furthermore, DSI is more valid than HE in terms of identifying usability problems accurately; however, DSI is slightly less valid than UT in terms of discovering usability problems accurately.

In conclusion, the adaptive framework was successful in generating DSI methods, which are more efficient, thorough and effective than either HE or UT. Hertzum (2006) aspired to develop UEMs in order to increase their effectiveness and efficiency and to identify the most acceptable approach for assessing such interactions; this research has achieved this goal.

6.2.4 Usability problem areas

A gap exists in some previous studies, as they failed to provide detail with regard to the specific usability areas that could be identified by UT and HE, or by any recently developed method (Hasan, 2009), particularly in the educational and social network domains; hence this research aims to address that gap. When the steps of the adaptive framework were applied on the educational and social network domains, five areas were identified in the former, and seven areas were identified in the latter. Each method was able to identify usability problems related to each problem area; these areas were used to provide a structure to explain and comprehend the identified usability problems. In the first experiment, the real problems were identified for the three methods after conducting the falsification test. Consequently, those problems were classified according to the identified usability problem areas for the educational domain, namely, user usability, motivational factors, content information and process orientation, learning process, and design and media usability. Table 5.54 shows that HE failed to expose any usability problems in two main areas, namely, motivational factors and learning process; furthermore it failed to identify a sufficient number of usability problems in the content information and process orientation areas. Moreover, UT performed better than DSI and HE in identifying usability problems in the user usability area but failed

to identify a sufficient number of usability problems in motivational factors and learning process. In contrast, DSI helped to identify large numbers of real usability problems in all usability areas on the three websites (61): 15 problems in user usability, 3 problems in motivational factors, 5 problems in content information and process orientation, 6 problems in learning process, and 32 problems in design and media usability. Overall, HE identified 25 real usability problems: 15 problems in user usability, no problems in motivational factors, 2 problems in content information and process orientation, no problems in learning process, and 8 problems in design and media usability. UT performed better in identifying usability problems (41) in all usability areas on the three websites: 23 problems in user usability, 2 problems in motivational factors, 4 problems in content information and process orientation, 2 problem in learning process, and 10 problems in design and media usability. Based on the above results, the research question is answered when the results show that there are differences between DSI, HE and UT in terms of finding usability problem areas. The DSI identified notably more usability problems in all five areas than either HE or UT.

In the second experiment, the real problems were again identified for the three methods after conducting the falsification test. Consequently, those problems were classified according to the identified usability problem areas for the social network domain, namely, layout and formatting, content quality, security and privacy, business support, user usability, sociability and management activities, accessibility and compatibility, and navigation system and search quality. Table 5.55 shows that HE failed to expose any usability problems in three main usability problem areas: security and privacy, business support, and accessibility and compatibility. Furthermore, UT failed to identify a sufficient number of usability problems in the accessibility and compatibility area. Overall, DSI helped to identify large numbers of real usability problems in all usability areas on the three websites (116): 15 problems in layout and formatting, 25 problems in content quality, 8 problems in security and privacy, 4 problems in business support, 40 problems in user usability, sociability and management activities, 4 problems in accessibility and compatibility, and 20 problems in navigation system and search quality. HE identified 47 real usability problems: 9 problems in layout and formatting, 4 problems in content quality, no problems in security and privacy, no problems in business support, 19 problems in user usability, sociability and management activities, no problems in accessibility and compatibility, and 15 problems in navigation system and search quality. UT performed better in discovering usability problems (79 in all

usability areas): 13 problems in layout and formatting, 12 problems in content quality, 4 problems in security and privacy, 8 problems in business support, 19 problems in user usability, sociability and management activities, 2 problems in accessibility and compatibility, and 21 problems in navigation system and search quality. In this regard, the DSI method outperforms HE and UT in terms of identifying more problems in usability problem areas. Based on the above results, the DSI discovered sufficiently more usability problems in all seven areas than HE and UT.

For a deeper analysis, the contents of the unique usability problems identified by UT and HE were compared to those of previous studies, UT identified usability problems related to user performance, a lack of clear feedback and help facilities, functionality and learnability problems, navigation, excessive use of complex terminology (technical jargon), inappropriate choice of font size, use of an inappropriate format for links, plus some consistency problems (Simeral and Branaghan, 1997); (Jeffries et al., 1991); (Doubleday et al., 1997); (Fu et al., 2002); (Law and Hvannberg, 2002); (Mariage and Vanderdonckt, 2001), (Hasan, 2009). However, HE identified usability problems related to interface features and interface quality, appearance or layout of an interface, inconsistencies with the interface, slow interface response time in displaying results, compatibility, security and privacy issues (Nielsen and Phillips, 1993); (Doubleday et al., 1997); (Nielsen, 1992a); (Law and Hvannberg, 2002); (Simeral and Branaghan, 1997); (Fu et al., 2002); (Tan et al., 2009); (Hasan, 2009). In this research, UT identified the majority of usability problems that related to user performance and links issues (i.e. misleading), as shown in Appendix N and S, whereas HE identified the majority of usability problems that related to interface features, and inconsistency problems with the interface, as shown in Appendix N and S. Consequently, the results of HE and UT are in line with those of previous research. Furthermore, DSI identified the usability problems identified by UT and HE but also identified problems related to interface quality, layout of an interface, compatibility, inappropriate choice of font size, and functionality and navigation, as shown in Appendix N and S.

In conclusion, the above findings illustrate the effectiveness of the three methods in terms of their ability to identify specific usability problem areas. HE failed to identify any problems in some usability problem areas. In addition, HE was unable to consider the user's role in the evaluation process and thus could not identify the problems met in the testing session. UT

identified a sufficient number of problems in all usability problem areas. However, DSI identified more problems in all usability problem areas, and was able to consider the user's role in the evaluation process. These results can be seen as making a good contribution to the field of UEMs, as the increased effectiveness of the DSI method has been demonstrated.

6.3 The usefulness of the adaptive framework

The literature showed that the previous studies used UT and HE as the most common usability evaluation methods in the evaluation of such websites. Subsequently, many studies have recognised the importance of enhancing the current usability methods; thus, the literature review includes many frameworks and models that have been since been published to update usability evaluation methods (Alias et al., 2013); (Gutwin and Greenberg, 2000). However, these frameworks and models are not applicable to all products, as they were developed to address certain aspects of usability in certain areas (Coursaris and Kim, 2011). Furthermore, those studies did not describe the benefits or drawbacks of either HE or UT (in terms of the specific areas of the usability problem that they could or could not detect) (Hasan, 2009), particularly, in educational and social network websites. Moreover, those studies did not use a context-specific method, and consider expert and user perspectives together. Indeed, there is a clear need for an effective and appropriate methodology for evaluating the emerging domains/technology in order to measure their levels of efficiency, effectiveness and satisfaction, and ultimately to improve their quality.

This finding and the criticality of website usability has encouraged the researcher to formulate the adaptive framework. This framework is applicable across numerous website domains. In other words, it is designed to be capable of adapting to any website domain, and can be applied without the need to conduct UT. However, developing and testing a method is not a quick process and should involve a number of key stages. Chapter 3 discussed the adaptive framework and Section 4.4.8 discussed the validation phases for the adaptive framework. It has been clearly shown that the adaptive framework was able to generate the DSI method for both domains and was able to identify more real problems than either UT or HE. Furthermore, DSI proved better in discovering catastrophic, major, minor, and cosmetic real problems. It appeared to guide the evaluators' thoughts in judging the usability of the website through clear guidelines that included all aspects of the educational and social network quality aspects of the websites, which were represented in the five and seven

usability areas; as a result, it is unsurprising that the DSI method revealed a number of problems not identified by the other two methods. One expert commented that ‘’the DSI method was appropriate for this task and its usability areas and terminology were more appropriate for the interfaces of the tested websites”. One more expert commented that ‘’from the first time I used the DSI method I had a doubt in the ability of the DSI, and now I feel more confident and willing to use it than before and this is due to its focused heuristics with classified problem areas’’. Also, this expert commented that ‘’before becoming involved in this experiment my opinion was all inspection methods are useless, however DSI helps to change my opinion because DSI is simple and able to find more problems’’. Another expert commented that ‘’the DSI method was easy to remember and it was very focused on the problems associated with usability for these websites’’. The HE method did not perform as well as either DSI or UT, based on the number of usability problems discovered during both experiments. The experts that used HE appeared to have their confidence undermined whilst performing the evaluation; for example, one expert commented that ‘’when I performed the evaluation, I found no readily applicable heuristic in the HE for evaluating some of the main functions in the websites so I recommended to extend the current heuristics by adding some heuristics’’. Another expert commented that ‘’the HE is a good method, but it is so difficult to use and requires more knowledge’’. Another expert commented that ‘HE is helpful in general and DSI is an extremely helpful tool for the product that it is designed for’’. Consequently, HE performed poorly in identifying 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 has been generated here for educational and social networks domains, should be employed. Furthermore, the adaptive framework provided optimal results regarding the identification of comprehensive ‘usability problem areas’ on the educational and social network websites, with minimal input in terms of cost and time spent in comparison with the employment of usability evaluation methods. Moreover, the DSI checklists developed from the adaptive framework as an adaptive tool for evaluating the usability of both domains support evaluators in the evaluation process. They also provide an opportunity for designers, developers, instructors and website owners to design an interactive interface, to assess the quality of existing systems, or to

choose the usability area(s) that they felt needed to be evaluated. Furthermore, they also afford anyone to adopt any area of usability or any principle to determine the usability problems related to the five or seven specific areas in both domains.

In order to evaluate the usefulness of the adaptive framework, there are two approaches. The first approach is to redesign the tested websites and to repeat the data gathering process by using the DSI method. The second approach is to take into consideration the usability problem report for DSI from an owner perspective. The second approach was chosen because it was difficult to ask the owners to redesign their websites, and this approach was also related to the timeframe of this research. The usability problem reports (as shown in Appendix N and S) were sent by email to the website owners and included a set of questions to obtain qualitative data from their feedback about the usefulness of the problems discovered by DSI method as shown in Appendix Z. As mentioned in Section 5.6.3.3, the websites have been changed and their usability has been improved after the reports were received. Many problems were fixed and the majority of these problems were found when using DSI. Also, positive responses were received from some website owners as shown in Appendix U1 and U2. This indicates that the adaptive framework is useful in terms of discovering a set of real problems that have to be fixed as a matter of high priority.

The adaptive framework provides steps on how to develop a domain-specific inspection method (DSI) and its checklist. The examples used in this research were in the domain of free educational and social networks websites, but the adaptive framework is general and can be used with other website domains. Based on the research results, it can be concluded that the adaptive framework is most valuable when evaluating many websites (based on the calculations in Table 6.2 and Figure 6.1, four website will be worth to use the adaptive framework). This means that if anyone wants to evaluate one website, it is better to use the traditional evaluation methods rather than the adaptive framework. Also, the adaptive framework is most valuable for dominant domains that are not too broad. Thus, it will not work if it is used to cover a domain that has number of websites that have different aims and massive features. For example, it will not work for evaluating whole topics of the science domain such as Biology, Astronomy, Chemistry, etc. Also, it will not work if the definitions and classifications of the chosen domain is not clear or identified. Additionally, it will not work on the websites that are out of the scope of the chosen domain. Additionally, this

framework will not work on the whole usability domain which including usability, accessibility, user experience (UX), and human factors. This due to that if all previous topics are considered, in the end it will find a huge data and many questions that will take more time and thus the study becomes too complex. For example, there is a needed to medical background, sensitive data about users and experts to obtain and analyse data for Web accessibility in terms of removing barriers that prevent interaction with, or access to websites, by people with disabilities. Also, it will not work on the too broad domain like human factor because it is umbrella term for numerous areas that has many areas related to human performance, technology, and design which they are out of the scope of the adaptive framework as mentioned in Section 1.5. Moreover, the step of the focus group or expert input (Step 3 in the adaptive framework) indeed can do work if it has data from previous steps which are user input (mini-user testing) and literature review. This is the reason that the expert input is the best to be in step three. Also, the second step of the user input or mini-user testing (Step 2 in the adaptive framework) offers more data about user experience that can be help to enrich the focus group session to formulate focused heuristics that are built based on the experiences of the real users and experts.

In addition, the results of this research found that the DSI method in both domains failed to discover some problems that were discovered by UT and HE. Thus, the lessons that have been learned about how best to apply the framework is to apply it in the dominant domains that have clear definitions and classifications and thus it is easy to scoping these domains. Also, the selected websites should be relevant to the subject of study. Furthermore, it is good to recruit more users in the context meeting and focus group to give more information and feedback, and thus formulating more focused heuristics. Moreover, it is recommended to increase the number of tested websites in the mini-user testing (four websites at least) which will give the participants the opportunity to evaluate more websites in the chosen domain that have different interface designs and features. This will help to uncover more problems and thus formulate focused heuristics from these problems.The five points below illustrate the adaptive framework steps;

1- Select clear and dominant domain. Then, select some websites that are relevant to the subject of study and in the scope of the chosen domain. Then, conduct extensive literature analysis to the related work on the selected domain. If there is limited literature or there is

no information, the content analysis with two experts is needed. This step helps to identify usability problem areas for the selected domain and to formulate specific heuristics and checklists for each usability problem area, as mentioned in the first step of the adaptive framework 'Development Step One' at Section 3.3.

2- Conduct mini-user testing with a number of real users for the selected domain. Before testing the context meeting should be arranged and, based on its results, the task scenarios should be identified and designed. Also, the real users are identified and recruited. The tasks should be designed to cover the main functionalities of the domain, and simpler tasks should be first and then the more difficult ones should be added. The structured problem report should be used. Pilot testing should be conducted on these tasks to make sure that all tasks can be performed. Also, the observation method should be used during the mini-user testing. Then, the mini-user testing is conducted. This step aims to learn from user error and behaviour to identify usability problem areas and also to formulate specific heuristics and to develop a checklist from the usability problems that are found by users, as mentioned in the second step in the adaptive framework 'Development Step Two' at Section 3.3.

3- Conduct a focus group with experts in the domain and in general usability to discuss issues arising from the above steps, and to incorporate the knowledge gained from experts to identify the usability problem areas and to formulate specific heuristics and checklists for each area. Reliability testing should be calculated to identify the final draft of usability problem areas between experts, as mentioned in the third step in the adaptive framework 'Development Step Three' at Section 3.3.

4- Combine the results achieved from each step. Thus, all discovered usability problems from step two, usability areas from step three, and the formulated heuristics from step two and three are listed. Then, the listed heuristics from step two and three are analysed and filtered. At that point, filtered heuristics are categorised according to appropriate usability areas. Consequently, a DSI method is established with explanations. After that, the results that have been collected from step one, the discovered usability problems from step two, and meeting report from step three are combined and analysed to produce a set of elements under each heuristics. Thus, to develop the DSI checklist. Finally, pilot test should be conducted on DSI and its checklist to make sure that all part can be performed, as

mentioned in the fourth step in the adaptive framework 'Development Step Four' at Section 3.3.

To sum up, this study has addressed the relative effectiveness of the three methods for evaluating user interfaces, and offers some insights into each (see Table 6.3). Overall, DSI, as applied here, produced good results; it found the greatest number of real problems, including more of the most serious ones, than either HE or UT, and at only a minimally higher cost. For more explanation, DSI method will be too expensive if it is used to evaluate one website only. This is because that the fixed cost and variable cost for one website will be much higher than the variable cost for the other methods. However, DSI's cost will be less when it is used to evaluate many websites this is due to that the fixed cost will be one only (there is no needed to repeat the steps of the adaptive framework every time) and the variable cost will be calculated based on the number of the evaluated websites. Also, more time and resources are needed for building a DSI method. It also needs an expert user to use it. HE failed to identify a large number of the most severe problems; however, it was quite effective at identifying cosmetic and minor problems. UT is the most expensive method, yet it failed to identify some severe problems; however, it helped to discover general problems and it assists, as does DSI, in defining the users' goals. Thus, these findings facilitate the decision-making process with regard to which method to employ, either on its own or in combination with another, in order to identify usability problems on the applicable websites. The selection of methods will depend on the types of problems identified by each of them. For convenience a please see Table 6.3 that summarizes these points.

Table 6.3: Summary of the study 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 * Usually conducted under lab conditions
Heuristics Evaluation (HE)	* Identifies some 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 evaluator performance * Identifies the real users' problems and helps define and achieve users' goals * The cost of the DSI including the fixed cost and variable cost will be cheaper than the variable cost of other methods, if many websites will be evaluated. * Developing DSI method that is characterized to be focused and involved experiences of the real users and experts.	* A little higher in cost than HE and cheaper than UT * Slightly higher in terms of time than HE * More resources are needed for building a DSI method (e.g. reviewing literature and recruiting users and experts). * The DSI method consumes more time for validation (e.g. time for conducting experiment, collecting data, analysing results, and reporting the results). * Needs an expert user to use it with knowledge in using content analysis, user testing, focus groups, interviews and observation methods. Also, the researcher needs the skills of designing tasks and questions and managing a meeting or a discussion group. * For evaluating one website only, the fixed cost of the DSI is much higher than the variable cost of other methods. Thus, this method is better to use in evaluating many websites. * The adaptive framework does not work in the broad domains such as whole topics of the science domain, it works on the dominant domains that are not too broad.

6.4 The identification of the sample size

Many companies struggle with limited budgets, therefore, usability experts recommend recruiting only five participants for usability testing, rather than the large sampling needed for experimental research; however, some experts are opposed to this figure. Thus, this has led to the establishment of many rules, as mentioned previously in Section 2.4.1 , such as the 10±2 rule, 20 users, and the 5 users for UT, and the 4±1 rule for HE (Hwang and Salvendy, 2010); (Nielsen and Molich, 1990). This research aims to readdress this issue and to quantify the sample size required based on empirical studies conducted on two different domains. In addition, it seeks to measure the effects that different sample sizes have on the number of usability problems found. In this research, for the first experiment, 20 users for each website

(total 60 users) were divided into two teams to investigate the 10±2 rule. The first team consists of the results of 8 users whereas the second team delivered the results of 12 users for the same group. This means that there were 6 teams overall; three teams consisting of 8 users and the other three teams consisting of 12 users for each website. Also, to investigate the 4±1 rule, the evaluator groups were divided into two teams with 4 evaluators in each team. In terms of the second experiment, the rule of 3 evaluators (including double and single), the rule of 20 users, and of 5 users (the 'magic number') were examined. Thus, 25 users were recruited for each website (totalling 75 users), which were divided into two teams. The first team delivered the results of 20 users whereas the second team delivered the results of 5 users for the same group. This means that there were 6 teams overall; three teams consisting of 20 users, and the other three teams consisting of 5 users for each website. Furthermore, to investigate the '3 evaluators' rule, the evaluator groups were divided into two teams, which included 3 evaluators in each team.

After analysing the results of both experiments, the UT's result in the first experiment were in line with Nielsen's claim that five users are enough to discover from 80% to 85% of usability problems. Also, the HE's result in the first experiment were in line with Nielsen's claim that four evaluators can discover 80% of usability problems. However, the results of the second experiment were completely different when the five users discovered only 37.6% of the total problems found. Also, the three evaluators discovered between 23% and 27% of usability problems. It is clear from the above results that the Nielsen's claim worked in the first experiment but it didn't work in the second experiments. Furthermore, this research found that the 16±4 rule of participants was valid in identifying over 90% of the usability problems in tested interfaces by using the UT method. The researcher here arrived at these results through conducting complex research experiments. Moreover, the figures in this study cannot be generalised to other domains because of the complexity and context of this study; it employed specific and highly targeted types of task, the websites were specifically chosen, and the differences between and among the users and evaluators in terms of their characteristics and knowledge may have been significant to these particular studies. Also, the participants were recruited from different cultures, which may imply that the interaction, communication and tested interfaces may be different for those from the same culture. This could explain the differences between the users' results. Table 6.4 shows the appropriate sample size purposes based on the findings of this study. It seems that there is no solid sample

size for finding all usability problems. Moreover, for studies where statistically significant findings are being sought, or for comparative studies, a group size of greater than or equal to 20 users is valid. This research strongly recommends considering the 20 users as the highest sample size and 12 users as the lowest sample size along with the study's complexity and the criticality of its context before commencing an evaluation study in order to achieve a successful evaluation. Furthermore, for the HE and DSI methods, this research recommends for HE that the 7±2 evaluators with mixed double and single evaluators are sufficient, and that 3 with mixed double and single evaluators are sufficient for the DSI method, as seen Table 6.5. In the recruitment of experts, one must consider that the number of evaluators and their expertise (double or single) can affect the results to a considerable degree, and probably more than the participant group size.

Table 6.4: Sample size estimation for various UT purposes

Main Purpose	# users
To find more cosmetic problems and problems relating to structure and content	5
To find fewer major and more minor problems. This is more appropriate for commercial studies and more problems in layout and formatting.	8
To find more catastrophic, major, minor and cosmetic problems; also, for finding more problems relating to design, navigation and the key aims and functions for which the system was designed. Moreover, it is more appropriate for comparative studies.	16±4
Appropriate for statistically significant studies and analysis of performance metrics, such as success rate.	≥ 20

Table 6.5: Sample size estimation for various HE and DSI purposes

Main Purpose	# experts
For HE, this sample with mixed double and single experts is sufficient to find 80% of usability problems, with applying user testing (UT) as a complementary method.	7±2
For DSI, this sample with mixed double and single experts is sufficient to find 90% of usability problems, without applying the user testing (UT) method.	3

In conclusion, it is challenging to determine the optimal sample size based on problem discovery or level of confidence and then to generalise this advice, as the result should be driven by the study's context. There is no solution to the challenge here. The above results provide evidence that the first viewpoint's affirmation, which states that a sample of 5 users will discover 80% of all usability problems (Nielsen, 2000b), is not likely to work on any experiment, whereas the 16 ±4 rule gains much validity for user testing. Table 6.6 shows the comparisons of five users' performances between different studies and those presented in this

study. Furthermore, this study also confirms the importance of involving the double expert evaluators in order to take advantage of their expertise in finding the main problems that might lead to product failure. The reality is that most usability methods will never discover all or most problems. Furthermore, even if all problems were identified, most of them would never be fixed because of their cost (Hornbæk and Frokjaer, 2004); (Hornbæk, 2010); (Hertzum, 2006). Consequently, there is no unique model for sample size estimation, as the sample size depends on the objective of each particular study, as mentioned in Table 6.4 and 6.5; hence, the group size should typically be increased along with the study's complexity and the criticality of its context. Care should be taken when seeking advice offered in the literature. Furthermore, it is appropriate to split the sample size into groups of users (the data can be analysed for each group); also, a study can be terminated in the early stages when its purpose has been achieved in order to save time and money.

Table 6. 6: Comparisons of five users' performances in different studies based upon (Alshamari, 2010)

Study	Results	Comments
(Nielsen, 2000b)	85%	Based on statistical formula
(Virzi, 1992)	80%	Claimed 3 users were enough to identify most severe problems
(Turner et al., 2006)	80% to 85%	3 to 5 users can detect most usability problems
(Bevan et al., 2003)	35%	Tested amongst 49 users on four websites
(Faulkner, 2003)	55%	5 users
(Molich et al., 2004)	75%	The top team was able to reveal this percentage
This research	15% to 37%	5 users for testing three social network websites

6.5 Recommendations

This research provides a set of recommendations based on its results. These recommendations are divided into two parts. The first part contains the recommendations regarding to the methods used in this research. The second part is a set of recommendations that highlights the specific types of usability problems for both domains and that were discovered using the three methods. This part gives suggestion on how the usability for the chosen websites could be improved. These are as follows:

6.5.1 Recommendations for methods used in this research

This research used three evaluation methods and each method has advantages and disadvantages that can help a development team and usability practitioners to choose which method they should use, which are as follows;

6.5.1.1 Heuristics Evaluation (HE)

The results of this method in terms of percentage of discovering the usability problems vary from 75% - 85% using four evaluators to 23% - 27% using three evaluators. This indicates that more expert evaluators should be recruited during the evaluation session to discover a high percentage of usability problems. In other words, to improve the efficiency of this method, the experience of number of expert evaluators should be incorporated to evaluate a chosen product. Another solution to improve its efficiency is for it to be complemented by the conducting of user testing because each one is complementary to the other. Moreover, the efficiency of this heuristics evaluation can be improved through the creation of well-designed and specific heuristics, as has been carried out in this research.

6.5.1.2 User Testing (UT)

In this research, UT is the second good method in terms of discovering a lot of usability problems. It was able to discover the percentage of problems ranging between of 37% to 80% using 5 and 8 users, respectively. Also, it was able to discover the percentage of problems ranging between of 92% to 98% using 12 and 20 users, respectively. This also indicates that more users should be recruited to discover a high percentage of usability problems. However, this makes this method more expensive.

6.5.1.3 Domain Specific Inspection (DSI)

In this research, DSI is good in terms of numbers of usability problems discovered and its efficiency and effectiveness. It was able discover the percentage of problems ranging between 98% to 99% using four evaluators and between 90% to 95% using three evaluators. This indicates that the recruitment of few expert evaluators during evaluation session can discover a high percentage of usability problems without applying the user testing method.

6.5.2 Specific types of usability problems found on the chosen domains

The below list of usability problem areas were sent by email to website owners as shown in Appendix Z. These usability problems are categorized into five problem areas according to the most common usability problems that were discovered in this research. The following is the explanation for these categories;

6.5.2.1 Navigation problems

This area is very important as the users can be confused or lost, and this leads them to leave the website. Consequently, designers should consider the links because there are seven main usability problems that can be faced when they are clicked. The first problem is misleading links, the second problem is broken links, the third problem is unclear link positioned, and the fourth problem is that pages do not have navigation links. The fifth problem is links which are not clickable. The sixth problem is that some sites do not provide and site map feature. The seventh problem is that some sites do not provide a breadcrumb to identify the path to the current location.

6.5.2.2 Content quality problems

This area is important as well. There are five main usability problems that can occur when users surf the website. The first problem is irrelevant content, whether this is in a particular page or throughout the site such as an advertisement or pornographic post. The second problem is a huge content on a page which makes a page is too long, or a lot of unnecessary required fields that make the process of such tasks is frustrating, such as on a registration page. The third problem is using unfamiliar terminology. The fourth problem is offering unapproachable content, such as unavailable videos or lessons. The fifth problem is unreadable content due to its small font.

6.5.2.3 Inconsistency and design usability problems

Both these areas are important. The most common inconsistency problem is that the navigation of top and bottom menu is not consistent throughout the site, thus there are two forms of problems here. Firstly, the location of the navigation menu is not positioned on the same place throughout the pages; secondly, some pages are without the navigation menu. Both of these forms lead to user confusion. Consequently, designers should consider these types of inconsistency problems. Regarding design usability problems, there are different forms of usability problem in this area. The first form is misleading images such as the logo of a homepage or any images throughout the site that do not have a link to the correct pages. The second form is non-clickable images when the users expect that they are clickable and that they will link users to other pages, or clickable Images which do not have a mouseover feature. The third form is inappropriate page design such as the use of a lot of links or advertisements on the pages, items not being logically grouped. Other sites do not use

minimal scrolling, do not use alternative text for explaining the images icons, contain inappropriate or an unattractive color scheme such as colors of background, menu and link. Some pages are without headings, or the colour of the selected item in the menu should be changed to another colour to give a clear indication of the current page is displayed. Also, the colour of the visited items should be changed to a different colour which shows the items that were recently visited. Finally, the required fields on the registration page are not identified or show the error messages without any indication to which field is missed.

6.5.2.4 earch quality problems

This area is very important. It has different forms of problems. The first form is that some sites do not provide an advanced search feature. The second form is that the search results are not as accurate as was expected. The third problem is that the search button and search input field are not placed across all pages and they are not clearly positioned on the top of the homepage.

6.5.2.5 Help Center problems

This area too is important. It has different forms of problems. The first problem is that some sites do not provide an FAQ feature. The second problem is that there is not enough information on the FAQ page or 'Help Center' page. The third problem is that there are many helping forums for each product on the same website such as Google plus website. The fourth problem is that some sites do not provide a 'Contact Us' link or it is not clearly positioned on the homepage and throughout all pages.

6.6 Conclusion

This chapter has discussed the results achieved from Chapter 5 and has compared their results in the light of the previous literature. Furthermore, it has illustrated how the research question of this research was answered. The above results show the effectiveness of the three methods used in this research in terms of time spent, number of usability problems and their severity, UEM metrics, and specific types of usability problem areas. This research facilitates decision making regarding the most appropriate method to use (i.e. UT, HE, DSI, or HE and UT together). The next chapter will discuss how the aims and objectives of this research have been addressed. Also, the research contributions will be outlined.

Chapter 7: Conclusions

7.1 Introduction

The previous chapters (3, 4, 5, and 6) presented the adaptive framework, methodology and design, preparations, procedures, collection, analysis, results and findings for two experiments, aimed at validating the adaptive framework via its generated method (DSI). Chapter 6 also included the discussions and comparisons between the results and findings achieved in this research with those of previous studies, and answering the research question. This chapter discusses the conclusions and contributions of this study to the field of UEMs. It also presents the study limitations and offers recommendations for future research.

7.2 Achieving the objectives

The aim of this research was the construction of an adaptive methodological framework that would be readily capable of adaptation to any domain. This was then evaluated by generating an evaluation method for assessing the usability of products in a particular domain (educational and social network domains). The evaluation method under study is the Domain Specific Inspection method (DSI); it is empirically, analytically and statistically tested by applying it on the two aforementioned domains. In addition, it aimed to explore the effect of sample size on the usability evaluation and to quantify the number of evaluators and users required in order to achieve the good evaluation results for the DSI, HE and UT methods. This aim will be achieved through meeting the eight objectives mentioned in Chapter 1, and has been successfully achieved through conducting this research, as is clarified in the following sections.

7.2.1 Objective One: Review the current issues in usability evaluation methods on dynamic websites

This research began with a thorough investigation of the Literature Review, which provided an integrated analysis of the studies of pioneers and other prominent individuals within the field of usability evaluation methods (UEMs). Such pioneers include Jakob Nielsen, Sherry Chen, Robert Macredie, Gitte Lindgaard, Jarinee Chattratichart, Jeff Sauro and Kasper Hornbæk. Other valuable contributors have not been overlooked however, and are included in the study to offer a wider understanding of this field. Therefore, their studies and contributions have facilitated building a solid foundation for conducting this research. Thus, the first objective was met on the basis of the extensive Literature Review undertaken in this study.

7.2.2 Objective Two: Construct the adaptive framework

Complementing the first objective, Chapter 2 identified that many published works have been conducted to enhance the effectiveness of UEMs. HE has been revised and extended for universal and commercial websites, such as HE-Plus and HE++, and HOMERUN heuristics (Nielsen, 2000a, Chattratichart and Lindgaard, 2008, Chattratichart and Brodie, 2002); however, some researchers found that their tested websites failed in certain respects according to these extended or modified heuristics (Alrobai et al., 2013); (Thompson and Kemp, 2009). Consequently, researchers sought to compare and contrast the efficiency of HE with other methods, such as UT, during which they found that HE discovered approximately three times more problems than UT; however, they also reported that more severe problems were discovered through UT when compared with HE (Doubleday et al., 1997); (Jeffries et al., 1991); (Liljegren and Osvalder, 2004); (Thyvalikakath et al., 2009). More recently, researchers' findings have been almost unanimous in one respect: HE is not readily applicable to many new domains with different goals and is too vague for evaluating new products, such as web products, as it was originally designed to evaluate screen-based products. Furthermore, it was developed several years before the web was involved in user interface design (Hart et al., 2008); (Hasan, 2009); (Ling and Salvendy, 2005a). Therefore, as each method appears to overcome the other method's limitations, researchers now recommend conducting UT together with HE, as they complement each other; combining the

two methods offers a superior picture of a targeted website's level of usability (Law and Hvannberg, 2002); (Nielsen, 1992a). In addition, 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, as they were developed to deal with certain aspects of usability in certain areas (Coursaris and Kim, 2011).

Having extensively reviewed the existing literature, and, based on the researcher's knowledge, 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 methods and usability assessment process, represents a missing area in evaluation research and practice. Thus, Chapter 3 explained and justified the components of the adaptive framework and the components for testing the adaptive framework was explained in Chapter 4 (Section 4.4.9) in detail. Accordingly, the second objective was achieved.

7.2.3 Objective Three: Test the practicality and the efficiency of the adaptive framework

This objective was initiated from the desire to develop a method that is context specific, productive, useful, usable, reliable and valid. In this research, this framework was used to evaluate the educational and social network domains. Therefore, this objective was achieved through following the steps of the adaptive framework, as mentioned in Sections 5.5, and produced a DSI for each domain, as mentioned in Appendix L4 and Q4.

7.2.4 Objective Four: Validate the outcomes of the adaptive framework

After generating the DSI methods for the educational and social network domains, the targeted websites in each domain were selected. Subsequently, the components for testing the adaptive framework were applied, which entailed an analytical assessment of the DSI method through empirical and statistical processes. This was achieved in each of the two domains through applying the three UEMS (HE, UT and DSI) on three websites in each domain. Chapters 5 explained the results of both experiments in detail. Thus, this objective was achieved.

7.2.5 Objective Five: Identify the usability problem areas for the educational and social network domains

This objective was a continuation to Objective Two; the data collected through Step One, Step Two and Step Three in the adaptive framework were analysed separately in both experiments. Subsequently, Cohen's kappa coefficient was used twice on the same focus group to enable a calculation of the reliability quotient for identifying usability problem areas. After that, the key areas of the usability problems achieved from each step were identified. Thus, five areas were identified for the educational domain, and seven areas were identified for the social network domain. Furthermore, in order to facilitate the process of evaluation and analysis, and to help designers and programmers identify the areas in their websites that needed improvement, the DSI methods and their checklists were established, closely focused on each domain and classified according to the usability problem areas detailed in Appendix M2 and R2. Chapters 5 summarise how the usability problems were identified and classified according to these areas, and how these areas facilitated analysis and reporting of the findings. One expert commented that ''many of the problems could be missed but the usability problem areas helped to rectify this matter". Another expert commented that'' the usability problem areas facilitated to identify the overlapping and made the comparisons between problems easy''. Thus, this objective was achieved.

7.2.6 Objective Six: Explore the effect of sample size on the usability evaluation and identify the sample size of for good evaluation results for DSI, HE and UT methods

The Literature Review Chapter reveals that many rules were established to determine the appropriate sample size for UEMs, such as the 10±2 rule, 20 users, and 5 users for UT, and also the 4±1 rule for HE (Hwang and Salvendy, 2010); (Nielsen and Molich, 1990). This issue was examined through conducting the two experiments. The correlations between the number of users and the usability problems found in each experiment were explored, as mentioned in Chapter 5. Chapter 6 outlines the results achieved in this study, revealing that 7±2 evaluators with mixed double and single evaluators are sufficient for HE, and that 3 with mixed double and single evaluators are sufficient for DSI. For UT, 16±4 participants are a valid number in discovering over 90% of the usability problems in tested interfaces. Thus, Objective Six was achieved.

7.2.7 Objective Seven: Explore further the correlations among UEM measures in this study

Examining the correlation among the UEM measures statistically assisted in understanding this study. The Pearson Correlation test was employed for this purpose, which led to identifying a number of significant relationships. Chapters 5 investigated the correlations among the UEM measures in both experiments. This study confirms statistically that a relationship between the number of usability problems found and time spent by the users and evaluators exists. This means that when participants spend more time, they will discover more problems. Thus, this objective was met.

7.2.8 Objective Eight: Propose a set of recommendations and suggestions in order to improve the usability of the chosen domains

This objective was achieved, as can be seen in Chapters 5 and 6, where a number of findings and recommendations were discussed in detail, which should be considered to increase the overall effectiveness of UEMs and to improve the usability for tested websites. For example, it appears that there is no usability evaluation method able to discover all usability problems in the targeted products. The research findings conclude that adding more users does not necessarily mean that more problems will be discovered. Also, the study found that it is not essential to conduct HE and UT methods and combine their results, as their results complement each other (Nielsen, 1992a), specifically in methods such as DSI, which were developed from the adaptive framework in this research. Furthermore, the results of usability evaluation and testing may be influenced by certain factors, such as sample size, evaluator characteristics, type of method used, targeted product, type and complexity of task designs, and types of think aloud used. Additionally, the research findings revealed that user satisfaction is not influenced by the number of usability problems and time spent. The design and services provided by targeted products can influence user satisfaction, even if they face a number of challenges when performing tasks. However, there is a relationship between evaluator satisfaction and method used (HE or DSI); the evaluators who used DSI delivered a much higher score compared to the evaluators who used HE, which delivered a lower overall score in both experiments. Moreover, the final report of usability problems that

included the results of the three methods was sent to each website to aid in understanding these problems and, thus, improving their websites. The response was received from BBC and LinkedIn websites, as shown in Appendix U1 and Appendix U2. Also, all of these websites have been changed and their usability has been improved after sending the usability report as it described this in detail in the section 5.6.3.3. Moreover, Section 6.5 provides a set of recommendation based on the research result.

7.3 Research contributions

This research offers a number of contributions that can be divided into three categories: Paractical contributions, Theoretical contributions, and Publications and personal outcomes.

7.3.1 Practical contributions

The main practical contribution in this study the creation and testing of the adaptive framework as following;

- In relation to the creation of the adaptive framework, this research has generated DSI which were specific for evaluating the educational and social network domains. Also, it has developed the checklists from the DSI methods as a tool that affords designers, developers, instructors, evaluators, and website owners the facility to design an interactive interface or assess the quality of existing websites. Furthermore, it has identified the usability problem areas in the educational domain (five areas) and the social network domain (seven areas). These areas provide designers and developers with insights into how interfaces can be designed to be more effective, efficient and satisfying; they also support a more uniform problem description and can guide expert evaluators in finding real usability problems, thereby, facilitating the evaluation process by assessing each area and page in the target product; it also allows anyone to adopt any area of usability or any principle to determine the usability problems related to the five or seven specific areas in the educational and social network domains.

- In relation to the testing the adaptive framework, this research compares the effectiveness of the different UEMs, namely, HE, DSI and UT, against a set of measures, and determines which method is the most appropriate for evaluating each usability problem area. Furthermore, it examines the relationships between these methods and the set of usability

measures. Moreover, it examines the impact of sample size on the findings of usability tests, and it determines the appropriate sample size for the domain specific context inspection (DSI) method, HE, and UT through empirical studies in the social network and educational domains.

- In relation to the research problem, this research offers systematic procedures to develop a framework to solve a particular problem through six steps. These procedures start with defining the efficiency problems in usability evaluation methods (UEMs), and this was explained in detail in Chapter One, Section 1.4. This is followed by defining the requirements and features that should be considered in the new method. This can be done by understanding the current issues and challenges inherent in the development of a new method, and this was explained in detail in Chapter Two, Section 2.4. The third step is to define the appropriate strategy to develop the new method, and this was achieved in Chapter Four through a review of the available methodologies and theories to find appropriate methods for supporting the development of the new method. The fourth step is to construct a framework to solve the problem and execute the proposed solution, and this was achieved through the construction of the adaptive framework to generate a domain specific inspection method (DSI) as described in Chapter Three. The fifth step is to validate practically the proposed framework to ensure that it achieves the goals that were established for it. This was achieved by proposing the components for testing the adaptive framework in Section 4.4.8, using the research question Section 1.8, and conducting two exploratory experiments as described in detail in Chapter 5. The sixth step is to present the proposed framework and its generated method in different journals, workshops and conferences for sharing knowledge with experts in this domain, and this was achieved by publishing eight papers, as shown in Appendix W.

7.3.2 Theoretical contributions

This research has significantly helped to develop the researcher's knowledge in the field of UEMs. This research contributes to the advancement of knowledge in the HCI field by:

1. Understanding of the comparative value of generic and domain-specific usability evaluation methods. For example, this research finds that the domain-specific inspection (DSI) method discovers more unique problems than the generic method (HE). Also, the

DSI performs better in terms of identifying catastrophic and major problems than the generic method (HE). Furthermore, the number of problems identified by each evaluator who used the HE method is always less than the number of problems identified by any evaluator using the DSI method. However, the generic method (HE) consumes less time and is cheaper than the DSI.

2. Further investigation into the effect of sample size and evaluator types on the test result. This research confirms that there is no specific model for sample size estimation, as the sample size is likely depends on the objective and complexity of each particular study. Also, the effect of recruiting different evaluator types has been confirmed in this study.

7.3.3 Publications and personal outcomes

During the period of this research, the researcher has published several conference and journal papers (16 published papers) and has taken part in various related scientific activities. Briefly, two conference papers and three journal papers from the first experiment on the educational domain were published. Furthermore, one conference paper and two journal papers have been published from the second experiment on the social network domain. In addition, one conference paper has been published from the results of both experiments regarding the same size of usability methods. The researcher was also involved in the production of eight papers in the field of usability methods addressing different aims. Consequently, these papers have increased the researcher's total number of publications and they have offered scientific knowledge to the researcher to provide assistance in research, as shown in Appendix W.

The researcher conducted two comprehensive experiments to evaluate six websites by using three usability methods, recruiting 135 users and 14 expert evaluators for the experiments, 5 users for each context meeting session (10 users in total), 10 users for each mini-user testings (20 users in total) , 5 expert evaluators for each focus group sessions (10 in total), 30 users for pilot studies, 4 expert evaluators for analysing the results, 5 users for each falsification test (10 in total). Consequently, the researcher’s skill regarding time management and communication was developed, as the users and evaluators were gathered from different cultures, speaking a range of languages. Furthermore, conducting a series of experiments requires targeted planning, preparation, implementation, analysis, and result reporting, which

are more complicated if more than one method is used. This research used HE and DSI as the two methods classified under the inspection method, and UT classified under the testing method. Additionally, the skill of presentation using PowerPoint software was developed when the researcher attended two Doctoral Consortiums and poster design. Regarding the Doctoral Consortiums, the researcher attended the HCI 2012 in Birmingham and the BCS Doctoral Consortium 2012 in London as shown in Appendix X. Regarding the posters, the researcher presented one poster during the 26th Annual Conference of the Specialist HCI group of the BCS, and the second was presented during a research day conducted at the School of Computing Science 2014 at the University of East Anglia (Alroobaea, 2013). These activities offered a cooperative forum for the researcher to learn how to do research and to discuss his work and receive constructive feedback. In addition, they offered relevant information on issues important to doctoral candidates and nurtured the community of researchers. Moreover, the researcher is now able to organise a conference based on the experience that has gained from volunteerism as shown in Appendix Y.

7.4 Limitations and future Research

As in any research, this study has a number of limitations; however, the researcher views these not as weaknesses but as opportunities for further research. As such, these possibilities can be divided into four key areas:

1. One measurement, namely, the Redesigning Step, could be added to the components for testing the adaptive framework to measure the usefulness of DSI. This step requires further investigation by testing the adaptive framework after applying this measurement to measure the efficiency of UT, HE and DSI in producing useful usability problems. It aims to redesign the tested products based on the problem report and to re-evaluate them in order to measure their usability and improvement; ultimately, to measure the efficiency of the adaptive framework and its generated method (DSI). Hornbæk (2010) pointed out that "the true utility of methods lies in their ability to influence the design of the application being evaluated". Thus, the evaluation in the context of design leads to more realistic results for UEMs and more connection to practical usability work (Hornbæk, 2010); (Wixon, 2003).

2. This research claimed that this framework is applicable for any product. The educational and social networks websites were chosen; however, it is unclear about other domains such as e-commerce websites, which require further research.

3. This research conducted two experiments and analysed their results. The study results found that the 16 ± 4 rule gained much validity for UT. For HE, the 7 rule with mixed double and single evaluators proved sufficient; also, 3 with mixed double and single evaluators were deemed sufficient for the DSI method. Consequently, these findings require diverse tests to be performed, and then the data compared to those processed here in order to verify these findings more conclusively.

4. User preference of using the adaptive framework to generate a DSI method needs further research. This aims to measure the helpfulness attribute in order to identify to what extent the user finds using the adaptive framework is more helpful in terms of generating DSI for a chosen product. Furthermore, the DSI checklists were used by experts in both experiments, what about the normal users? This needs further research in the future.

5. To improve the quality of the DSI method, the results of the validation process can be employed to analyse in depth the unique problems discovered by UT and HE.This permits an investigation into why these problems were not identified when DSI was used. If it is found that the heuristics are not accurately specified or that there are no proper heuristics to reveal these problems, then the unspecified heuristics can be properly specified or new heuristics can be added.

6. Determining sample size and its statistical power provides a different method to carry out empirical studies. This method employs significance testing and is essential for the planning of studies, for the interpretation of study results, and for the validity of study conclusions. As part of my future interests in continuing to research in this area, determining sample size and testing for its statistical power will be strongly considered.

7.5 Concluding remarks

Contrary to the majority of efforts made to construct and test enhanced usability methods, this research has made more explicit the process for doing so. The adaptive framework includes the views of users and usability experts to help generate a context-specific method for evaluating any chosen domain. The findings presented here illustrate and evaluate this process for the generation of the DSI method to assess the usability of educational and social network websites. DSI outperformed both HE and UT, even when taken together. This clearly represents a progressive step forward. The process for construction of a DSI based on the adaptive framework can take considerable resources, but once the DSI constructed, it is relatively fast, productive and cheap to use.

References

Abuzaid, R. A. S. 2010. Bridging the Gap between the E-Learning Environment and E-Resources: A case study in Saudi Arabia. *Procedia-Social and Behavioral Sciences,* 2**,** 1270-1275, 10.1016/j.sbspro.2010.03.186.

AcademicEarth. 2012. *AcademicEarth* [Online]. Available: http://academicearth.org/ [Accessed 3/4/2012.

Al-Badi, A. H. 2014. The adoption of social media in government agencies: Gulf Cooperation Council case study. *Journal of Technology Research,* 5**,** 1-26.

Al-Badi, A., Okam, M., Alroobaea, R. & Mayhew, P. 2013. Improving Usability of Social Networking Systems: A Case Study of LinkedIn. *Journal of Internet Social Networking & Virtual Communities,* 2013**,** 23, 10.5171/2013.889433.

Al-Razgan, M. S., Al-Khalifa, H. S., Al-Shahrani, M. D. & AlAjmi, H. H. 2012. Touch-Based mobile phone interface guidelines and design recommendations for elderly people: a survey of the literature. Neural Information Processing. Springer, 568-574, 10.1007/978-3-642-34478-7_69.

Alam, T. & Ali, M. 2010. *The Challenge of Usability Evaluation of Online Social Networks with a Focus on Facebook.* Master, Blekinge Institute of Technology.

Albert, W. & Tullis, T. 2013. *Measuring the user experience: collecting, analyzing, and presenting usability metrics,* Morgan Kaufmann.

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,* 1.

Alkhattabi, M., Neagu, D. & Cullen, A. 2010. Information quality framework for e-learning systems. *Knowledge Management & E-Learning: An International Journal (KM&EL),* 2**,** 340-362.

Alleva, E. & Branchi, I. 2011. An evolutionary perspective for contemporary psychiatric research. *Rivista di psichiatria,* 46**,** 288-291, 10.1708/1009.10973.

Alrobai, A. A., Alroobaea, R. S., Al-Badi, A. H. & Mayhew, P. J. 2013. Investigating the Usability of E-Catalogues Systems: Modified Heuristics vs. User Testing. *User Testing (July 4, 2013),* 4**,** 1-29, 10.2139/ssrn.2416586.

Alroobaea, R. 2013. *School of Computing Sciences,. Postgraduate Research Day 2013: Book of Abstracts* [Online]. University of East Anglia. Available: https://www.uea.ac.uk/documents/429378/1126774/abstract-bookletvs.pdf/54260e9a-01f7-48d7-a953-e6ee47e267ba.

Alshamari, M. 2010. *Task Formulation in Usability Testing.* Doctor of Philosophy, University of East Anglia.

Alsumait, A. & Al-Osaimi, A. 2009. Usability heuristics evaluation for child e-learning applications. Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services. ACM, 425-430, 10.1145/1806338.1806417.

Andre, T. S., Rex Hartson, H., Belz, S. M. & McCreary, F. A. 2001. The user action framework: a reliable foundation for usability engineering support tools. *International Journal of Human-Computer Studies,* 54**,** 107-136, 10.1006/ijhc.2000.0441.

Andreasen, M. S., Nielsen, H. V., Schrøder, S. O. & Stage, J. 2007. What happened to remote usability testing?: an empirical study of three methods. Proceedings of the SIGCHI conference on Human factors in computing systems, New York. ACM, 1405-1414, 10.1145/1240624.1240838.

Antill, L. 1985. Selection of a research method. *Research methods in information system.* Elsevier Science Publishers,194-204.

Ardito, C., Costabile, M. F., De Angeli, A. & Lanzilotti, R. 2006. Systematic evaluation of e-learning systems: an experimental validation. Proceedings of the 4th Nordic conference

on Human-computer interaction: changing roles. ACM, 195-202, 10.1145/1182475.1182496.

Astani, M. & Elhindi, M. 2008. An empirical study of university websites. *Issues in Information Systems,* 9**,** 460-465.

Bahiss, K., Cunningham, S. J. & Smith, T. 2010. Investigating the usability of social networking sites for teenagers with autism. Proceedings of the 11th International Conference of the NZ Chapter of the ACM Special Interest Group on Human-Computer Interaction. ACM, 5-8, 10.1145/1832838.1832840.

Bangor, A. 2009. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies,* 4**,** 114-123.

Bangor, A., Kortum, P. T. & Miller, J. T. 2008. An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction,* 24**,** 574-594, 10.1080/10447310802205776.

Benbasat, I., Goldstein, D. K. & Mead, M. 1987. The case research strategy in studies of information systems. *MIS quarterly,* 11**,** 369-386, 10.2307/248684.

Bennett, M. 2004. A review of the literature on the benefits and drawbacks of participatory action research. *First Peoples Child & Family Review,* 1**,** 19-32.

Bengtsson, M. 2016. How to plan and perform a qualitative study using content analysis. *NursingPlus Open,* 2**,** 8-14.

Berg, K. E. & Latin, R. W. 2008. *Essentials of research methods in health, physical education, exercise science, and recreation,* Philadelphia, Lippincott Williams & Wilkins.

Bernérus, A. & Zhang, J. 2010b. *A Peek at the Position of Pedagogical Aspects in Usability Evaluation of E-learning System-A Literature Review of Usability Evaluation of E-learning System conducted since 2000.* Bachelor University of Gothenburg.

Bernsen, N. O. & Dybkjær, L. 2009. *Multimodal usability,* Denmark, Springer-Verlag New York, LLC, 10.1007/978-1-84882-553-6.

Bevan, N. 1995. *Measuring usability as quality of use,* UK, Kluwer Academic Publishers, 10.1007/BF00402715.

Bevan, N. & Azuma, M. 1997. Quality in use: Incorporating human factors into the software engineering lifecycle. Software Engineering Standards Symposium and Forum, Emerging International Standards. ISESS 97., Third IEEE International, Walnut Creek, CA. IEEE, 169-179, 10.1109/SESS.1997.595963.

Bevan, N., Barnum, C., Cockton, G., Nielsen, J., Spool, J. & Wixon, D. 2003. The magic number 5: is it enough for web testing? CHI'03 extended abstracts on Human factors in computing systems, New York, NY, USA. ACM, 698-699, 10.1145/765891.765936.

Beynon-Davies, P. 2002. *Information systems: An introduction to informatics in organisations,* Basingstoke, Palgrave Basingstoke, 10.1007/978-1-349-14931-5_2.

Bhattacharyya, E., Patil, A. & Sargunan, R. A. 2010. Methodology in Seeking Stakeholder Perceptions of Effective Technical Oral Presentations: An Exploratory Pilot Study. *Qualitative Report,* 15**,** 1549-1568.

Bias, R. G. 1994. The pluralistic usability walkthrough: coordinated empathies. *Usability inspection methods***,** 63-76.

Blandford, A. & Green, T. R. 2008. Methodological development. *In:* CAIRNS, P. A. A. C., A.L., (ed.) *Research Methods for Human Computer Interaction.* Cambridge, UK: Cambridge University Press, 158-174.

Blomkvist, J. & Holmlid, S. 2011. Existing Prototyping Perspectives: Considerations for Service Design. *Nordic Design Research Conference 2011.* Helsinki: Nordes.

Brady, K. P., Holcomb, L. B. & Smith, B. V. 2010. The use of alternative social networking sites in higher educational settings: A case study of the e-learning benefits of Ning in education. *Journal of Interactive Online Learning,* 9**,** 151-170.

Braun, V. & Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative research in psychology,* 3**,** 77-101.

Brinck, T., Gergle, D. & Wood, S. D. 2001. *Usability for the Web: designing Web sites that work,* francesco, USA, Morgan Kaufmann, 10.1.55860-658-0.

Brooke, J. 1996. *SUS-A quick and dirty usability scale,* UK, Taylor & Francis ltd.

Brown, M. A. 2009. *Developing Usability Heuristics for Computer Game Design.* PhD, National University of Ireland.

Bryman, A. 1992. Quantitative and qualitative research: further reflections on their integration. *Mixing methods: Quantitative and qualitative research*,57-78.

Bryman, A. 2012. Social research methods, New York, Oxford university press, 10987654321.

Cacioppo, J. T., Semin, G. R. & Berntson, G. G. 2004. Realism, instrumentalism, and scientific symbiosis: psychological theory as a search for truth and the discovery of solutions. *American Psychologist,* 59**,** 214, 10.1037/0003-066X.59.4.214.

Cairns, P. 2007. HCI... not as it should be: inferential statistics in HCI research. Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI... but not as we know it-Volume 1, British Computer Society Swinton, UK,. British Computer Society, 195-201.

Cairns, P. & Cox, A. L. 2008. *Research methods for human-computer interaction,* UK, Cambridge University Press.

Capra, M. G. 2006. *Usability problem description and the evaluator effect in usability testing.* PhD, Virginia Polytechnic Institute and State University.

Card, S. K., Newell, A. & Moran, T. P. 1983. *The psychology of human-computer interaction*, L. Erlbaum Associates Inc. Hillsdale, NJ, USA.

Carroll, J. M. 2003. *HCI models, theories, and frameworks: Toward a multidisciplinary science,* USA, Morgan Kaufmann.

Carroll, J. M., Singley, M. K. & Rosson, M. B. 1992. Integrating theory development with design evaluation. *Behaviour & Information Technology,* 11**,** 247-255, 10.1080/01449299208924345.

Castillo, J. C., Hartson, H. R. & Hix, D. 1998. Remote usability evaluation: can users report their own critical incidents? CHI 98 Cconference Summary on Human Factors in Computing Systems, 1998-04-01, New York, NY, USA. ACM, 253-254, 10.1145/286498.286736.

Charness, G., Gneezy, U. & Kuhn, M. A. 2012. Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization,* 81**,** 1-8, 10.1016/j.jebo.2011.08.009.

Chattratichart, J. & Brodie, J. 2002. Extending the heuristic evaluation method through contextualisation. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. SAGE Publications, 641-645, 10.1177/154193120204600509.

Chattratichart, J. & Brodie, J. 2004. Applying user testing data to UEM performance metrics. CHI'04 extended abstracts on Human factors in computing systems, 2004-04-24, Austria. ACM, 1119-1122, 10.1145/985921.986003.

Chattratichart, J. & Lindgaard, G. 2008. A comparative evaluation of heuristic-based usability inspection methods. CHI'08 extended abstracts on Human factors in computing systems, 2008-04-05, Italy ACM, 2213-2220, 10.1145/1358628.1358654.

Chen, F. 2006. *Designing human interface in speech technology,* USA, Springer Science & Business Media.

Chen, S. Y. & Macredie, R. D. 2005. The assessment of usability of electronic shopping: A heuristic evaluation. *International Journal of Information Management,* 25**,** 516-532.

Chen, W. & Hirschheim, R. 2004. A paradigmatic and methodological examination of information systems research from 1991 to 2001. *Information Systems Journal,* 14**,** 197-235, 10.1111/j.1365-2575.2004.00173.x.

Chinthakayala, K. C., Zhao, C., Kong, J. & Zhang, K. 2013. A comparative study of three social networking websites. *World Wide Web,* 17**,** 1233-1259, 10.1007/s11280-013-0222-8.

Chisnell, D. E., Redish, J. C. G. & Lee, A. 2006. New heuristics for understanding older adults as web users. *Technical Communication,* 53**,** 39-59.

Cho, N. & Park, S. 2001. Development of electronic commerce user-consumer satisfaction index (ECUSI) for Internet shopping. *Industrial Management & Data Systems,* 101**,** 400-406, 10.1108/EUM0000000006170.

Cockton, G. 2007. Make evaluation poverty history. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2007-04-29, USA ACM, 10.1145/1240624.2180964.

Cockton, G. & Woolrych, A. 2001. Understanding inspection methods: lessons from an assessment of heuristic evaluation. *People and Computers XV—Interaction without Frontiers.* London: Springer, 171-191, 10.1007/978-1-4471-0353-0_11.

Cockton, G. & Woolrych, A. 2002. Sale must end: should discount methods be cleared off HCI's shelves? *interactions,* 9**,** 13-18, 10.1145/566981.566990.

Cockton, G., Woolrych, A., Hall, L. & Hindmarch, M. 2004a. Changing analysts' tunes: the surprising impact of a new instrument for usability inspection method assessment. *People and Computers XVII—Designing for Society.* London: Springer, 145-161, 10.1007/978-1-4471-3754-2_9.

Cockton, G., Woolrych, A. & Hindmarch, M. 2004b. Reconditioned merchandise: extended structured report formats in usability inspection. CHI'04 extended abstracts on Human factors in computing systems, 2004-04-24, Austria ACM, 1433-1436, 10.1145/985921.986083.

Collins, J. & Hussey, R. 2013. *Business Research: a practical guide for undergraduate and postgraduate students,* UK, Palgrave Macmillan.

Connell, I. W. & Hammond, N. V. 1999. Comparing usability evaluation principles with heuristics: problem instances vs. problem types. Proceedings of INTERACT\'99-Human Computer Interaction. citeulike.org, 621-629, 0 9673355 0 7.

Corti, L., Day, A. & Backhouse, G. 2000. Confidentiality and informed consent: Issues for consideration in the preservation of and provision of access to qualitative data archives. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research,* 1.

CosmoLearning. 2012. *CosmoLearning* [Online]. Available: http://cosmolearning.org/ [Accessed 23/01/2016.

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

Cook, D. A. & Dupras, D. M. 2004. A Practical Guide To Developing Effective Web-based Learning. *Journal of general internal medicine,* 19**,** 698-707, 10.1111/j.1525-1497.2004.30029.x.

Creswell, J. W. & Clark, V. L. P. 2011. Designing and conducting mixed methods research. *Australian and New Zealand Journal of Public Health,* 31**,** 388–389, 10.1111/j.1753-6405.2007.00097.x.

Crowe, S., Cresswell, K., Robertson, A., Huby, G., Avery, A. & Sheikh, A. 2011. The case study approach. *BMC medical research methodology,* 11**,** 100, 10.1186/1471-2288-11-100.

De Vaus, D. A. & de Vaus, D. 2001. *Research design in social research,* SAGE Publications Ltd.

Delone, W. H. 2003. The DeLone and McLean model of information systems success: a ten-year update. *Journal of management information systems,* 19**,** 9-30, 10.1080/07421222.2003.11045748.

Denscombe, M. 2010. *The Good Research Guide: For Small-Scale Social Research Projects: For small-scale social research projects,* England, McGraw-Hill International.

Desurvire, H., Kondziela, J. & Atwood, M. 1992a. What is Gained and Lost when Using Evaluation Methods Other than Empirical Testing Practical Evaluation Methods for Improving a

Prototype. Proceedings of the HCI'92 Conference on People and Computers VII 1992 p. 89, UK. ACM

Desurvire, H., Kondziela, J. & Atwood, M. E. 1992b. What is gained and lost when using methods other than empirical testing. Posters and short talks of the 1992 SIGCHI conference on Human factors in computing systems. ACM, 125-126.

Di Lucca, G. A. & Fasolino, A. R. 2006. Testing Web-based applications: The state of the art and future trends. *Information and Software Technology,* 48**,** 1172-1186, 10.1016/j.infsof.2006.06.006.

Dix, A. 2009. *Human-computer interaction,* Pearson Education.

Doubleday, A., Ryan, M., Springett, M. & Sutcliffe, A. 1997. A comparison of usability techniques for evaluating design. Proceedings of the 2nd conference on Designing interactive systems: processes, practices, methods, and techniques, 1997-08-01, New York, NY, US. ACM, 101-110, 10.1145/263552.263583.

Downe‐Wamboldt, B. 1992. Content analysis: method, applications, and issues. Health care for women international, 13, 313-321.

Dray, S. & Siegel, D. 2004. Remote possibilities?: international usability testing at a distance. *interactions,* 11**,** 10-17, 10.1145/971258.971264.

Du Toit, M. & Bothma, C. 2009. Evaluating the usability of an academic marketing department's website from a marketing student's perspective. *International Retail and Marketing Review,* 5**,** 25-37.

Dumas, J. S. & Loring, B. A. 2008. *Moderating usability tests: Principles and practices for interacting,* Morgan Kaufmann.

Dumas, J. S. & Redish, J. 1999. *A practical guide to usability testing,* USA, Intellect Ltd.

Dwyer, C., Hiltz, S. & Passerini, K. 2007. Trust and privacy concern within social networking sites: A comparison of Facebook and MySpace. Americas Conference on Information Systems (AMCIS) Proceedings, 12-31-2007, Colorado. AIS Electronic Library (AISeL), 339.

Dykstra, D. J. 1993. *A Comparison of Heuristic Evaluation and Usability Testing: The Efficacy of a Domain-Specific Heuristic Checklist,* USA, A & M University, Texas.

Easterby-Smith, M., Thorpe, R. & Jackson, P. 2012. *Management research,* London, SAGE Publication Ltd, 10.4135/9781412950589.n521.

Ebling, M. R. & John, B. E. 2000. On the contributions of different empirical data in usability testing. Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques, 2000-08-01, USA ACM, 289-296, 10.1145/347642.347766.

Elliott, R. & Timulak, L. 2005. *Descriptive and interpretive approaches to qualitative research,* UK, OXFORD UNIVWESITY PRESS.

Ellison, N. B. 2007. Social network sites: Definition, history, and scholarship. *Journal of Computer‐Mediated Communication,* 13**,** 210-230, 10.1111/j.1083-6101.2007.00393.x.

Estes, J., Schade, A. & Nielsen, J. 2009. *Social Media User Experience* [Online]. Nielsen Norman Group. Available: http://www.nngroup.com/reports/social-media-user-experience/ 2012].

Faulkner, L. 2003. Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments, & Computers,* 35**,** 379-383, 10.3758/BF03195514.

Federoff, M. A. 2002. *Heuristics and usability guidelines for the creation and evaluation of fun in video games.* Master, Indiana University.

Felder, R. M. & Silverman, L. K. 1988. Learning and teaching styles in engineering education. *Engineering education,* 78**,** 674-681.

Feng, J., Lazar, J., Kumin, L. & Ozok, A. 2010. Computer usage by children with down syndrome: Challenges and future research. *ACM Transactions on Accessible Computing (TACCESS),* 2**,** 13, 10.1145/1714458.1714460.

Finstad, K. 2006. The system usability scale and non-native english speakers. *Journal of usability studies,* 1**,** 185-188.

Flickr. *Flickr* [Online]. Available: https://www.flickr.com [Accessed 23/01/2015].

Fontana, A. & Frey, J. 1994. *The art of science,* Sage Publications.

Fox, D. & Naidu, S. 2009. Usability evaluation of three social networking sites. *Usability News,* 11**,** 1-11.

Fraenkel, J. R. & Wallen, N. E. 2012. *How to design and evaluate research in education,* McGraw Hill Humanitie.

Fraser, D. M. 1998. *Action research for curriculum improvement in pre-registration midwifery education.* University of Nottingham.

Freeman, T. 2006. 'Best practice'in focus group research: making sense of different views. *Journal of Advanced Nursing,* 56**,** 491-497, 10.1111/j.1365-2648.2006.04043.x.

Fu, F., Liu, L. & Wang, L. 2008. Empirical analysis of online social networks in the age of Web 2.0. *Physica A: Statistical Mechanics and its Applications,* 387**,** 675-684, 10.1016/j.physa.2007.10.006.

Fu, L., Salvendy, G. & Turley, L. 2002. Effectiveness of user testing and heuristic evaluation as a function of performance classification. *Behaviour & Information Technology,* 21**,** 137-143, 10.1080/02699050110113688.

Gall, J. P., Gall, M. & Borg, W. R. 2009. *Applying Educational Research: How To Read, Do, And Use Research To Solve Problems Of Practice Author,* Addison-Wesley Longman, Incorporated.

Gallant, L. M., Boone, G. M. & Heap, A. 2007. Five heuristics for designing and evaluating Web-based communities. *First Monday,* 12.

Galliers, R. D. 1992. *Choosing appropriate information systems research approaches: In Information Systems Research: Issues, Methods and Practical Guidelines,* Blackwell Scientific Publications, Oxford.

Gamber, L. & Valent, E. 2001. *Web usability today: Theories, approach and methods,* Towards Cyberpsychology: Mind, Cognition, and Society in the Internet Age, IOS Press.

Garrett, J. J. 2010. *Elements of User Experience, The: User-Centered Design for the Web and Beyond*, Pearson Education.

Gerring, J. 2007. Case study research. *In:* SWANN, J. & PRATT, J. (eds.) *Educational Research in Practice.* London A&C Black, 111, 8.

Ghasemi, A. & Zahediasl, S. 2012. Normality tests for statistical analysis: a guide for non-statisticians. *International journal of endocrinology and metabolism,* 10**,** 486, 10.5812/ijem.3505.

Gillham, B. 2008. *Developing a questionnaire,* A&C Black.

Granka, L. A., Joachims, T. & Gay, G. 2004. Eye-tracking analysis of user behavior in WWW search. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, 2004-07-25, UK. ACM, 478-479, 10.1145/1008992.1009079.

Gray, W. D. & Salzman, M. C. 1998. Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human–Computer Interaction,* 13**,** 203-261, 10.1207/s15327051hci1303_2.

Greenberg, S. & Buxton, B. 2008. Usability evaluation considered harmful (some of the time). Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2008-04-06, Italy ACM, 111-120, 10.1145/1357054.1357074.

Gutwin, C. & Greenberg, S. 2000. The mechanics of collaboration: Developing low cost usability evaluation methods for shared workspaces. Enabling Technologies: Infrastructure for Collaborative Enterprises, 2000.(WET ICE 2000). Proeedings. IEEE 9th International Workshops on, 14/7/ 2000, Gaithersburg. IEEE, 98-103, 10.1109/ENABL.2000.883711.

Hackman, G. S. & Biers, D. W. 1992. Team usability testing: Are two heads better than one? Proceedings of the Human Factors and Ergonomics Society Annual Meeting. SAGE Publications, 1205-1209, 10.1177/154193129203601605.

Hanna, L., Risden, K. & Alexander, K. 1997. Guidelines for usability testing with children. *Magazine Interactions,* 4**,** 9-14, 10.1145/264044.264045.

Harper, E. R., Rodden, T., Rogers, Y., Sellen, A. & Human, B. 2008. Human-Computer Interaction in the Year 2020. 7 J J Thomson Avenue, Cambridge, CB3 0FB, England: Microsoft Research Ltd.

Harrison, G. W., Harstad, R. M. & Rutström, E. E. 2004. Experimental methods and elicitation of values. *Experimental economics,* 7**,** 123-140, 10.1023/B:EXEC.0000026975.48587.f0.

Hart, J., Ridley, C., Taher, F., Sas, C. & Dix, A. 2008. Exploring the facebook experience: a new approach to usability. Proceedings of the 5th Nordic conference on Human-computer interaction: building bridges, 2008-10-20, Sweden. ACM, 471-474, 10.1145/1463160.1463222.

Hartson, H. R., Andre, T. S. & Williges, R. C. 2003. Criteria for evaluating usability evaluation methods. *International Journal of Human-Computer Interaction,* 15**,** 145-181, 10.1207/S15327590IJHC1501_13.

Hasan, L. 2009. *Usability evaluation framework for e-commerce websites in developing countries.* PhD, Loughborough University.

Hasan, L. & Abuelrub, E. 2013. Common Usability Problems on Educational Websites. Proceedings of the 2013 International Conference on Education and Educational Technologies, Greece. 172-178.

Henninger, S. 2000. A methodology and tools for applying context-specific usability guidelines to interface design. *Interacting with computers,* 12**,** 225-243, 10.1016/S0953-5438(99)00013-2.

Henninger, S., Lu, C. & Faith, C. 1997. Using organizational learning techniques to develop context-specific usability guidelines. Proceedings of the 2nd conference on Designing interactive systems: processes, practices, methods, and techniques, 1997-08-01. ACM, 129-136, 10.1145/263552.263594.

Hersen, M., Turner, S. M. & Beidel, D. C. 2011. *Adult psychopathology and diagnosis,* New Jersey, John Wiley & Sons.

Hertzum, M. 2006. Problem prioritization in usability evaluation: From severity assessments toward impact on design. *International Journal of Human-Computer Interaction,* 21**,** 125-146, 10.1207/s15327590ijhc2102_2.

Hertzum, M. & Jacobsen, N. E. 2001. The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction,* 13**,** 421-443, 10.1207/S15327590IJHC1304_05.

Hertzum, M., Molich, R. & Jacobsen, N. E. 2014. What you get is what you see: revisiting the evaluator effect in usability tests. *Behaviour & Information Technology,* 33**,** 144-162, 10.1080/0144929X.2013.783114.

Hollingsed, T. & Novick, D. G. 2007. Usability inspection methods after 15 years of research and practice. Proceedings of the 25th annual ACM international conference on Design of communication, 2007-10-22, Texas, USA. ACM, 249-255, 10.1145/1297144.1297200.

Holzinger, A. 2005. Usability engineering methods for software developers. *Communications of the ACM - Interaction design and childrenof the ACM,* 48**,** 71-74, 10.1145/1039539.1039541.

Hooper, C. J. & Dix, A. 2012. Web science and human-computer interaction: when disciplines collide. Proceedings of the 3rd Annual ACM Web Science Conference, 2012-06-22, Evanston, United States. ACM, 128-136, 10.1145/2380718.2380736.

Hornbæk, K. 2006. Current practice in measuring usability: Challenges to usability studies and research. *International journal of human-computer studies,* 64**,** 79-102, 10.1016/j.ijhcs.2005.06.002.

Hornbæk, K. 2010. Dogmas in the assessment of usability evaluation methods. *Behaviour & Information Technology,* 29**,** 97-111, 10.1080/01449290801939400.

Hornbæk, K. & Frokjaer, E. 2004. Usability inspection by metaphors of human thinking compared to heuristic evaluation. *International Journal of Human-Computer Interaction,* 17**,** 357-374, 10.1207/s15327590ijhc1703_4.

Hornbæk, K. & Frøkjær, E. 2008. Comparison of techniques for matching of usability problem descriptions. *Interacting with Computers,* 20**,** 505-514, 10.1016/j.intcom.2008.08.005.

Hornbæk, K. & Law, E. L.-C. 2007. Meta-analysis of correlations among usability measures. Proceedings of the SIGCHI conference on Human factors in computing systems, 2007-04-29 USA ACM, 617-626, 10.1145/1240624.1240722.

Howe, K. R. 1988. Against the quantitative-qualitative incompatibility thesis or dogmas die hard. *Educational researcher,* 17**,** 10-16, 10.3102/0013189X017008010.

Hvannberg, E. T., Law, E. L.-C. & Lérusdóttir, M. K. 2007. Heuristic evaluation: Comparing ways of finding and reporting usability problems. *Interacting with computers,* 19**,** 225-240, 10.1016/j.intcom.2006.10.001.

Huang, Z. & Benyoucef, M. 2013. From e-commerce to social commerce: A close look at design features. *Electronic Commerce Research and Applications,* 12**,** 246-259.

Human Factors Society. 2016. *Definitions of Human Factors and Ergonomics* [Online]. Available: http://www.hfes.org/Web/EducationalResources/HFEdefinitionsmain.html#Website [Accessed 12/08/2016 2016].

Hwang, W. & Salvendy, G. 2010. Number of people required for usability evaluation: the 10±2 rule. *Communications of the ACM,* 53**,** 130-133, 10.1145/1735223.1735255.

Inostroza, R., Rusu, C., Roncagliolo, S., Jimenez, C. & Rusu, V. 2012. Usability heuristics for touchscreen-based mobile devices. Information Technology: New Generations (ITNG), 2012 Ninth International Conference on, 16/4/ 2012, Las Vegas, NV. IEEE, 662-667, 10.1109/ITNG.2012.134.

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

ISO 2010. ISO/IEC 9241 - 210: 2010 Ergonomics of human-system interaction -- Part 210: Human-centred design for interactive systems.

ISO 2011. ISO/IEC 25010: Systems and software engineering -- Systems and software Quality Requirements and Evaluation (SQuaRE) -- System and software quality models.

ISO, D. 1994. IS0 DZS 8402: Quality Vocabulary.

ISO, W. 1998b. 9241-11. Ergonomic requirements for office work with visual display terminals (VDTs). *The international organization for standardization*.

Jabbar, H. S., Gopal, T. & Aboud, S. J. 2007. An Integrated Quantitative Assessment Model For Usability Engineering. *Journal of Computer Science,* 3**,** 345.

Jain, P., Dubey, S. K. & Rana, A. 2012. SOFTWARE USABILITY EVALUATION METHOD. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET),* 1**,** pp: 28-33.

Jamal, A. & Cole, M. 2009. A heuristic evaluation of the Facebook's advertising tool beacon. Information Science and Engineering (ICISE), 2009 1st International Conference on, 26/12/ 2009, Nanjing. IEEE, 1527-1530, 10.1109/ICISE.2009.45.

Janesick, V. J. 2000. The choreography of qualitative research design. *Handbook of Qualitative Research.***,** 379-399.

Jeffries, R. 1994. Usability problem reports: Helping evaluators communicate effectively with developers. *Usability inspection methods.* John Wiley & Sons, Inc., 273-294.

Jeffries, R. & Desurvire, H. 1992. Usability testing vs. heuristic evaluation: was there a contest? *ACM SIGCHI Bulletin,* 24**,** 39-41, 10.1145/142167.142179.

Jeffries, R., Miller, J. R., Wharton, C. & Uyeda, K. 1991. User interface evaluation in the real world: a comparison of four techniques. Proceedings of the SIGCHI conference on Human factors in computing systems, 1991-04-27, USA ACM, 119-124, 10.1145/108844.108862.

Jeng, J. 2005. What is usability in the context of the digital library and how can it be measured? *Information technology and libraries,* 24**,** 3-12.

Jensen, C. & Potts, C. 2007. Experimental evaluation of a lightweight method for augmenting requirements analysis. Proceedings of the 1st ACM international workshop on Empirical assessment of software engineering languages and technologies: held in conjunction with the 22nd IEEE/ACM International Conference on Automated Software Engineering (ASE) 2007, 2007-11-05. ACM, 49-54, 10.1145/1353673.1353684.

Jiang, N. 2009. *A usability approach to improving the user experience in web directories.* Queen Mary University of London

John, B. E. & Kieras, D. E. 1996. The GOMS family of user interface analysis techniques: Comparison and contrast. *ACM Transactions on Computer-Human Interaction (TOCHI),* 3**,** 320-351, 10.1145/235833.236054.

Johnson, G. I. 1996. The usability checklist approach revisited. *Usability evaluation in industry.* UK: Taylor & Francis, 179-188, 20.

Johnston, J., Eloff, J. H. & Labuschagne, L. 2003. Security and human computer interfaces. *Computers & Security,* 22**,** 675-684, 10.1016/S0167-4048(03)00006-3.

Kantner, L. & Rosenbaum, S. 1997. Usability studies of WWW sites: heuristic evaluation vs. laboratory testing. Proceedings of the 15th annual international conference on Computer documentation, 1997-10-01, USA. ACM, 153-160, 10.1145/263367.263388.

Khajouei, R., Hasman, A. & Jaspers, M. W. 2011. Determination of the effectiveness of two methods for usability evaluation using a CPOE medication ordering system. *international journal of medical informatics,* 80**,** 341-350, 10.1016/j.ijmedinf.2011.02.005.

Kirk, R. E. 1982. *Experimental design: Procedures for the Behavioural Sciences,* Belmont, CA, USA, Brooks/Cole: Publishing Company.

Kitzinger, J. 1994. The methodology of focus groups: the importance of interaction between research participants. *Sociology of health & illness,* 16**,** 103-121, 10.1111/1467-9566.ep11347023.

Klenke, K. 2008. *Qualitative research in the study of leadership,* UK, Emerald group publishing.

Koshy, V. 2005. *Action research for improving practice: A practical guide*, Sage.

Kostaras, N. & Xenos, M. 2007. Assessing educational web-site usability using heuristic evaluation rules. Proceedings of 11th Panhellenic Conference in Informatics, 18-20 May 2007, Greece.: Citeseer, 543-550.

Krahmer, E. & Ummelen, N. 2004. Thinking about thinking aloud: A comparison of two verbal protocols for usability testing. *IEEE Transactions on Professional Communication* 47**,** 105-117, 10.1109/TPC.2004.828205.

Krueger, R. A. & Casey, M. A. 2000. *Focus groups: A practical guide for applied research* India, SAGE Piblications Asia Ltd.

KS3bitesize, B. 2012. *BBCKS3bitesize* [Online]. Available: http://www.bbc.co.uk/schools/ks3bitesize/ [Accessed 3/4/2012.

Kukulska-Hulme, A. & Shield, L. 2004. Usability and pedagogical design: Are language learning websites special? ED-MEDIA 2004, Switzerland. AACE Digital Library, 4235-4242.

Lancaster, G. A., Dodd, S. & Williamson, P. R. 2004. Design and analysis of pilot studies: recommendations for good practice. *Journal of evaluation in clinical practice,* 10**,** 307-312, 10.1111/j..2002.384.doc.x.

Lavery, D., Cockton, G. & Atkinson, M. P. 1997. Comparison of evaluation methods using structured usability problem reports. *Behaviour & Information Technology,* 16**,** 246-266, 10.1080/014492997119824.

Law, E. L.-C. & Hvannberg, E. T. 2004. Analysis of combinatorial user effect in international usability tests. Proceedings of the SIGCHI conference on Human factors in computing systems, 2004-04-25 Austria ACM, 9-16, 10.1145/985692.985694.

Law, L.-C. & Hvannberg, E. T. 2002. Complementarity and convergence of heuristic evaluation and usability test: a case study of universal brokerage platform. Proceedings of the second Nordic conference on Human-computer interaction, 2002-10-19, Denmark. ACM, 71-80, 10.1145/572020.572030.

Lazar, J. 2005. *Web usability: A user-centered design approach,* USA, Addison-Wesley Longman Publishing Co., Inc.

Lazar, J., Feng, J. H. & Hochheiser, H. 2010. *Research methods in human-computer interaction*, John Wiley & Sons.

Lee, J.-W. 2010. Online support service quality, online learning acceptance, and student satisfaction. *The Internet and Higher Education,* 13**,** 277-283, 10.1016/j.iheduc.2010.08.002.

Lee, Y. & Kozar, K. A. 2012. Understanding of website usability: Specifying and measuring constructs and their relationships. *Decision Support Systems,* 52**,** 450-463, 10.1016/j.dss.2011.10.004.

Lenhart, A. & Madden, M. 2007. *Social networking websites and teens: An overview,* Pew/Internet.

Lewis, C., Polson, P. G., Wharton, C. & Rieman, J. 1990. Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces. Proceedings of the SIGCHI conference on Human factors in computing systems, 1990-03-01, USA. ACM, 235-242, 10.1145/97243.97279.

Lewis, J. R. 1994. Sample sizes for usability studies: Additional considerations. *Human Factors: The Journal of the Human Factors and Ergonomics Society,* 36**,** 368-378, 10.1177/001872089403600215.

Lewis, J. R. 2001. Evaluation of procedures for adjusting problem-discovery rates estimated from small samples. *International Journal of Human-Computer Interaction,* 13**,** 445-479, 10.1207/S15327590IJHC1304_06.

Lewis, J. R. 2006. Sample sizes for usability tests: mostly math, not magic. *interactions,* 13**,** 29-33, 10.1145/1167948.1167973.

Liang, T.-P., Ho, Y.-T., Li, Y.-W. & Turban, E. 2011. What drives social commerce: The role of social support and relationship quality. *International Journal of Electronic Commerce,* 16**,** 69-90, 10.2753/JEC1086-4415160204.

Liljegren, E. & Osvalder, A.-L. 2004. Cognitive engineering methods as usability evaluation tools for medical equipment. *International Journal of Industrial Ergonomics,* 34**,** 49-62, 10.1016/j.ergon.2004.01.008.

Lindgaard, G. 2006. Notions of thoroughness, efficiency, and validity: Are they valid in HCI practice? *International journal of industrial ergonomics,* 36**,** 1069-1074, 10.1016/j.ergon.2006.09.007.

Lindgaard, G. & Chattratichart, J. 2007. Usability testing: what have we overlooked? CHI '07 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2007-04-29, USA ACM, 1415-1424, 10.1145/1240624.1240839.

Ling, C. & Salvendy, G. 2005a. Extension of heuristic evaluation method: a review and reappraisal. *Ergonomia-An International Journal of Ergonomics and Human Factors (IJE&HF),* 27**,** 179-197.

Ling, C. & Salvendy, G. 2005b. Extension of heuristic evaluation method: a review and reappraisal. *Ergonomia IJE & HF,* 27**,** 179-197.

Macefield, R. 2009. How to specify the participant group size for usability studies: A practitioner's guide. *Journal of Usability Studies,* 5**,** 34-45.

Mack, R. L. & Nielsen, J. 1994. Usability inspection methods. Proceeding CHI '94 Conference Companion on Human Factors in Computing Systems, USA ACM, 413-414, 10.1145/259963.260531.

Macleod, M. 1994. Usability in context: Improving quality of use. Human Factors in Organizational Design and Management–IV (Proceedings of the International Ergonomics Association 4th International Symposium on Human Factors in Organizational Design and Management, Stockholm, 29 May 1994, North Holland. G Bradley and HW Hendricks.

Magno, G., Comarela, G., Saez-Trumper, D., Cha, M. & Almeida, V. 2012. New kid on the block: Exploring the google+ social graph. IMC '12 Proceedings of the 2012 ACM conference on Internet measurement conference, 2012-11-14, USA ACM, 159-170, 10.1145/2398776.2398794.

Magoulas, G. D., Chen, S. Y. & Papanikolaou, K. A. 2003a. Integrating layered and heuristic evaluation for adaptive learning environments. Proceedings of the Second Workshop on Empirical Evaluation of Adaptive Systems, held at the 9th International Conference on User Modeling UM2003, Pittsburgh. 5-14.

Magoulas, G. D., Chen, S. Y. & Papanikolaou, K. A. 2003b. Integrating layered and heuristic evaluation for adaptive learning environments. UM2001. 5-14.

Maguire, M. 2001a. Context of use within usability activities. *International Journal of Human-Computer Studies,* 55**,** 453-483, 10.1006/ijhc.2001.0486.

Maguire, M. 2001b. Methods to support human-centred design. *International journal of human-computer studies,* 55**,** 587-634, 10.1006/ijhc.2001.0503.

Mahatody, T., Sagar, M. & Kolski, C. 2010. State of the art on the cognitive walkthrough method, its variants and evolutions. *Intl. Journal of Human–Computer Interaction,* 26**,** 741-785, 10.1080/10447311003781409.

Malinen, S. & Ojala, J. 2011. Applying the heuristic evaluation method in the evaluation of social aspects of an exercise community. Proceedings of the 2011 Conference on Designing Pleasurable Products and Interfaces, 2011-06-22, Italy ACM, 15, 10.1145/2347504.2347521.

Mankoff, J., Dey, A. K., Hsieh, G., Kientz, J., Lederer, S. & Ames, M. 2003. Heuristic evaluation of ambient displays. Proceedings of the SIGCHI conference on Human factors in computing systems, 2003-04-05 USA ACM, 169-176, 10.1145/642611.642642.

Manzari, L. & Trinidad-Christensen, J. 2013. User-centered design of a web site for library and information science students: Heuristic evaluation and usability testing. *Information technology and libraries,* 25**,** 163-169.

Mao, Y., Shen, H. & Sun, C. 2011. Google+ facebook: a social-network-optimized web search approach. The 6th International Workshop on Ubiquitous and Collaborative Computing (iUBICOM'11), Newcastle.

Mariage, C. & Vanderdonckt, J. 2001. A comparative usability study of electronic newspapers. *Tools for Working with Guidelines.* London: Springer, 325-337, 10.1007/978-1-4471-0279-3_31.

Markopoulos, P. & Bekker, M. 2003. On the assessment of usability testing methods for children. *Interacting with computers,* 15**,** 227-243, 10.1016/S0953-5438(03)00009-2.

Martin, R., Al Shamari, M., Seliaman, M. E. & Mayhew, P. 2014. Remote Asynchronous Testing: A Cost-Effective Alternative for Website Usability Evaluation. *International Journal of Computer and Information Technology* 3.

Mata, F.J. and Quesada, A., 2014. Web 2.0, social networks and e-commerce as marketing tools. *Journal of theoretical and applied electronic commerce research*, *9*(1), pp.56-69, 10.4067/S0718-18762014000100006.

McDaniel, C. & Gates, R. 2004. *Marketing research essential*, Wiley, New York.

Mendoza, V. & Novick, D. G. 2005. Usability over time. Proceedings of the 23rd annual international conference on Design of communication: documenting & designing for pervasive information, 2005-09-21, United Kngdm. ACM, 151-158, 10.1145/1085313.1085348.

Meredith, J. 1998. Building operations management theory through case and field research. *Journal of operations management,* 16**,** 441-454, 10.1016/S0272-6963(98)00023-0.

Miller, M. J. 2005. Usability in e-learning. *Learning circuits,* 48.

Molich, R. 2010. A critique of" How to specify the participant group size for usability studies: a practitioner's guide" by Macefield. *Journal of Usability Studies,* 5**,** 124-128.

Molich, R. & Dumas, J. S. 2008. Comparative usability evaluation (CUE-4). *Behaviour & Information Technology,* 27**,** 263-281, 10.1080/01449290600959062.

Molich, R., Ede, M. R., Kaasgaard, K. & Karyukin, B. 2004. Comparative usability evaluation. *Behaviour & Information Technology,* 23**,** 65-74, 10.1080/0144929032000173951.

Molich, R. & Nielsen, J. 1990. Improving a human-computer dialogue. *Communications of the ACM,* 33**,** 338-348, 10.1145/77481.77486.

Monk, A. 2002. Noddy's guide to usability. *British HCI Group,* 50**,** 31-33.

Muir, A., Shield, L. & Kukulska-Hulme, A. 2003. The pyramid of usability: A framework for quality course websites. Proceedings of EDEN 12th Annual Conference of the European Distance Education Network, The Quality Dialogue: Integrating Quality Cultures in Flexible, Distance and eLearning, 15 June 2003, Greece. European Distance Education Network (EDEN), 188-194.

Muller, M. J., Dayton, T. & Root, R. 1993. Comparing studies that compare usability assessment methods: an unsuccessful search for stable criteria. INTERACT'93 and CHI'93 Conference Companion on Human Factors in Computing Systems, 1993-04-01, Netherlands ACM, 185-186, 10.1145/259964.260197.

Myers, M. D. & Avison, D. 1997. Qualitative research in information systems. *Management Information Systems Quarterly,* 21**,** 241-242.

Myers, M. D. & Newman, M. 2007. The qualitative interview in IS research: Examining the craft. *Information and organization,* 17**,** 2-26, 10.1016/j.infoandorg.2006.11.001.

Nayak, P. R. 1991. Managing rapid technological development. *Boston, MA: Arthur D. Little*.

Nayebi, F., Desharnais, J.-M. & Abran, A. 2012. The state of the art of mobile application usability evaluation. Electrical & Computer Engineering (CCECE), 2012 25th IEEE Canadian Conference, April 29 2012-May 2 2012, Montreal, QC. IEEE, 1-4, 10.1109/CCECE.2012.6334930.

Nayebi, F., Desharnais, J.-M. & Abran, A. 2013. An Expert-based Framework for Evaluating iOS Application Usability. Software Measurement and the 2013 Eighth International Conference on Software Process and Product Measurement (IWSM-MENSURA), 2013 Joint Conference of the 23rd International Workshop on, Ankara. IEEE, 147-155, 10.1109/IWSM-Mensura.2013.30.

Napierala, M. A. 2012. What is the Bonferroni correction. *AAOS Now,* 6**,** 40-40.

Neto, E. V. & Campos, F. F. 2014. Evaluating the usability on multimodal interfaces: a case study on tablets applications. *Design, User Experience, and Usability. Theories, Methods, and Tools for Designing the User Experience.* Springer International Publishing, 484-495, 10.1007/978-3-319-07668-3_47.

Newman, I. 1998. *Qualitative-quantitative research methodology: Exploring the interactive continuum*, SIU Press.

Nielsen, J. 1992a. Finding usability problems through heuristic evaluation. Proceedings of the CHI '92 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 1992-06-01, USA. ACM, 373-380, 10.1145/142750.142834.

Nielsen, J. 1992b. The usability engineering life cycle. *Computer,* 25**,** 12-22, 10.1109/2.121503.

Nielsen, J. 1993. *Iterative User Interface Design* [Online]. Nielsen Norman Group. Available: http://www.nngroup.com/articles/iterative-design/ [Accessed 27/01/2015.

Nielsen, J. 1994a. *Guerrilla HCI: Using Discount Usability Engineering to Penetrate the Intimidation Barrier* [Online]. Nielsen Norman Group. Available: http://www.nngroup.com/articles/guerrilla-hci/ [Accessed 27/01/2015.

Nielsen, J. 1994b. Heuristic evaluation. *Usability inspection methods*, 24**,** 413.

Nielsen, J. 1994c. *Usability engineering*, California, Elsevier.

Nielsen, J. 1994d. Usability inspection methods. CHI '94 Conference Companion on Human Factors in Computing Systems. ACM, 413-414, 10.1145/259963.260531.

Nielsen, J. 1995a. *How to Conduct a Heuristic Evaluation* [Online]. Available: http://www.nngroup.com/articles/how-to-conduct-a-heuristic-evaluation/ [Accessed 27/01/5015.

Nielsen, J. 1995b. *Severity Ratings for Usability Problems* [Online]. Nielsen Norman Group. Available:http://www.nngroup.com/articles/how-to-rate-the-severity-of-usability-problems/.

Nielsen, J. 1995c. *Summary of Usability Inspection Methods* [Online]. Available: http://www.nngroup.com/articles/summary-of-usability-inspection-methods/ [Accessed 27/01/2015.

Nielsen, J. 2000a. *HOMERUN Heuristics for Commercial Websites*, [Online]. Available: www.useit.com.

Nielsen, J. 2000b. *Why You Only Need to Test with 5 Users* [Online]. Nielsen Norman Group. Available:http://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/ [Accessed 27/01/2015.

Nielsen, J. 2001. *Is Poor Usability Killing E-Commerce?* [Online]. Nielsen Norman Group. Available:http://www.nngroup.com/articles/did-poor-usability-kill-e-commerce/ [Accessed 27/01/2015.

Nielsen, J. 2003. *Recruiting Test Participants for Usability Studies* [Online]. Nielsen Norman Group. Available: http://www.nngroup.com/articles/recruiting-test-participants-for-usability-studies/ [Accessed 15/06/2011.

Nielsen, J. 2005. *Authentic Behavior in User Testing* [Online]. Nielsen Norman Group. Available:http://www.nngroup.com/articles/authentic-behavior-in-user-testing/ [Accessed 27/01/2015.

Nielsen, J. 2006. *Quantitative Studies: How Many Users to Test?* [Online]. Nielsen Norman Group. Available:http://www.nngroup.com/articles/quantitative-studies-how-many-users/ 2013].

Nielsen, J. 2009. *Discount Usability: 20 Years* [Online]. Nielsen Norman Group. Available: http://www.nngroup.com/articles/discount-usability-20-years/ [Accessed 28/01/2015.

Nielsen, J. 2012. *Usability 101: Introduction to Usability* [Online]. Nielsen Norman Group. Available:http://www.nngroup.com/articles/usability-101-introduction-to-usability/ [Accessed 27/01/2015.

Nielsen, J. & Landauer, T. K. 1993. A mathematical model of the finding of usability problems. Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems, 1993-05-01, Netherlands ACM, 206-213, 10.1145/169059.169166.

Nielsen, J. & Loranger, H. 2006. *Prioritizing web usability*, Pearson Education.

Nielsen, J. & Molich, R. 1989. Teaching user interface design based on usability engineering. *ACM SIGCHI Bulletin*, 21**,** 45-48, 10.1145/67880.67885.

Nielsen, J. & Molich, R. 1990. Heuristic evaluation of user interfaces. Proceedings of the SIGCHI conference on Human factors in computing systems. ACM, 249-256.

Nielsen, J. & Phillips, V. L. 1993. Estimating the relative usability of two interfaces: heuristic, formal, and empirical methods compared. Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems. 1993: ACM, 214-221.

Noiwan, J. & Norcio, A. 2000. A comparison analysis on web heuristic usability between Thai academic web sites and US academic web sites. Proceedings of SGI, World Multi Conference on Systems, Cybermetrics and Informatics, USA. 536-541.

Norman, D. A. 2002. *The design of everyday things*, Basic books.

Olesen, V. L., Bryant, A. & Charmaz, K. 2007. *Feminist qualitative research and grounded theory: Complexities, criticisms, and opportunities*, The SAGE handbook of grounded theory.

O'Murchu, I., Breslin, J. G. & Decker, S. 2004. Online Social and Business Networking Communities. ECAI Workshop on Application of Semantic Web Technologies to Web Communities.

Oehlert, G. W. 2000. A first course in design and analysis of experiments. *The American Statistician,* 57**,** 66-67, 10.1198/tas.2003.s210.

Oppenheim, A. N. 2000. Questionnaire design, interviewing and attitude measurement. *Journal of the Market Research Society*.

Oren, A., Nachmias, R., Mioduser, D. & Lahav, O. 1998. *Learnet: A model for virtual learning communities in the world wide web*, Tel-Aviv University, School of Education, Knowledge Technology Lab.

Osterbauer, C., Köhle, M., Grechenig, T. & Tscheligi, M. 1999. *Web usability testing: a case study of usability testing of chosen sites (banks, daily newspapers, insurances)* [Online]. Available: http://ausweb.scu.edu.au/aw2k/papers/osterbauer/paper.html [Accessed 01/12/2014.

Owens, J. W., Lenz, K. & Speagle, S. 2009. *Trick or Tweet: How Usable is Twitter for First-Time Users* [Online]. The Software Usability Research Lab (SURL) Available: http://usabilitynews.org/trick-or-tweet-how-usable-is-twitter-for-first-time-users/ [Accessed 2/11/2015.

Oztekin, A., Kong, Z. J. & Uysal, O. 2010. UseLearn: A novel checklist and usability evaluation method for eLearning systems by criticality metric analysis. *International Journal of Industrial Ergonomics,* 40**,** 455-469, 10.1016/j.ergon.2010.04.001.

Perfetti, C. & Landesman, L. 2001. Eight is not enough. *User Interface Engineering*.

Pessagno, R. 2010. *Design and usability of social networking web sites.* BS in Graphic Communication, California Polytechnic State University.

Petrie, H., Hamilton, F., King, N. & Pavan, P. 2006. Remote usability evaluations with disabled people. Proceedings of the SIGCHI conference on Human Factors in computing systems, 2006-04-22, Canada ACM, 1133-1141, 10.1145/1124772.1124942.

Pinelle, D., Wong, N. & Stach, T. 2008. Heuristic evaluation for games: usability principles for video game design. CHI '08 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2008-04-06, Italy. ACM, 1453-1462, 10.1145/1357054.1357282.

Polson, P. G., Lewis, C., Rieman, J. & Wharton, C. 1992. Cognitive walkthroughs: a method for theory-based evaluation of user interfaces. *International Journal of man-machine studies,* 36**,** 741-773, 10.1016/0020-7373(92)90039-N.

Ponterotto, J. G. 2005. Qualitative research in counseling psychology: A primer on research paradigms and philosophy of science. *Journal of counseling psychology,* 52**,** 126, 10.1037/0022-0167.52.2.126.

Preece, J. 2001. Sociability and usability in online communities: Determining and measuring success. *Behaviour & Information Technology,* 20**,** 347-356, 10.1080/01449290110084683.

Preece, J. & Maloney-Krichmar, D. 2003. Online communities: focusing on sociability and usability. *Handbook of human-computer interaction***,** 596-620.

Punch, K. F. 2013. *Introduction to social research: Quantitative and qualitative approaches*, Sage.

Raward, R. 2001. Academic library website design principles: development of a checklist. *Australian Academic & Research Libraries,* 32**,** 123-136, 10.1080/00048623.2001.10755151.

Redish, J. G., Bias, R. G., Bailey, R., Molich, R., Dumas, J. & Spool, J. M. 2002. Usability in practice: formative usability evaluations-evolution and revolution. CHI'02 extended abstracts on Human factors in computing systems. ACM, 885-890, 10.1145/506443.506647.

Reeves, T. C., Benson, L., Elliott, D., Grant, M., Holschuh, D., Kim, B., Kim, H., Lauber, E. & Loh, S. 2002. Usability and Instructional Design Heuristics for E-Learning Evaluation. World Conference on Educational Multimedia, Hypermedia & Telecommunications, June 24-29, 2002, Colorado. Association for the Advancement of Computing in Education (AACE).

Reips, U.-D. 2000. The Web experiment method: Advantages, disadvantages, and solutions. *Psychological experiments on the Internet.* Academic Press

Riihiaho, S. 2002. The pluralistic usability walk-through method. *Ergonomics in Design,* 10**,** 23-30.

Ritchie, J., Lewis, J., Nicholls, C. M. & Ormston, R. 2013. *Qualitative research practice: A guide for social science students and researchers*, Sage.

Rogers, Y., Sharp, H. & Preece, J. 2011. *Interaction design: beyond human-computer interaction*, Wiley Publishing.

Rowlands, I., Nicholas, D., Russell, B., Canty, N. & Watkinson, A. 2011. Social media use in the research workflow. *Learned Publishing,* 24**,** 183-195, 10.1087/20110306.

Rubin, J. & Chisnell, D. 2008. *Handbook of usability testing: howto plan, design, and conduct effective tests*, John Wiley & Sons.

Ruxton, G. & Colegrave, N. 2011. *Experimental design for the life sciences*, Oxford University Press.

Salvucci, D. D. & Goldberg, J. H. 2000. Identifying fixations and saccades in eye-tracking protocols. Proceedings of the 2000 symposium on Eye tracking research & applications, 2000-11-08, USA ACM, 71-78, 10.1145/355017.355028.

Sandelowski, M. 2000. Combining qualitative and quantitative sampling, data collection, and analysis techniques in mixed-method studies. *Research in nursing & health,* 23**,** 246-255.

Sauro, J. 2006. *Sample Size Calculator for Discovering Problems in a User Interface* [Online]. MeasuringU. Available: http://www.measuringu.com/problem_discovery.php 2015].

Sauro, J. 2009. *Is There A Difference In Usability Data From Remote Unmoderated Tests And Lab-Based Tests?* [Online]. Available: http://www.measuringu.com/unmoderated-testing.php [Accessed 27/01/2015.

Sauro, J. 2010. A Practical Guide to Measuring Usability: 72 Answers to the Most Common Questions about Quantifying the Usability of Websites and Software. A Measuring Usability LLC.

Sauro, J. 2011a. *10 Things to Know about Task Times* [Online]. MeasuringU. Available: http://www.measuringu.com/blog/task-times.php [Accessed 11/11/2014.

Sauro, J. 2011b. *10 Tips For Benchmark Usability Tests* [Online]. Available: https://www.measuringu.com/blog/benchmark-tips.com [Accessed 27/01/2015.

Sauro, J. 2011c. *The Four Corners Of Usability Measurement* [Online]. Available: http://www.measuringu.com/blog/four-corners.php [Accessed 27/01/2015.

Sauro, J. 2012a. *How Effective Are Heuristic Evaluations?* [Online]. MeasuringU. Available: http://www.measuringu.com/blog/effective-he.php [Accessed 27/01/2015.

Sauro, J. 2012b. *Triangulate For Better User Research* [Online]. Colorado and Redwood City, California, USA: Measuring Usability LLC. Available: http://www.measuringu.com/blog/triangulate-ux.php [Accessed 17/10/2014].

Sauro, J. & Kindlund, E. 2005. A method to standardize usability metrics into a single score. Proceedings of the SIGCHI conference on Human factors in computing systems, 2005-04-02, USA ACM, 401-409, 10.1145/1054972.1055028.

Sauro, J. & Lewis, J. R. 2012. *Quantifying the user experience: Practical statistics for user research,* USA, Elsevier.

Schmettow, M. 2012. Sample size in usability studies. *Communications of the ACM,* 55**,** 64-70, 10.1145/2133806.2133824.

SchoolsWorld. *SchoolsWorld* [Online]. Available: http://www.schoolsworld.tv/ [Accessed 23/01/2016.

Sears, A. 1997. Heuristic walkthroughs: Finding the problems without the noise. *International Journal of Human-Computer Interaction,* 9**,** 213-234, 10.1207/s15327590ijhc0903_2.

Sears, A. & Hess, D. J. 1999. Cognitive walkthroughs: Understanding the effect of task-description detail on evaluator performance. *International Journal of Human-Computer Interaction,* 11**,** 185-200, 10.1207/S15327590IJHC1103_1.

Sekaran, U. & Bougie, R. 2010. *Research methods for business: A skill building approach*, John Wiley & Sons, 10.1080/08832323.1993.10117635.

Seliger, H. W. 1989. *Second language research methods*, Oxford University Press.

Shackel, B. & Richardson, S. J. 1991. *Human factors for informatics usability*, Cambridge university press.

Shih, F. J. 1998. Triangulation in nursing research: issues of conceptual clarity and purpose. *Journal of advanced nursing,* 28**,** 631-641, 10.1046/j.1365-2648.1998.00716.x.

Shrivastava, M., Paperwala, T. & Dave, K. 2011. Trends in web technologies: Web 1.0 to Web 3.0 & beyond. The International Information Systems Conference (iiSC) 2011 Sultan Qaboos University, Muscat, Sultanate of Oman. 73.

Silius, K., Kailanto, M. & Tervakari, A.-M. 2011. Evaluating the quality of social media in an educational context. Global Engineering Education Conference (EDUCON), 2011 IEEE, Amman. IEEE, 505-510, 10.1109/EDUCON.2011.5773183.

Silva, P. A. & Dix, A. 2007. Usability: not as we know it! Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI... but not as we know it-Volume 2, 2007-09-03, UK. British Computer Society, 103-106.

Simeral, E. J. & Branaghan, R. J. 1997. A comparative analysis of heuristic and usability evaluation methods. ANNUAL CONFERENCE-SOCIETY FOR TECHNICAL COMMUNICATION. Citeseer, 307-309.

Sierkowski, B. Achieving web accessibility. Proceedings of the 30th annual ACM SIGUCCS conference on User services, 2002 New York. ACM, 288-291.

Skoool. 2012. *Skoool* [Online]. Available: http://lgfl.skoool.co.uk// [Accessed 3/4/2012.

Skov, M. B. & Stage, J. 2005. Supporting problem identification in usability evaluations. Proceedings of the 17th Australia conference on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future, 2005-11-21, Australia Computer-Human Interaction Special Interest Group (CHISIG) of Australia, 1-9.

Snyder, C. 2003. *Paper prototyping: The fast and easy way to design and refine user interfaces*, Morgan Kaufmann.

Somekh, B. 2005. *Action Research: A Methodology For Change And Development: a methodology for change and development*, McGraw-Hill Education (UK).

Spool, J. & Schroeder, W. 2001. Testing web sites: Five users is nowhere near enough. CHI'01 extended abstracts on Human factors in computing systems, 2001-03-31, USA ACM, 285-286, 10.1145/634067.634236.

Squires, D. & Preece, J. 1996. Usability and learning: evaluating the potential of educational software. *Computers & Education,* 27**,** 15-22, 10.1016/0360-1315(96)00010-3.

Ssemugabi, S. & De Villiers, M. 2007. Usability and learning: A framework for evaluation of web-based e-learning applications. Proceedings of EdMedia: World Conference on Educational Media and Technology 2007 Montgomerie & J. Seale (Eds.), 906-913.

Stanton, B., Theofanos, M. & Joshi, K. P. 2014. Framework for Cloud Usability. *Human Aspects of Information Security, Privacy, and Trust.* Springer International Publishing, 664-671, 10.1007/978-3-319-20376-8_59.

Stephanidis, C. 2001. *User interfaces for all: New perspectives into human-computer interaction*, CRC Press.

Stevenson, M. P. & Liu, M. 2012. Learning a language with web 2.0: Exploring the use of social networking features of foreign language learning websites. *Calico Journal,* 27**,** 233-259.

Stracke, C. M. & Hildebrandt, B. 2007. Quality Development and Quality Standards in e Learning: Adoption, Implementation, and Adaptation. Proceedings of EdMedia: World Conference on Educational Media and Technology 2007, Canada Association for the Advancement of Computing in Education (AACE), 4158-4165.

Tan, W.-s., Liu, D. & Bishu, R. 2009. Web evaluation: Heuristic evaluation vs. user testing. *International Journal of Industrial Ergonomics,* 39**,** 621-627, 10.1016/j.ergon.2008.02.012.

Tashakkori, A. & Creswell, J. W. 2007. Editorial: The new era of mixed methods. *Journal of mixed methods research,* 1**,** 3-7.

Thelwall, M. 2009. Social network sites: Users and uses. *Advances in computers,* 76**,** 19-73, 10.1016/S0065-2458(09)01002-X.

Thomas, C. & Bevan, N. 1996. Usability context analysis: a practical guide. *Loughborough University Institutional Repository*.

Thompson, A.-J. & Kemp, E. A. 2009. Web 2.0: extending the framework for heuristic evaluation. Proceedings of the 10th International Conference NZ Chapter of the ACM's Special Interest Group on Human-Computer Interaction, 2009-07-06, New Zealand. ACM, 29-36, 10.1145/1577782.1577788.

Thovtrup, H. & Nielsen, J. 1991. Assessing the usability of a user interface standard. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 1991-04-27, USA. ACM, 335-341, 10.1145/108844.108937.

Thyvalikakath, T. P., Monaco, V., Thambuganipalle, H. & Schleyer, T. 2009. Comparative study of heuristic evaluation and usability testing methods. *Studies in health technology and informatics,* 143**,** 322.

Triacca, L., Bolchini, D., Botturi, L. & Inversini, A. 2004. MiLE: Systematic usability evaluation for e-learning web applications. World Conference on Educational Multimedia, Hypermedia and Telecommunications. Switzerland, 4398-4405.

Trivedi, M. C. & Khanum, M. A. 2012. Role of context in usability evaluations: A review. *Human-Computer Interaction,* 3**,** 10, 10.5121/acij.2012.3208.

Tuckman, B. W. & Harper, B. E. 2012. *Conducting educational research*, Rowman & Littlefield Publishers.

Tullis, T., Fleischman, S., McNulty, M., Cianchette, C. & Bergel, M. 2002. An empirical comparison of lab and remote usability testing of web sites. Usability Professionals Association Conference.

Turner, C. W., Lewis, J. R. & Nielsen, J. 2006. Determining usability test sample size. *International encyclopedia of ergonomics and human factors,* 3**,** 3084-3088.

Usability.gov. 2016. *User Experience Basics* [Online]. Available: https://www.usability.gov/what-and-why/user-experience.html [Accessed 12/08/2016 2016].

Van den Haak, M. J., de Jong, M. D. & Schellens, P. J. 2004. Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: a methodological comparison. *Interacting with computers,* 16**,** 1153-1170, 10.1016/j.intcom.2004.07.007.

Vassar, A. 2012. *The effect of personality in sample selection for usability testing.* Masters University of New South Wales.

Virzi, R. A. 1992. Refining the test phase of usability evaluation: how many subjects is enough? *Human Factors: The Journal of the Human Factors and Ergonomics Society,* 34**,** 457-468, 10.1177/001872089203400407.

Walliman, N. 2006. *Social research methods*, Sage.

Webb, E. J., Campbell, D. T., Schwartz, R. D. & Sechrest, L. 1966. Unobtrusive measures: Nonreactive research in the social sciences. NCJRS Database.

Wentz, B. & Lazar, J. 2011. Are separate interfaces inherently unequal?: an evaluation with blind users of the usability of two interfaces for a social networking platform. Proceedings of the 2011 iConference, 2011-02-08. ACM, 91-97, 10.1145/1940761.1940774.

Wikipedia. 2016. *List of educational video websites* [Online]. Wikipedia. Available: https://en.wikipedia.org/wiki/List_of_educational_video_websites [Accessed 01/05/2016 2016].

Wikipedia. 2016. *Human factors and ergonomics* [Online]. Available: https://en.wikipedia.org/wiki/Human_factors_and_ergonomics [Accessed 12/08/2016 2016].

Wild, P. J. & Macredie, R. D. 2000. Usability evaluation and interactive systems maintenance. Proceedings of 2000 Annual Conference for the ComputerHuman Interaction Special Interest Group of the Ergonomics Society of Australia: Interfacing Reality in the New Millennium, Australia. Citeseer.

Wilson, C. 2007. Taking usability practitioners to task. *interactions,* 14**,** 48-49, 10.1145/1189976.1190004.

Wixon, D. 2003. Evaluating usability methods: why the current literature fails the practitioner. *interactions,* 10**,** 28-34, 10.1145/838830.838870.

Wolf, C., Carroll, J., Landauer, T., John, B. & Whiteside, J. 1989. The role of laboratory experiments in HCI: help, hindrance, or ho-hum? CHI '89 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 1989-03-01. ACM, 265-268, 10.1145/67450.67500.

Woolrych, A. & Cockton, G. 2001. Why and when five test users aren't enough. Proceedings of IHM-HCI 2001 conference, Cépadèus Toulouse,, France. 105-108.

Woolrych, A., Cockton, G. & Hindmarch, M. 2004. Falsification testing for usability inspection method assessment. Proceedings of HCI. 137-140.

Zaharias, P. & Poylymenakou, A. 2009. Developing a usability evaluation method for e-learning applications: Beyond functional usability. *Journal of Human–Computer Interaction,* 25**,** 75-98, 10.1080/10447310802546716.

Zapata, C. & Pow-Sang, J. A. 2012. Sample size in a heuristic evaluation of usability. *SOFTWARE ENGINEERING: METHODS, MODELING, AND TEACHING***,** 37.

Zaphiris, P. & Kurniawan, S. 2007. *Human computer interaction research in web design and evaluation*, Idea Group Inc (IGI).

Zhang, D. & Adipat, B. 2005. Challenges, methodologies, and issues in the usability testing of mobile applications. *International Journal of Human-Computer Interaction,* 18**,** 293-308, 10.1207/s15327590ijhc1803_3.

Zhu, X. & Liao, J. 2007. Web usability: A user-centered design approach. *Journal of the American Society for Information Science and Technology,* 58**,** 1066-1067, 10.1002/asi.20588.

APPENDICES

Appendix A1: Introductory script for evaluator (educational/social network websites)

Dear Evaluator,

My name is Roobaea AlRobaea, and I am a PhD student in the School of Computing Sciences at the University of East Anglia. As part of my studies, I am conducting a research project concerned with generating a domain specific inspection or evaluation method through an adaptive framework. To validate the newly developed Domain Specific Inspection (DSI) method, it should be compared against Nielsen's heuristics to establish which method delivers the most efficient results. Therefore, this research is aims to determine the extent to which these DSI guidelines help evaluators to discover usability problems in a website, and your contribution will greatly assist me in this research. I would like you to evaluate three websites by using the two sets of guidelines (DSI and Nielsen's heuristics). It should take about an hour to an hour and a half to complete. Please do not rush through the website - take your time. Also, please express your personal observations and opinions after finishing the evaluation on both methods. The main aim of this evaluation is to identify any problems that the websites have through using these methods; however, both positive and negative comments are very welcome. The researcher shall store the data in a safe place, and only I and my supervisor have permission to see them. The data will be used only in this study and it will be deleted permanently at the end of the study. So, if you are happy please read and sign the Consent Form. If you are not happy to do the experiment, please fill in the Withdrawal Form. I look forward to your reply via email. If you need more information please contact either me or my supervisor.

Thank you.
Yours sincerely,
Roobaea AlRoobaea
School of Computing Sciences
University of East Anglia (UEA), UK
R.Alrobaea@uea.ac.uk
Dr Pam Mayhew
P.mayhew@uea.ac.uk

Appendix A2: Introductory script for users (educational/social network websites)

- Welcome and Purpose

Dear Participants,

First of all, I would like to thank you very much for participating in this testing session. Let me tell you who I am and why I have asked you to come in today. My name is Roobaea AlRobaea, and I am a PhD student in the School of Computing Sciences at the University of East Anglia. As part of my studies, I am conducting a research project concerned with generating a domain specific inspection or evaluation method through an adaptive framework. The purpose of our work today is to explore three educational/social network website interfaces, by using lab usability testing, to find usability problems that could affect their efficiency, effectiveness and satisfaction. I am also interested in collecting your subjective responses relating to the usability of the targeted websites. The results of this experiment will be compared later to the results of another experiment by using heuristic evaluation and the newly developed method (DSI) to establish which method delivers the most efficient results. Thus, I am looking to you for your help.

- Procedures

You will be asked to use one website to perform a set of different tasks. While performing these tasks, you will also be asked to 'think aloud' as you work. The testing session should not take more than one hour to complete. Also, I am going to play the role of observer during the session and will be taking notes. The aim of the observation is to ensure that I have accurately understood what you have done during the session. After finishing the testing, you will be asked to complete a post-test questionnaire. These questions are designed to help me understand your feedback about the usability of the targeted website, and whether you like its interface and why. So, it is important that you answer truthfully and honestly based on your experience of using a website in the testing session.

- Risks/Discomfort

There are no risks in this experiment. Before the actual testing session, you can explore the website you are going to evaluate independently for 10 minutes; you may also read all the tasks and procedures, and you are welcome to ask any questions. Also, during the actual testing session you can stop your work without penalty, ask any questions you like, or

withdraw from the session. Furthermore, you can skip a task that you get stuck on or are unable to complete, and move on to the subsequent one. I will not be able to offer any suggestions or hints, but from time to time, I may ask you to clarify what you have said or ask you for information on what you are looking for or what you expect to happen. You may become fatigued while performing the tasks, so please do not hesitate to ask for a break and refreshment. Your results will be kept anonymous and protected in secure storage.

- Benefits

It is hoped that your results will be useful in validating the newly developed method (DSI), and in improving it to become an efficient method for evaluating the design of user interfaces, which could help people to design and use dynamic websites more effectively.

- Alternatives to participation

Your participation in this study is voluntary. You are free to withdraw or discontinue participation in any time.

- Cost and Compensation

Participation in this study will involve no cost to you. You will be paid for your participation.

- Confidentiality

The information provided by you will be treated as confidential and will be used only for research purposes. You will be identified through identification numbers only. No publications or reports will include identifying information on any participant. Do you have any questions?

If no;

(a) Please read and sign the Consent Form.

(b) Please fill out the pre-test questionnaire.

(c) If you are not happy to do the experiment, please fill in the Withdrawal Form.

(d) Please do not rush through the website - take your time.

Appendix A3: Consent Form

- Please read and sign this form.

	Tick as appropriate ☑
I confirm that I have read and understood the Introductory script.	
I have understood and I am happy with the evaluation/testing processes.	
I am happy with how the data will be collected and stored during and after the study.	
I understand that only the researcher will look at the data.	
I am satisfied that the data will be used according to the study needs only.	
I understand that the data will be deleted permanently at the end of the study.	
I confirm that all my questions have been answered.	
I understand that my participation is completely voluntary and I have the right to withdraw at any time, without giving a reason.	

- Please note that while you have the option to withdraw from the experiment at any time, it would be preferable to notify the researcher in advance. If you have any questions regarding this study please contact me through this email (R.Alrobaea@uea.ac.uk) or my supervisor Dr Pam Mayhew through this email (P.mayhew@uea.ac.uk).

Participant's/evaluator's signature…… Date ……

I appreciate your participation.

Appendix A4: Withdrawal form

Identification number:……

- Please read and sign this form.

	Tick as appropriate ☑
The test/evaluation is too long.	
The test/evaluation procedures are not clear enough.	
The researcher doesn't provide enough help.	
Poor communication with the researcher.	
I don't want to give a reason.	
Other reason(s):	

- If you have any questions, please contact me through this email (R.Alrobaea@uea.ac.uk) or my supervisor Dr Pam Mayhew through this email (P.mayhew@uea.ac.uk).

Appendix B1: Pre-test questionnaire for evaluators for educational websites

❖ Please answer the following questions about your background and experience

Section 1: Background and Experience

➢ Personal Information

- What is your name?

- What is your nationality?

- What is your first language?

- Which level of education do you have?

- What is your job? Is it related to usability issues?

➢ Skills

- How many years' experience do you have in Usability Engineering? …….
- Have you taken a Usability Engineering course?
  - ☐ Yes
  - ☐ No
- Have you taken a Human Computer Interaction course?
  - ☐ Yes
  - ☐ No
- Have you participated in a Usability Experiment before?
  - ☐ Yes If yes, how many ……
  - ☐ No

Section 2: Usability Experience in Education Websites

- Do you have a degree or any experience in the educational domain/websites?
  - ☐ Yes If yes, what is it? ……
  - ☐ No
- How frequently do you use educational websites?
  - ☐ Daily
  - ☐ Less often Examples of these websites……
- Do you have/have you had membership in any educational websites?
  - ☐ Yes If yes, which ones? ……
  - ☐ No
- How many educational websites have you evaluated before?
  - ☐ 1-2
  - ☐ 3-4
  - ☐ More than 5

Appendix B2: Pre-test questionnaire for evaluators for social network websites

❖ Please answer the following questions about your background and experience

Section 1: Background and Experience

➢ Personal Information

- What is your name?

- What is your nationality?

- What is your first language?

- Which level of education do you have?

- What is your job? Is it related to usability issues?

➢ Skills

- How many years' experience do you have in Usability Engineering? …….
- Have you taken a Usability Engineering course?
  - ☐ Yes
  - ☐ No
- Have you taken a Human Computer Interaction course?
  - ☐ Yes
  - ☐ No
- Have you participated in a Usability Experiment before?
  - ☐ Yes If yes, how many ……
  - ☐ No

Section 2: Usability Experience in Social Networks

- Do you have a degree or any experience in the social network domain/websites?
  - ☐ Yes If yes, what is? ……
  - ☐ No
- How frequently do you use social network websites?
  - ☐ Daily
  - ☐ Less often Examples of these websites……
- Do you have/have you had membership in any social network websites?
  - ☐ Yes If yes, which ones? ……
  - ☐ No
- How many social network websites have you evaluated before?
  - ☐ 1-2
  - ☐ 3-4
  - ☐ More than 5

Appendix B3: Pre-test questionnaire for users for educational websites

❖ Please answer the following questions about your background and experience

Section 1: Background and Experience

➢ Personal Information

- What is your age?

- What is your nationality?

- What is your first Language?

- Which level of education and occupation do you have?

➢ Computer Experience

- How many years have you been using a computer?
  - ☐ Less than 1 year
  - ☐ 1 to 4 years
  - ☐ More than 4 years
- How many daily hours do you use a computer?
  - ☐ Less than 2 hours
  - ☐ 2 to 4 hours
  - ☐ More than 4 hours

➢ Internet Experience

- Which browser do you use?
  - ☐ Internet Explorer
  - ☐ Google Chrome
  - ☐ Firefox Mozilla

  Other ……
- How many daily hours do you use the Internet?
  - ☐ Less than 2 hours
  - ☐ 2 to 4 hours
  - ☐ More than 4 hours
- How many years have you been using the Internet?
  - ☐ Less than 1 year
  - ☐ 1 to 4 years
  - ☐ More than 4 years

Section 2: Online Education Experience

- What is the value of these websites? .....
- How frequently do you use educational websites?
  - ☐ Daily
  - ☐ Less often Examples of these websites……
- Do you have/have you had membership in any educational websites? ☐ Yes ☐ No

If yes, which ones? …….

- What kinds of difficulties do you face on these websites? …….
- What kinds of activity do you do on these websites? .....

Appendix B4: Pre-test questionnaire for users for social network websites (SNSs)

❖ Please answer the following questions about your background and experience

Section 1: Background and Experience

➢ Personal Information

- What is your age?

- What is your nationality?

- What is your first language?

- Which level of education and occupation do you have?

➢ Computer Experience

- How many years have you been using a computer?
  - ☐ Less than 1 year
  - ☐ 1 to 4 years
  - ☐ More than 4 years
- How many daily hours do you use a computer?
  - ☐ Less than 2 hours
  - ☐ 2 to 4 hours
  - ☐ More than 4 hours

➢ Internet Experience

- Which browser do you use?
  - ☐ Internet Explorer
  - ☐ Google Chrome
  - ☐ Firefox Mozilla

  Other ……
- How many years have you been using the Internet?
  - ☐ Less than 1 year
  - ☐ 1 to 4 years
  - ☐ More than 4 years
- How many daily hours do you use the Internet?
  - ☐ Less than 2 hours
  - ☐ 2 to 4 hours
  - ☐ More than 4 hours

Section 2: Social Network Sites (SNSs) Experience

- What is the value of these websites? .....
- How frequently do you use SNSs?
  - ☐ Daily
  - ☐ Less often Examples of these websites ……
- Do you have/have you had membership in any SNSs? ☐ Yes ☐ No

If yes, which ones? …….

- What kinds of difficulty do you face on these websites? …….
- What kinds of activity do you do on these websites? .....

Appendix B5: Post- evaluation questionnaire for evaluators on both methods

- Evaluator's identification number: ……
- Group Name: ……..

❖ Feedback and improvement questionnaire This questionnaire is designed for your feedback on the methods you have just used on the three websites.

1. How much time did you spend on the actual evaluation?

No	Website Name	Method (HE / DSI)	Time
1	Website 1		
2	Website 2		
3	Website 3		

2 Did you find the structured problem report useful, and why?

☐ Yes ☐ No …….

3 Did you find the description of Heuristic Evaluation useful, and was it easy to use?

……

4 Did you find the description of DSI useful, and was it easy to use?

……

5. Which method helped you to discover more usability problems in these websites, and why?

☐ Heuristics Evaluation method ☐ DSI method …….....

6. Which method would you prefer to evaluate any other websites in same domain, and why?

☐ Heuristics Evaluation method ☐ DSI method …….....

7. Do you think that the length of time spent using the DSI method impacted on your evaluation results, and why?

☐ Yes ☐ No …….

8. Please give any suggestion that could improve the DSI method?

……

❖ Rating scale questionnaire for HE

This questionnaire is designed to reflect your feedback while performing the heuristic evaluation method. Please select the rating scale that most clearly expresses your view about each statement.

1. Strongly Disagree
2. Probably Disagree
3. Not sure
4. Probably Agree
5. Strongly Agree

#	Statements	1	2	3	4	5
1	I think that I would like to use the heuristic evaluation method frequently when I evaluate these websites.					
2	I found the heuristic evaluation method unnecessarily complex.					
3	I think the heuristic evaluation method was easy to use.					
4	The evaluation of these websites can be performed in a straightforward manner by using heuristic evaluation.					
5	I found the various principles in the heuristic evaluation method to be well integrated and specific for these websites.					
6	I think there was too much inconsistency in the heuristic evaluation method.					
7	I would imagine that most people would learn to use the heuristic evaluation method very quickly.					
8	I felt very confident using the heuristic evaluation method.					
9	I needed to learn a lot of things before I could get going with the heuristic evaluation method.					

❖ Rating scale questionnaire for DSI
This questionnaire is designed to reflect your feedback while performing the DSI method. Please select the rating scale that most clearly expresses your view about each statement.

1. Strongly Disagree
2. Probably Disagree
3. Not sure
4. Probably Agree
5. Strongly Agree

#	Statements	1	2	3	4	5
1	I think that I would like to use the DSI method frequently when I evaluate these websites.					
2	I found the DSI method unnecessarily complex.					
3	I think the DSI method was easy to use.					
4	The evaluation of these websites can be performed in a straightforward manner by using DSI method.					
5	I found the various principles in the DSI method to be well integrated and specific for these websites.					
6	I think there was too much inconsistency in the DSI method.					
7	I would imagine that most people would learn to use the DSI method very quickly.					
8	I felt very confident using the DSI method.					
9	I needed to learn a lot of things before I could get going with the DSI method.					

Appendix B6: Post-test questionnaire for user

Observer's Name: ……. Participant's Identification Number: ….. Date: ……..

Session starts at: ..... Session ends at: .....

Task #: ……..... Start Time: ……..... End Time: …….....

❖ Satisfaction Scale

The table below is for you to give a satisfaction level on the design features of each website, where 7 is highly satisfactory and 1 is high unsatisfactory.

1	2	3	4	5	6	7

- Could you please explain why have chosen this satisfaction rate?

  ……

- What do you like best about the site?

  ……

- What do you like least about the site (in terms of features provided)?

  ……

- If you were a website developer, what would be the first thing you would do to improve the website?

  ……

- Is there anything that you feel is missing on the site? (Probe: content or site features/functions)?

  ……

❖ Users' comments

The table below is designed to explain in detail any comments related to your feeling about the website or testing session.

No.	(Feedback and Recommendation)

Thank you for your cooperation in this research study

Appendix B7: Usability Test Observation Sheet

Participant's Identification Number: ……; Date: …….;

Session starts at: .....; Session ends at: .....;

Task #: ……....; Start Time: ……; End Time: ……....;

No.	Usability Problems observed

Notes:

……

.....

Appendix B8: Recording permission form

I agree to participate in this experiment, and understand that recordings will be made of my session but that I may leave this session at any time. Therefore, I grant my permission for Mr. Roobaea AlRoobaea to use these recordings for the purposes mentioned in the introductory script. Furthermore, I relinquish my right to review or inspect the recordings, and I understand that the recordings may be used by Mr. Roobaea Alrobaea without further permission.

Print Name: _____

Signature: _____ Date: _____

Appendix B9: Email sent to the owner of the chosen websites

Dear Mr/Ms ……,

My name is Roobaea AlRobaea, and I am a PhD student in the School of Computing Sciences at the University of East Anglia. As part of my studies, I am conducting a research project concerned with evaluating the usability of educational/social network websites. As your website is one of the websites that has fulfilled the criteria that were specified in our research study, I would like you to be involved in the context meeting, which aims to understand your website and to identify the user types and tasks that are performed in your website. The result of this meeting should assist in recruiting realistic participants, and in design representative tasks for the user testing sessions. Also, I would like to offer you the opportunity of being included in my study, which will involve three evaluation methods: user testing, heuristic evaluation and Domain Specific Inspection (DSI). The output of this study is a report that includes all the discovered usability problems and their severity, identifying usability problem areas and offering advice on how to improve your site. If you agree to cooperate, then all the relevant gathered data will be made freely available to you. Additionally, I undertake to keep all data confidential to yourselves; be assured that any data referenced in my thesis will be kept anonymous. This research represents a great opportunity for you to obtain very useful data for free, and thus I hope you will accept this offer. I look forward to your reply via email. If you need more information please contact either me or my supervisor.

Thank you.

Yours sincerely,

Roobaea Alrobaea
School of Computing Sciences
University of East Anglia (UEA), UK
R.Alrobaea@uea.ac.uk
Dr Pam Mayhew
P.mayhew@uea.ac.uk

Appendix B10: Interview agenda for context meeting

❖ Setting

- Interviewee Identification number:
- Date:
- Time:
- Location:

❖ Aim of the interview:

1. To understand the websites in the chosen domain.
2. To understand the context of use for these websites.
3. To identify any current problems or critical components that could affect usability, and to set overall usability goals for these websites.
4. To generate alternative solutions to each of the above problems (if there are any).
5. To define the user types who use these websites, for the actual evaluation.
6. To define task scenarios for the actual evaluation.
7. To specify the users' requirements.

❖ Interview questions:

- Why was the website developed?
- What are the overall objectives for this website?
- How will it be judged that this website is a success?
- What are the usability problems in the website that lead to low levels of user satisfaction, efficacy and effectiveness?
- What is the solution for each problem?
- Who are the intended user types for this website?
- What are users' expected experience and expertise in using the website's main functions?
- What tasks do users generally perform when they use the website?
- What are the users' requirements?
- What key functionality is needed to support the users' needs?

Appendix C: The First Experiment Tasks (Educational Domain)

❖ Task Scenarios for Website 1 (KS3 Bitesize)

Your teacher has recommended that, over the weekend, you visit Website 1, which is an online educational website; it offers many lessons in the English module with examples and tests (assessment). Kindly visit this website and take 10 minutes to explore and learn more about it, and to practice the Think Aloud protocol.

Task 1: < Motivational Factors >	
Task Goal	Browsing and gaining an impression of the website.
User's Task	You are in a browser and have typed 'ks3bitesize'. You feet that you want to know more about this website, so you have decided to surf it and get an idea about this website; what do you do?
Task 2: < Design and Media Usability >	
Task Goal	To measure the functionality of the searching feature and website's navigation design.
User's Task	You feel you want to get more information on the title 'Formal and informal writing'; try to find this title.
Task 3: < Learning Process >	
Task Goal	To measure the features of assessment, interactivity, resources and learning management.
User's Task	Having accessed the above title, take 2 minutes to self-study and get more information on this title. Then, you want to test your understanding of the lesson by conducting an assessment test; what do you do?
Task 4: < User Usability >	
Task Goal	To examine whether the website supports user tasks, avoids difficult concepts, and provides feedback and support services.
User's Task	• You feel you have difficulty in understanding some examples in the lesson 'Formal and informal writing, and you want ask other users of the website for help. Try to find the open discussion area and write your questions in there. What do you do? • If you want help, try to find a contact or the help page.

❖ Task Scenarios for Website 2 (Skoool)

Your teacher has recommended that, over the weekend, you visit Website 2, which is online educational website; it offers many lessons in the mathematics module with examples and assessments. Kindly visit this website and take 10 minutes to explore and learn more about it, and to practice the Think Aloud protocol.

Task 1: < Motivational Factors >	
Task Goal	Browsing and gaining an impression of the website
User's Task	You were in a browser and typed 'skoool'. You feet that you want to know more about this website, so you have decided to surf it and gain an idea about this website. What do you do?
Task 2: < Design and Media Usability >	
Task Goal	To measure the functionality of the searching feature and the website's navigation design.
User's Task	You feel you want to obtain more information under the title 'Integer'. Try to find this title.
Task 3: < Learning Process >	
Task Goal	To measure the features of assessment, interactivity, resources, and learning management.
User's Task	Once you get the above title, take 2 minutes to self-study and gather more ideas on this title. Then, you want to test your understanding of the lesson by conducting an assessment. What do you do?
Task 4: < User Usability >	
Task Goal	To examine whether the website supports user tasks, avoids difficult concepts, and provides feedback and support services.
User's Task	• You feel you want to download this lesson and send it to your friends. What do you do? If you want help, try to find a contact or the help page.

❖ Task Scenarios for Website 3 (Academicearth)

Your teacher has recommended that, over the weekend, you visit Website 3, which is online educational website; it offers many lessons with examples. Kindly visit this website and take 10 minutes to explore and learn more about it, and to practice the Think Aloud protocol.

Task 1: < Motivational Factors >	
Task Goal	Browsing and gaining an impression of the website.
User's Task	You were in a browser and typed 'academicearth'. You feet that you want to know more about this website, so you have decided to surf it and gain an idea about this website. What do you do?
Task 2: < Design and Media Usability >	
Task Goal	To measure the functionality of the searching feature and the website's navigation design.
User's Task	• Try to find Computer Science, and then select the topic 'Building Dynamic Websites'. After that, you feel you want read the course description; what do you do? • Once you have the above topic, try to find the Course Index and select the XML lesson. Listen to the video and subscribe to the podcast. What do you do?
Task 3: < Learning Process >	
Task Goal	To measure the features of assessment, interactivity, resources, and learning management.
User's Task	• Try to access the lecture videos of the instructor Courtenay Raia. • You feel you want download all the documents on the about Effective Computing course.
Task 4: < User Usability >	
Task Goal	To examine whether the website supports user tasks, avoids difficult concepts, and provides feedback and support services.
User's Task	• Task 1: What political science courses are offered by MIT University? • Task 2: Try to find a really interesting lecture among the political science courses. • You have a problem with one of the videos and you want report it; what do you do?

Appendix D: The Second Experiment Tasks (Social Network Domain)

❖ Task Scenarios for Website 1 (LinkedIn)

You read in a magazine about social network websites that offer many things such as contacting and keeping in touch with work colleagues, uploading pictures and videos, playing games, and sending private messages. Then, you decide to join some of these websites. Kindly visit this website and take 10 minutes to explore and learn more about it, and to practice the Think Aloud protocol.

Task 1: <Layout and Formatting>	
Task Goal	To measure the simplicity of the user interface and the consistency of the design.
User's Task	• You feel you want to gather more information on the LinkedIn website and then to join it. You have time to do so; what do you do?
Task 2: <User usability, Sociability and Management activities>	
Task Goal	To measure the ability of users to manage their profiles, to customize their settings, to explore the level of freedom and control, and to explore sociability on the website.
User's Task	• You have decided to add new information about yourself to your profile (personal info, experience, education), and you have time to do so; what do you do? • You want to be friend to someone, so add a new contact and send him/her a welcome message. You have time to do so; what do you do? • Suddenly, you want to revoke the recent connection, so what do you do? • You feel you want to change the notifications option for messages to include all opportunities available. You have time to do so; what do you do? • You feel you want to turn off the 'suggested people' option. You have time to do so; what do you do?
Task 3: < Business Support>	
Task Goal	To measure the services provided for supporting user businesses
User's Task	• You want to share a new movie with your friends, so they ask you to upload the picture and link to the movie trailer. You have time to do so; what do you do? • You feel you want to find information on an interesting company and to join an interesting RSS group. You have time to do so; what do you do? • You want to look for a job by uploading your CV, ask for recommendation, and placing an advert with interested groups. You have time to do so; what do you do? • You want to support your posting for the job by asking your contacts for recommendations (at school or in the workplace). You have time to do so; what do you do?
Task 4: < Security and Privacy >	
Task Goal	To measure the ability to change the privacy setting and to access the Privacy Policy.

User's Task	• You want change your privacy settings such that everyone can view your profile. What do you do? • You feel you want to know how the website will gather and use information on users who visit it. You have time to do so; what do you do?
Task 5: < Accessibility and Compatibility>	
Task Goal	To measure the accessibility and compatibility of the website.
User's Task	• You want to access the site map or 'contact us' page. You have time to do so; what do you do? • You feel you want use other devices (e.g. mobile and iPad) for accessing LinkedIn. You have time to do so; what do you do?
Task 6: < Navigation of Website and Search Quality>	
Task Goal	To measure ease of navigation and functionality of search on the website
User's Task	You feel you want to get more information of advertising in LinkedIn, You have got time to do so, and what do you do?

❖ Task Scenarios for Website 2 (Google Plus)

You read in a magazine about social network websites that offer many things such as contacting and keeping in touch with work colleagues, uploading pictures and videos, playing games, and sending private messages. Then, you decide to join some of these websites. Kindly visit this website and take 10 minutes to explore and learn more about it, and to practice the Think Aloud protocol.

Task 1: <Layout and Formatting>	
Task Goal	To measure the simplicity of the user interface and the consistency of the design.
User's Task	• You feel you want to gather more information on the Google Plus website and then to join it. You have time to do so; what do you do?
Task 2: <User Usability, Sociability and Management Activities>	
Task Goal	To measure the ability of users to manage their profiles, to customize their settings, to explore the level of freedom and control, and to explore sociability on the website.
User's Task	• You have decided to add new information about yourself to your profile (personal info, experience, education). You have time to do so; what do you do? • You want to be friend to someone, so add a new contact and send him/her a welcome message. You have time to do so; what do you do? • Suddenly, you want to revoke the recent connection, so what do you do? • You feel you want to change the notifications option for messages to include all opportunities available. You have time to do so; what do you do? • You feel you want to turn off the 'suggested people' option. You have time to do so; what do you do? • You feel you want to download data from Google+ (e.g. save a backup of your photos, profile information). What do you do?
Task 3: < Business Support>	
Task Goal	To measure the services provided for supporting user businesses.
User's Task	• You want to share a new movie with your friends, so they ask you to upload the picture and link to the movie trailer. You have time to do so; what do you do? • You feel you want to find information on an interesting company and then to join an interesting RSS group. You have time to do so; what do you do? • You want to make a circle of friends and then to send an image to the recently created circle; what do you do? • You want to make video call to your friend, so what do you do?
Task 4: < Security and Privacy >	
Task Goal	To measure the ability to change the privacy setting and to access the Privacy Policy.

User's Task	• You want change your privacy settings such that everyone can view your profile, so what do you do? • You feel you want to know how the website will gather and use information on users who visit it. You have time to do so; what do you do?
Task 5: < Accessibility and Compatibility>	
Task Goal	To measure the accessibility and compatibility of the website.
User's Task	• You want to access the site map or 'contact us' page. You have time to do so; what do you do? • You feel you want use other devices (e.g. mobile and iPad) for accessing Google Plus. You have time to do so; what do you do?
Task 6: < Navigation of Website and Search Quality>	
Task Goal	To measure the ease of navigation and the functionality of the search tool on the website.
User's Task	• You feel you want to gather more information on advertising in Google Plus. You have time to do so; what do you do? • Search for the best post. You have time to do so; what do you do?

❖ Task Scenarios for Website 2 (Ecademy)

You read in a magazine about social network websites that offer many things such as contacting and keeping in touch with work colleagues, uploading pictures and videos, playing games, and sending private messages. Then, you decide to join some of these websites. Kindly visit this website and take 10 minutes to explore and learn more about it, and to practice the Think Aloud protocol.

Task 1: <Layout and Formatting>	
Task Goal	To measure the simplicity of the user interface and the consistency of the design.
User's Task	• You feel you want to gather more information on the ecademy website and then to join it. You have time to do so; what do you do?
Task 2: <User Usability, Sociability and Management Activities>	
Task Goal	To measure the ability of users to manage their profiles, to customize their settings, to explore the level of freedom and control, and to explore sociability on the website.
User's Task	• You have decided to add new information about yourself to your profile (personal info, experience, education). You have time to do so; what do you do? • You want to be friend to someone, so add a new contact and send him/her a welcome message. You have time to do so; what do you do? • Suddenly, you want to revoke the recent connection, so what do you do? • You feel you want to change the notifications option for messages to include all opportunities available. You have time to do so; what do you do? • You feel you want to turn off the 'Suggested People' option. You have time to do so; what do you do?
Task 3: < Business Support>	
Task Goal	To measure the services provided for supporting user businesses.
User's Task	• You want to share a new business with your friends. You have time to do so; what do you do? • You feel you want to find information on an interesting company and then to join an interesting RSS group. You have time to do so; what do you do? • You want to create your blog for your company, so what do you do?
Task 4: < Security and Privacy >	
Task Goal	To measure the ability to change the privacy setting and to access the Privacy Policy.
User's Task	• You want change your privacy settings such that everyone can view your profile; so, what do you do? • You feel you want to see only the members in your research results who are online; so, what do you do? • You feel you want to know how the website will gather and use information on users who visit it. You have time to do so; what do you do?

Task 5: < Accessibility and Compatibility>	
Task Goal	To measure the accessibility and compatibility of the website.
User's Task	• You want to access the site map or the 'contact us' page. You have time to do so; what do you do? • You feel you want use other devices (e.g. mobile and iPad) for accessing Ecademy. You have time to do so; what do you do?
Task 6: < Navigation of Website and Search Quality>	
Task Goal	To measure the ease of navigation and the functionality of the search tool on the website.
User's Task	• You feel you want to gather more information on advertising in Ecademy. You have time to do so; what do you do? • Search for any business. You have time to do so; what do you do?

Appendix E: Usability problem report description

- Number: The numeric identifier of the problem.
- Heuristic name: The 10 heuristics of HE; the 21 DSI heuristics for the educational domain; the 26 DSI heuristics for social network domain.
- Problem description: Describing what is wrong and needs to be fixed and justifying why it is problematic.
- Problem context: Describing the context of the discovered problem, such as when the problem occurs, its impact, and the solution.
    - The impact is by a rating on a four-point scale:
      "(1) no problem, (2) a minor problem, that is a brief delay, (3) a serious problem, that is a significant delay but users eventually complete their task, and (4) a disaster, that is users voice strong irritation, are unable to solve the task, or solve it incorrectly". Also, impact can be rated as; "(1) Users quickly learn to get around the problem. (2) Users only learn to get around the problem after encountering it several times. (3) Users never learn how to get around the problem."
- Problem area: How the discovered problem is related to any usability problem areas based on the DSI method (the 5 usability problem areas in the educational domain, and the 7 usability problem areas in the social network domain).
- Problem severity: This to classify and prioritize the severity of a discovered problem, which means;
    - (0) I don't agree that this is a usability problem at all.
    - (1) Cosmetic problem only: need not be fixed unless extra time is available on the project.
    - (2) Minor usability problem: fixing this should be given low priority.
    - (3) Major usability problem: important to fix, so should be given high priority.
    - (4) Usability catastrophe: imperative to fix this before product can be released.

NO	Heuristic name	Problem description	Problem context	Problem area	Problem severity
1					
2					

Appendix F: Heuristics - evaluation and their explanation

Nielsen's heuristics	Explanation
Visibility of system status	The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.
Match between system and the real world	The system should speak the users' language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.
User control and freedom	Users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.
Consistency and standards	Users should not have to wonder whether different words, situations or actions mean the same thing. Follow platform conventions.
Error prevention	Even better than good error messages is a careful design that prevents a problem from occurring in the first place. Either eliminate error-prone conditions or check for them and present users with a confirmation option before they commit to the action.
Recognition rather than recall	Minimize the user's memory load by making objects, actions and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.
Flexibility and efficiency of use	Accelerators -- unseen by the novice user -- may often speed up the interaction for the expert user such that the system can cater to both

	inexperienced and experienced users. Allow users to tailor frequent actions.
Aesthetic and minimalist design	Dialogues should not contain information which is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.
Help users recognize, diagnose, and recover from errors	Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.
Help and documentation	Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large

Appendix G: Email sent to recruit participants for mini- user testing and user testing method

Dear Participant,

My name is Roobaea Alroobaea, and I am a student in the School of Computing Sciences at the University of East Anglia. I am currently doing a PhD degree in Computer Sciences, under the supervision of Dr Pam Mayhew. As part of my study, I am doing a research project concerned with evaluating educational/social network websites by using lab-based usability testing. Thus, I am looking to your help as volunteers to undertake the testing experiments. You will be asked to do several short tasks using a website. You will also be asked questions about your experience and perceptions of the website. The testing session should take no more than half hour. There will be £10 voucher for each participant for taking part in this study.

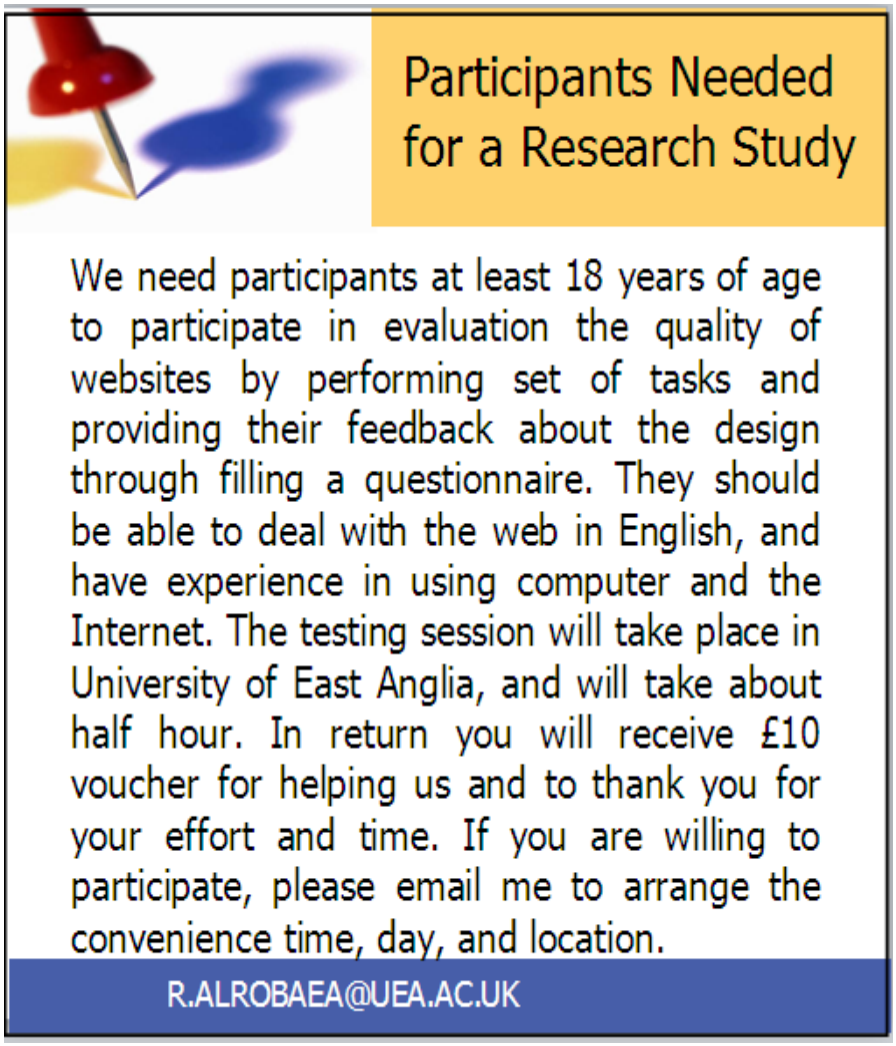
If you are willing to participate, please reply to my email and I will ask you some questions to help us to determine if you qualify for the study. After that, we will arrange a convenient date, time and place.

If you have any questions regarding this study please contact me by this email R.Alrobaea@uea.ac.uk

Thank you for your interest,

Yours sincerely,
Roobaea Alrobaea
School of Computing Sciences
University of East Anglia (UEA), UK

Appendix H: An advertisement to recruit users



Appendix I: Confirmation email for participating in our usability study

Dear Participant,

Thank you for agreeing to participate in our testing session. As I mentioned, you will be asked to perform a set of tasks on the targeted website, and to give us your thoughts about your experience. You won't need to prepare anything before the session.

You are scheduled to participate as follows:

Date: ……
Time: …….
Place: ……..

❖ A few key reminders:

- You will be given £10 voucher in exchange for your participation.
- During the study, we will ask you to perform some tasks using the website. You will be asked to talk aloud while you are thinking, so that the facilitator/observer can follow along.
- Your personal data will not be used for any purpose beyond this session.

Also, we have only one person scheduled at a time for these sessions, so if you find that you cannot participate on your scheduled day, please tell me as soon as possible so that I can reschedule your session.

Thanks again!

Yours sincerely,

Roobaea Alrobaea

School of Computing Sciences

University of East Anglia (UEA), UK

R.Alrobaea@uea.ac.uk

Appendix J: Questions after training session

- Have you heard about this website?

□ Yes
□ No

- Tell me what you know about this website?

……
……

- From looking at this site, what kinds of information do you think you could get from this site? Please be specific.

……
……

- Did you face any difficulties when you were talking aloud whilst looking for answers to the test questions?

……
……

Appendix K: Sets of tasks for Step Two 'User Input' in the adaptive framework for educational domain

❖ Task Scenarios for Educational Websites

Someone has asked that, over the weekend, you visit any online educational website you like; it should offer many lessons with examples and assessments. Kindly surf this website and take 10 minutes to explore and learn more about it, and to practice the Think Aloud protocol.

Task 1: < User Input >	
Task Goal	Mini-user testing to elicit feedback on educational websites from the real users
User's Task	• Browse the homepage • Click on Register on the homepage, and fill in the fields (if there are any) • Log out • Log in again using your username and password • Click on any link you like (e.g. site map, etc.) • Scroll down any page you are interested in • Go to the homepage • Go to a discussion board and add, edit and delete your question (if you can) • Go to "Contact us"; fill in a form to contact the Admin and send it • Type key words into "Search" and, using different search criteria, check the results • Try to enjoy some lessons that you are interested in and conduct assessment tests • Ty to customise a video lesson • Do you have a clear idea about the website? (Yes, No) • Did you face any problems? (Yes, No) • Do you have any recommendations or features that you think should be added to improve the website?

Appendix L1: The results of Step one ‘ Familiarization’ on the adaptive framework for educational domain

Table 1:Usability and instructional designs heuristics for evaluation of e-learning programs (Reeves et al., 2002)

No.	Heuristics
1	Visibility of system status
2	Match between system and the real world
3	Eror recovery and exiting
4	Consistency and standards
5	Eror prevention
6	Navigation suport
7	Help Documentation
8	Aesthetics
9	Interactivity
10	Message Design
11	Learning Design
12	Media Integration
13	Instructional Assessment
14	Resources
15	Feedback

Table 2: Evaluation criteria for web-based learning (Ssemugabi and De Villiers, 2007)

No.	Heuristics
1	Visibility of system status
2	Match between the system and the real world
3	User control and freedom
4	Consistency and adherence to standards
5	Error prevention, specifically prevention of peripheral usability-related errors
6	Recognition rather than recall
7	Flexibility and efficiency of use
8	Authenticity and minimalism in design
9	Recognition, diagnosis, and recovery from errors
10	Help and documentation
11	Simplicity of site navigation, organisation and structure
12	Relevance of site content to user
13	Clarity of goals, objectives and outcomes
14	Collaborative learning
15	Appropriateness of the level of learner control
16	Support for personally significant approaches to learning
17	Cognitive error recognition, diagnosis and recovery
18	Feedback, guidance and assessment
19	Context meaningful to domain and learner
20	Motivation, creativity and active learning

Table 3: Usability factors category (Bernérus and Zhang, 2010a)

No.	Heuristics
1	Visibility of system status
2	Match between the system and the real world
3	User control and freedom
4	Consistency and standards
5	Error Management
6	Learn-ability
7	Cognition facilitation, recognition & Memorability
8	Flexibility and Efficiency of Use
9	GUI
10	Help and documentation
11	Navigation and Exiting
12	Accessibility
13	Learning Content Design
14	Assessment
15	Motivation to Learn & interactivity
16	learning/authoring supportive tools

Table 4: Child e-learning usability heuristics (Alsumait and Al-Osaimi, 2009)

No.	Heuristics
1	Visibility of system status
2	Match between system and the real world
3	User Control and Freedom
4	Consistency and standards
5	Eror prevention
6	Recognition Rather than Recall
7	Flexibility and Efficiency of Use
8	Aesthetic and Minimalist Design
9	Help Users Recognize, Diagnose, and Recover from Errors
10	Help and Documentation
12	Design Attractive Screen layout
13	Use Appropriate Hardware Devices
14	Challenge the Child
15	Evoke Child Mental Imagery
16	Support Child Curiosity
17	Learning Content Design
18	Assessment
19	Motivation to Learn
20	Interactivity
21	Accessibility

Table 5: Categories with their items that are concerned with educational issues (Squires and Preece, 1996)

No.	Category
1	Content
	The content is accurate The content has educational value The content is free of race, ethnic, sex, and other stereotypes.
2	Instructional quality
	The purpose of the package is well defined The package achieves its defined purpose Presentation of content is clear and logical The level of difficulty is appropriate to the target audience Graphics/color/sound are used for appropriate instructional reasons Use of the package is motivational The package effectively stimulates student creativity Feedback on student responses is effectively employed The learner controls the rate and sequence of presentation and review Instruction is integrated with previous student experience Learning is generalisable to an appropriate range of situations
3	Technical quality
	The user support materials are comprehensive The user support materials are effective Information displays are effective Intended users can easily and independently operate the program Teachers can easily employ the package The program appropriately uses relevant computer capabilities The program is reliable in normal use

Appendix L2: The result of Step two ' User Input' on the adaptive framework for the educational domain

Table 1: Results of the context meeting on the education domain

❖ Aim of the interview:

8. To understand the websites in the chosen domain.
9. To understand the context of use for these websites.
10. To identify any current problems or critical components that could affect usability, and to set overall usability goals for these websites.
11. To generate alternative solutions to each of the above problems (if there are any).
12. To define the user types who use these websites, for the actual evaluation.
13. To define task scenarios for the actual evaluation.
14. To specify the users' requirements.

- Interview questions and answers:
- Why was the website developed?

It was developed to create a safe learning environment for students and an exchange of experiences between them and the teachers. Also, to provide distance education for students who cannot attend school.

- What are the overall objectives for this website?
- *It is a structured process carried out by the teacher in order to transfer his information and knowledge to others.*
- *Provides online support for learners who have difficulty in understanding or finding solutions to a certain problem.*
- *Creates an interactive environment which encourages the learner to participate in education, for example, interactive video, online assessment and discussion groups.*
- *Problem solving and real-life situations within the school environment, and the use of network resources to deal with them and resolve them.*
- *Gives learners independence and self-reliance in the search for knowledge.*
- *Develops a positive trend towards information technology through the use of the network by learners and communities, and thus establishes an advanced information society.*

- How will it be judged that this website is a success?
- *Based on the number of members.*
- *Based on the number of posts.*
- *Based on the number of users who browse this website*

- What are the usability problems in the website that lead to low levels of user satisfaction, efficacy and effectiveness?
- *Quality of content*
- *Quality of videos and audios*
- *Quality of navigation, structure and layout*

- What is the solution for each problem?
- *Updating the content frequently*
- *Avoiding useless content*
- *Adding references for the presented content to increase its credibility*
- *Using flash and photoshop in increase the quality of videos*
- *Applying the rules of the International Organization for Standardization (ISO) and World Wide Web Consortium (W3C) in terms of the navigation, structure and layout*

- Who are the intended user types for this website?
- *Students*

- *Teachers*
- *Parents*
- *Advisers and researchers*

- What are users' expected experience and expertise in using the website's main functions?

The ability to use the Internet, computer, laptop, mobile, ipad and their softwares.

- What tasks do users generally perform when they use the website?
- *Review some lessons*
- *Watch some explanation videos*
- *Participate in some puzzles or games related to a lesson*
- *Do some tests to evaluate learner understanding for a lesson*
- *Answer some questions or search for an answer*

- What are the users' requirements?

- *Content is always updated*
- *Provides interactive tools such as videos, chat، discussion forum، games and puzzles*
- *Provides external resources and archives*
- *Simple design*
- *Website is accessible by mobile devices*
- *Secure website*
- *Provides privacy feature*
- *Provides help centre and FAQ*

Table 2: List of usability problems discovered in mini-user testing

No	Problems discovered
1	The design is not attractive and encouraging
2	Background and text colours are not appropriate
3	Font text is small
4	Some videos are not working and some take a long time to work
5	Image size is not appropriate to website page
6	Some links or buttons are not logically grouped
7	The layout of the website is not consistent
8	It is difficult to find what I am looking for because the elements of the website are not structured well
9	Sometimes the top bar or left menu is not positioned in each page
10	The links' colour or button colour does not change after they have been visited
11	Some links are misleading and others do not work
12	The link or button to get back to the homepage is not positioned on each page
13	The logo is not positioned on each page
14	In some cases the error message is not available and some cases is not clear how to sort out problems
15	The 'Contact Us' link is not clearly positioned. Even it is provided, there is no contact by email or phone, and it is just by post to their address
16	FAQ page or help forum is not available
17	The website does not provide a search tool
18	When I login I cannot see https which means that the website is not secure
19	There are many advertisements which is annoying
20	Some information is not relevant to the displayed content and the website's aim
21	I visited empty pages

Table 3: Features that should be added to improve the educational website

No	Features	Frequency out of 10 users
1	Interactive design	5
2	Available resources	2
3	Simple layout	6
4	High quality content with frequent updating	9
5	Using videos, games and graphics instead of huge texts	7
6	Well structured navigation with high capabilities	10
7	Good and accurate search engine	6
8	Well grouped of links, buttons, and menus instead of scattered in different places on the homepage	8
9	Fast downloaded time	3
10	Social tools such as chatting and posting	7
11	Easy to reporting any problem and quickly get help	5
12	Support different luguages	4

Table 4: Developed heuristics based on results of mini- user testing

No	Heuristics
1	The user become engaged with the website through activities that challenge them
2	Use suitable colours and graphics that promote navigation
3	The font choice, colours and sizes are consistent with good user screen design.
4	The website should provide sufficient feedback (audio, video) to the users in order to provide corrective directions
5	All control items are logically labelled and grouped in a control panel
6	The font choices, colours and sizes are consistent with good user screen design.
7	The screen layout is efficient and visually pleasing (it should appear simple and consistent, i.e., uncluttered, readable and memorable)?
8	Content is logically structured in different sections and levels with enough space between the individual items
9	The colours and graphics are used suitable for promoting navigation.
10	All functionality is clearly labelled
11	The user's current position in the system is clearly labelled, and adequate 'back buttons'(to previous pages) are provided.
12	Navigation objects and tools are kept in particular, clearly defined positions, and they are an adequately viewable size.
13	The user can clearly identify where to start on the system's Homepage.
14	The site navigation is consistent, and the search engine is accurate.
15	The system display only information that is relevant to its purposes.
16	Sensitive areas are protected by passwords and an SSL protocol (e.g., VeriSign™) against hackers.
17	Adequate FAQ is offered?

Table 5: The result of content analysis regarding to identifying usability problem areas

No.	Areas from users	Strongly agree	Agree	Neither	Disagree	Strongly disagree
1	Adding references	100%				
2	Review lessons	60%	40%			
3	Videos quality	100%				
4	Puzzles	20%	60%	20%		
5	Game			20%	40%	40%
6	Assessment	100%				
7	Content updating	20%	60%		20%	
8	Interactive tools	60%	40%			
9	Chat	100%				
10	External resources	100%				
11	Simple design	100%				
12	Accessible	100%				
13	Privacy		40%	40%	20%	
14	Help centre	100%				
15	FAQ	100%				
16	Encouraging	100%				
17	Background colour	100%				
18	Font size	100%				
19	Videos not working	100%				
20	Videos take a long time	100%				
21	Image size and quality	100%				
22	Grouping of links and buttons	100%				
23	Layout consistent	100%				
24	Website structure	100%				
25	Top/bottom bar menu	100%				
26	Links' colours	100%				
27	Misleading links	100%				
28	Not working links	100%				
29	Links are not positioned obviously	100%				
30	Error message is not clear	100%				
31	Search tool	100%				
32	Not secured	100%				
33	Advertisements annoying	100%				
34	Relevant content	100%				
35	Empty pages	100%				
NO.	Areas from experts					
36	Provide descriptive tasks	40%	60%			
37	Breadcrumbs	60%	40%			
38	Minimal clicks	100%				
39	Easy bookmark	20%	60%	20%		
40	Meaningful feedback	80%	20%			
41	Site map	60%	20%	20%		
42	Support mental models	100%				
43	e-stories and role-playing	20%	40%	20%	20%	
44	Reliable institution	20%	60%		20%	
45	Control panel	40%		20%	40%	
46	Different language	40%	40%		20%	
47	Pop-up windows	20%	60%			20%

Appendix L3: The result of Step three ' Expert Input' on the adaptive framework for the educational domain

Table 1: Summary of focus group results

No	Advice
1	User login and registration should be easy to complete, and should avoid filling in more fields. Also, using validation tools when filling in the fields, such as email validation, would help to prevent potential errors.
2	To present the lesson elements, they should start gradually from the easy points to the difficult points to support the user's cognitive curiosity and to avoid the user's frustration.
3	The consistency of the layout, font choices, colours and sizes, logo, above and below the navigation bar, the undo and redo features and the size of the page represent essential rules that should be considered when designing any website.
4	The problems that I noticed when I visited some educational websites are a small font size with inappropriate colours in the background, content which is not designed for portable devices, and using traditional ways to deliver the educational content and not using e-stories, animations, simulations and games: these attract attention and increase the motivation of the user to learn and spend more time on the website.
5	Based on my experience, these websites should support users' tasks and make every single thing clear. For example, make important activities larger than others and make them logically labelled and grouped in a clear place.
6	The important point is that the designers should ensure that their website is easy to use. They should therefore keep the all items visible when they should be hidden from view, and vice versa, anticipate the user's next activity correctly, provide the minimal number of clickable actions, required selections, and need for scrolling to complete one main task.
7	From my point of view, using breadcrumbs to show where the user is and where the user last was is a very important tool, although some designers may not care to use it. Also, matching the menu structure to the task structure can help the users to know where they are going.
8	Better images, friendly design and using clear and simple language with correct spelling and grammar are the main characteristics for making the website understandable and easy to remember for novice users.
9	The error message must be meaningful and the help tab should appear to give a solution or hints on how to solve the problem.
10	The 'Contact Us' page, the help centre page, and the FAQ page are the most important pages after the main pages.
11	I find it annoying when I send a question to get help and the answer to the question is not clear or I do not receive an email informing me that the question has been received. These websites should provide feedback with meaningful information concerning their current level of achievement within the website and the feedback should be related to the user's task.
12	The low budget educational websites suffer from old content which has not been updated. Also, they also provide repetitive information and limited lessons.
13	Some educational websites offer too much content which makes it difficult to distinguish between options, and the content on these pages is not appropriate to their length.
14	The lesson pages are difficult to bookmark on some educational websites, and others do not provide an overview of the work that has been completed by the user, through devices such as tests after lessons.
15	Some educational websites provide incorrect information and this is because they are managed by inexperienced people.
16	Any educational website should provide mini exams after completion of each lesson to make sure that the user understands the lesson correctly. Also, it is better if a record of progress or report is kept which allows parents to monitor their children.
17	It is better to adopt social network tools such as chat, edit and add comments, upload, download, share, retrieve and organise. This makes the learning environment more motivational, enjoyable, easy to learn from, and interactive.

Appendix L4: Establishing the DSI method for educational

Usability problem area	The adaptive Domain Specific Inspection (DSI)
User usability	Supports modification and progress of evaluation Explanation: An educational website should prompt the user toward the next activity or work to be completed by the user. Keys that are important in helping users to perform their tasks should be highlighted.
	Supports user tasks and avoids difficult concepts Explanation: An educational website should be in user-understandable form (clear, simple language, graphics, correct spelling and grammar). It should provide descriptive tasks (brief, unambiguous) and easy, approachable items (view hidden items). It must use breadcrumbs (secondary navigation scheme) to show users their current position. It should help users in completing their main task (with minimal clicks, infrequent selection/scrolling, easy bookmark). Its menu and task structure must match, so that the user can distinguish between options and contents.
	Feedback and support services Explanation: An educational website should provide helpful and meaningful feedback on time. There should be extended feedback options with FAQ and performance support tools to help user (i.e. site map, contact us, help centre).
	Error Prevention Explanation: An educational website must prevent users from errors and provide ways to recover (undo, redo, and validation field) or minimize errors.
	Easy to remember Explanation: An educational website should support the mental models of users and help them to reduce their memory load.
	Supports learner curiosity

Motivational factors	Explanation: The website should support the user's curiosity through surprises, paradoxes and humour. It should facilitate the user in posting their queries and difficulties, and in seeking solutions.
	Learning content design and Attractive screen design Explanation: The website should use good terminology and vocabulary with organised content and learning objects. It should have an efficient, simple and visually pleasing layout, with good background, colours, font and user screen design to help users achieve their primary goals easily.
	Motivation to learn Explanation: An educational website must provide learning activities along with e-stories, simulations, discussion messages and role-playing to motivate users. It must facilitate users with different difficulty levels and action rewards (by audio/video/ text/animation). Its learning sessions must be designed so as to minimize user fatigue.
Content information and process orientation	Relevant, correct and adequate information Explanation: The website should provide concise, non- repetitive, relevant and updated information and content, suitable to page length, with readable text size.
	Reliability and Validity Explanation An educational website must be reliable, stable, provide continuity of learning and be built by a reliable institution. It must provide a link to go to the source page. It must provide a means by which other users and their content can be validated.
	Privacy and Security Explanation: An educational website should provide complete protection, using passwords and SSL protocol.
Learning process	Assessment Explanation:

	The website should provide user assessment reports (audio/ video/text/graph) with feedback, corrective directions and instructions (with evaluation and tracking reports).
	Interactivity Explanation: An educational website should be designed in a manner that engages users with activities and challenges. It should offer an easy interactive approach that helps users to respond quickly and to gain in confidence.
	Evokes mental images for the users Explanation: An educational website should support the users' mental model and allow them to use their imagination to enhance comprehension. It should adopt characters or contexts from the user's own culture, so that the user can interpret and recognise them.
	Resources Explanation: The website should provide a wide range of resources and up-to-date links.
	Learning management Explanation: The website should support additional guidance (chat, edit, seek instruction, etc.) and learning styles that support synchronous and asynchronous modes. Lessons with easy upload, download, share, retrieve and organise, help users to learn with ease. Learning programs must be designed in a manner so that users can manage all the activities with ease. The control panel must have logically labelled and grouped items.
	Learnability Explanation: The website should be designed in a manner that the user finds easy to use.
Design and media usability	Multimedia representations Explanation: An educational website should support interactive multimedia (audio, video etc.) with customised audio, video and difficulty level settings to make learning enjoyable. Multimedia should include surprises, humour and interesting representations

	with meaningful feedback and hints (and a skip option).
	Accessibility and compatibility of hardware devices Explanation: An educational website must be compatible with various platforms and hardware devices, and should be matched with the necessary computer skills. To prevent users making input errors, devices and buttons with no functionality must be disabled. Easily accessible lessons with different language contents help users.
	Functionality Explanation: An educational website must have well-defined labels to compete the task without leaving the current environment. The website must provide a clear status for each task on all pages.
	Navigation and Visual clarity Explanation: An educational website should have consistent navigation tools and objects with a homepage, back links and an accurate search engine. It must have a map, a table of contents, labelled menus, buttons and links with clear functionality and mouse-over or pop-up windows. The website's content should be logically structured in different sections, and it should avoid unnecessary flash/animation.

Appendix M1: Example on how to develop DSI checklist for educational domain

<table>
<tr><th>Usability problem area</th><th>Resource</th><th>From</th><th>The Adaptive Domain Specific Inspection (DSI) Checklist</th></tr>
<tr><td rowspan="11">Content information and process orientation</td><td>Relevance of site content to user</td><td>(Ssemugabi and De Villiers, 2007)</td><td rowspan="11"><u>Heuristic1</u>: Relevant, correct and adequate information:
o Does the website display only information that is relevant to the purposes of it?

o Does the website update the content constantly?

o Does the website display only the available lesson, and the content is suitable to page length?

o Does the website provide concise and non- repetitive information?

o There is no too much information on a page, and all text is viewable size.</td></tr>
<tr><td>Learning Content Design</td><td>(Alsumait and Al-Osaimi, 2009)</td></tr>
<tr><td>Presentation of content is clear and logical</td><td>(Squires and Preece, 1996)</td></tr>
<tr><td>Quality of content</td><td>Context meeting interview</td></tr>
<tr><td>Font text is small in the content</td><td rowspan="2">Mini-user testing Problems</td></tr>
<tr><td>Some information are not relevant to the displayed content and the website aim.</td></tr>
<tr><td>High quality content with frequent updating</td><td>User testing post questionnaire</td></tr>
<tr><td>The system display only information that is relevant to its purposes.</td><td rowspan="3">Developed heuristics based on results of mini- user testing</td></tr>
<tr><td>Content is logically structured in different sections and levels with enough space between the individual items</td></tr>
<tr><td>The font choice, colours and sizes are consistent with good user screen design.</td></tr>
<tr><td>The problems that I noticed when I visited some educational websites are that small font size with suitable colours to the background, content is not designed to portable devices, using traditional way to deliver the educational content such as do not use e-stories, animations, simulations and games to get attention and increase the motivation of the user to learn and spent more time on the website.</td><td rowspan="2">Summary of focus group results</td></tr>
<tr><td>Some educational websites provide wrong information and this because they are managed by inexperienced.</td></tr>
</table>

Appendix M2: Establishing the DSI checklist for educational domain

Usability problem area	The Adaptive Domain Specific Inspection (DSI) Checklist
User usability	*Supports modification and progress of evaluation:* o Does a website make important keys larger than other keys? o Does a website anticipate the user's next activity correctly? o Does a website allow the user to initiate actions? o Does a website provide an overview of the work process that has been completed by the users?
	o *Supports user tasks and avoids difficult concepts:* o Does a website provide constructive, brief, unambiguous descriptions of the task when needed? o Does a website match the menu structure to the task structure? Can user distinguish between options and content on the pages? Are there breadcrumbs to show where user is and where user last was? o Does a website use clear, simple language for questions and answers? o Does a website provide correct spelling and grammar, and understandable graphic symbols? o Does a website provide the few number of clickable actions, infrequent selection, and infrequent scrolling to complete one main task? Are users easy to bookmark a page of lesson? o Is an item visible when it should be hidden from the view, vice versa?
	Feedback and support services: o Is the feedback given at any specific time tailored to the content or problem being studied by the learner? o Does feedback provide the users with meaningful information concerning their current level of achievement within the program? And status of message helping is related to the user task. o Does the website program provide users with opportunities to access extended feedback from instructors through email and Internet communication? Or Offer adequate FAQ. o Is performance support tools provided that mimic their access in the real world?
	Error Prevention: o Do error messages prevent potential errors from happening? o Does website provide solutions help users to recover from error, such as providing undo and redo features? o Can be errors averted or minimized when possible?
	Easy to remember : o Are the casual users able to return to using the website after some period without having to learn everything all over again? All functions and information are well presented to support memorability.
Motivational factors	*Supports leaner curiosity:* o Does the website support the user's cognitive curiosity through surprises, paradoxes, humor, and deals with topics that already interest the users?
	Learning content design and Attractive screen design: o The vocabulary and terminology used are appropriate and are presented with a good background, giving suitable examples. o The organization of the content pieces and learning objects is suitable for achieving the primary goals of the system. o Are the similar learning objects organized in a similar style? o The screen layout is efficient and visually pleasing. It should appear simple, i.e., uncluttered, readable, and memorable.

	○ The font choice, colours and sizes are consistent with good user screen design.
	Motivation to learn: ○ Does the website use e-stories, simulations, discussion messages, role playing, and activities to gain the attention and to maintain the motivation of users to learn more. ○ Does the website provide the users with frequent and varied learning activities that increase learning success? ○ Are the user's actions rewarded by audio, video, text, or animations and the rewards are meaningful. ○ Is the website easy to learn, but hard to master? Is the website paced to apply pressure but not frustrate the users? The difficulty level varies so that the users have greater challenges as they develop mastery. ○ Is the user's fatigue minimized by varying activities and difficulties during learning sessions?
Content information and process orientation	*Relevant, correct and adequate information:* ○ Does the website display only information that is relevant to the purposes of it? ○ Does the website update the content constantly? ○ Does the website display only the available lesson, and the content is suitable to page length? ○ Does the website provide concise and non- repetitive information? ○ There is no too much information on a page, and all text are viewable size.
	Reliability and Validity: ○ Is there a link provided to the homepage? Look for a reliable institution. ○ Are reliability, stability and continuity of learning in the website guaranteed?
	Privacy and Security: ○ Protected areas inaccessible using passwords and SSL protected by "Verisign" to avoid any hacking.
Learning process	*Assessment:* ○ Does the website include self-assessment for each module, e.g. audio, video & writing, and keeps a record of progress. ○ Does the website provide sufficient feedback (audio, video) to the users in order to provide corrective directions? ○ Does the website provide the instructor with users' evaluation and tracking reports?
	Interactivity : ○ Does the user become engaged with the website program through activities that challenge them? ○ Does the user able to respond to the program at leisure. Does Lessons presentation promote engagement to users? ○ Des learning become easier with an interactive approach wherein users are taught to respond to the program. Does user gain confidence by doing so? ○ Does the user have confidence that the website is interacting and operating in the way it was designed to?
	Evokes mental images for the users: ○ Does the website allow the users to use their imagination, which enhances their comprehension? ○ Does the website appeal to the imagination and encourages recognition in order for the users to create unique interpretations of the characters or contexts. ○ Are the users interested in the website characters because they are drawn from the user's own culture?
	Resources: ○ Does a website provide access to all range of resources (e.g. examples and real date archives) appropriate to the learning context? ○ If the website includes links to external w.w.w. or intranet resources, are the links kept up-to-date?

	Learning management: o Does the user manage all the activity pertaining to the learning program easily, clearly understand everything, and perceive options for additional guidance, chat, edit, add, instruction, or other forms of assistance when needed. All control items are logically labelled and grouped in control panel. o Lessons are easy to upload, download, shared, retrieval, organise, support various learning styles, and support synchronous and asynchronous modes.
	Learnability: o The capability of the website to enable the users to learn how to use it.
Design and media usability	*Multimedia representations:* o Does multimedia help users in all aspects to learn interactively by playing videos, audios, audio mock tests and making it enjoyable to learn? o Does the website include sound and visual effects? These effects should provide meaningful feedback, hints, or stir particular emotions. o Does the website include surprises, humor and interesting representations for the learner, and avoids unnecessary multimedia representations as they can confuse a user that has just started to work with the system. o -Allows the users to skip non-playable and frequently repeated content in videos or learning games. o Allows the users to customize video and audio settings, and difficulty level.
	Accessibility and compatibility of hardware devices: o Compatibility of website on various platforms and different hardware. Also, its features are adaptable on individual user preferences. o Potential e-users have all the necessary computer skills to use the application. There should be consistency between the motor effort and skills required by the hardware and the developmental stage of the learner audience. o All input devices/buttons that have no functionality are disabled to prevent user input errors. o Are the lessons accessible to users with physical impairments, and their content transcribed to various languages?
	Functionality: o All necessary functionality of the website is available without leaving the site, and works correctly. o All functionality is clearly labelled, easy to complete a task, and website status of each task is clear on the pages.
	Navigation and Visual clarity: o Navigation objects and tools are kept in particular, clearly defined positions, and viewable size. o Unnecessary animation and Flash is avoided. o Content logically structured in different sections and levels with enough space between items design. Also, use suitable colours and graphics that promote navigation. o Menus understandable and straightforward and items are logically grouped and labelled. All buttons, links, and features have 'mouseover' or pop up window which can provide meaningful feedback. o Site map and/or table of contents are available, as well as, calendar. o Consistent navigation throughout site, and search engine accurate. o Clear label of current position on system, and users distinguish where to start on the Homepage of the system, and provide adequate back button to a previous page. o Reducing the need for scrolling action on a page to find items. o All functions, buttons, and links are labels meaningful, and their intended functionality is clear.

Appendix N: The three methods' performances in discovering usability problems for the educational domain

No:	Usability problems discovered	Website	Area	Severity rating	Method
1	Page load takes too long to run videos.	Skoool	Lessons page	2	HE
2	Inconsistency in design of main menu.	Skoool	Homepage	1	HE
3	Inconsistency in the heading of a page.	Skoool	Homepage	1	HE
4	Misleading links	Skoool	Whole website	2	HE
5	Inconsistency in colour of page design	Skoool	Whole website	1	HE
6	There is an empty content in the one of the external pages.	Skoool	Homepage	3	HE
7	Some graphics took a long time to finish, and for text to appear in the videos.	Skoool	Lessons page	2	HE & DSI & UT
8	Help or 'contact us' is not stuck at the bottom of the page.	Skoool	Whole website	4	HE & DSI & UT
9	Site map is not provided in this website.	Skoool	Whole website	2	HE & DSI & UT
10	Some button icons on the homepage for external pages are not working correctly and other are not clickable.	Skoool	Homepage	2	HE & DSI & UT
11	There is too much content in some pages.	AcademicEarth	Whole website	1	HE & DSI & UT
12	It does not hold the user's info after registration.	AcademicEarth	Registration page	3	HE & DSI & UT
13	There are no context labels to show the user what the link has or which page it is on.	AcademicEarth	Whole website	2	HE & DSI & UT
14	Some of the terminology used is inappropriate.	AcademicEarth	Whole website	2	HE & DSI & UT
15	Difficult to comment on without having gone through an entire course.	AcademicEarth	Course page	2	HE & DSI & UT
16	The link provided to the homepage is not obvious.	AcademicEarth	Homepage	1	HE & DSI & UT
17	The website does not anticipate the user's next activity correctly.	AcademicEarth	Whole website	3	HE & DSI & UT
18	The website does not contain a site map link.	AcademicEarth	Whole website	2	HE & DSI & UT
19	The reporting link on the videos is not clearly positioned.	AcademicEarth	Course page	1	HE & DSI & UT

20	The advertisement covers the homepage and is not relevant to the website content.	AcademicEarth	Homepage	1	HE & DSI & UT
21	Some course links and some videos are not working.	AcademicEarth	Course page	3	HE & DSI & UT
22	The videos take a long time to work.	AcademicEarth	Course page	3	HE & DSI & UT
23	Registration form is too long.	AcademicEarth	Registration page	1	HE & DSI & UT
24	The website does not support testing or activities for some lessons.	BBC KS3bitesize	Course page	2	HE & DSI & UT
25	No specific help, only general BBC help. So, it not easy to get what you want from the contact page.	BBC KS3bitesize	Whole website	3	HE & DSI & UT
26	Videos have to be closed using the window rather than a close button.	Skoool	Lessons page	1	DSI
27	On certain web pages, there are multiple choices in some videos without any questions.	Skoool	Lessons page	3	DSI
28	The task is limited in some lessons (no test after the lesson.)	Skoool	Lessons page	1	DSI
29	The website does not support user curiosity because it displays lessons in a traditional manner. Also, the website does not provide e-stories, games, simulations, role-playing and activities to hold attention during the lessons or to use during leisure time.	Skoool	Lessons page	2	DSI
30	The search engine does not exist in some websites such as Skoool for Yemen and Egypt.	Skoool	Whole website	2	DSI
31	The website does not provide access to a wide range of resources that are appropriate to the learning context.	Skoool	Whole website	1	DSI
32	The website is not compatible with mobile devices.	Skoool	Whole website	3	DSI
33	There are not enough assessment questions with different difficulty levels.	Skoool	Lessons page	1	DSI
34	The home page link does not exist on all other pages.	Skoool	Whole website	4	DSI
35	In the world map, some Skoool sites do not work, such as in Saudi Arabia.	Skoool	Homepage	4	DSI
36	The website does not update the content constantly (the content is fixed three weeks).	Skoool	Whole website	1	DSI
37	The website does not provide a FAQ page.	Skoool	Whole website	3	DSI
38	Testing results are not displayed, for example, how many scores the user got correctly.	Skoool	Test page	2	DSI
39	There are a lot of pop-up windows when the user selects many lessons.	Skoool	Lessons page	2	DSI

40	All subjects page does not have a clear headline; it just uses, for example, “Grade 10”.	Skoool	Lessons page	1	DSI
41	In the Yemen website, the "التعلم" and "المراجعه“ links should be changed to "الدرس" and "الاهداف والخلاصة" respectively.	Skoool	Lessons page	1	DSI
42	Do not allow the user to customize the video and audio settings.	AcademicEarth	Course page	2	DSI
43	The characters in the videos are real people, so it is not really necessary to use imagination. Also, the website does not support the learner’s cognitive curiosity through surprises, paradoxes, humour, e-stories, games, simulations or role-playing, and deals with topics that already interest the learner (it is just play and watch videos, so users are not able to respond to the program in leisure time). Thus, learner fatigue is not minimized by varying the activities and difficulty levels during learning sessions. The website does not include self-assessment for each module.	AcademicEarth	Course page	2	DSI
44	The search engine is not accurate or in clearly defined positions.	AcademicEarth	Whole website	2	DSI
45	The website does not update the content constantly (the content is fixed for two weeks).	AcademicEarth	Whole website	2	DSI
46	There is no error message on the data entry field in the register page or in the login page. Also, the required fields are not identified.	AcademicEarth	Registration page	3	DSI
47	The user’s current position in the website is not clearly labelled, which leads to user confusion.	AcademicEarth	Whole website	3	DSI
48	There are no ‘back buttons’ to the previous page or the homepage when the user visits certain courses that are provided by universities (i.e. there is a need to leave the site).	AcademicEarth	Whole website	4	DSI
49	The registration and login links are not clearly positioned.	AcademicEarth	Homepage	1	DSI
50	The courses, universities and instructors items are not logically grouped and labelled in the homepage.	AcademicEarth	Homepage	2	DSI
51	The website displays unavailable lessons. Also, there are some courses without materials and videos, and others are for the next academic year (or have already gone).	AcademicEarth	Course page	2	DSI
52	The scrolling down for all pages takes a long time (it is not kept to a minimum).	AcademicEarth	Whole website	1	DSI

53	The similar questions on the FAQ page are not grouped together, which leads to this page being overly long with repeated questions.	AcademicEarth	FAQ page	1	DSI
54	The titles in the data entry screen appear in a small font and without validation (e.g. email) and without support SSL protected.	AcademicEarth	Registration and login pages	4	DSI
55	It is not clear that the user needs to register to get some features (e.g. receive email).	AcademicEarth	Registration page	3	DSI
56	All audios are classified under the video section.	AcademicEarth	Course page	1	DSI
57	Abbreviated names for some universities are mentioned on the list of universities names such as MIT University.	AcademicEarth	Homepage	3	DSI
58	The cancellation link on the login page does not return to the homepage for BBC KS3bitesize.	BBC KS3bitesize	Login page	2	DSI
59	The home link does not work correctly because it is for BBC and not specific to this website.	BBC KS3bitesize	Homepage	4	DSI
60	A 'back button' to previous pages is not provided. Also, there is the same problem with the 'next button'. So, there are poor structures in supporting undoe and redo.	BBC KS3bitesize	Whole website	4	DSI
61	The 'Sign in' link is named 'BBC iD' in the main menu, which is not a familiar name.	BBC KS3bitesize	Homepage	2	DSI
62	The user is not allowed to customize video games.	BBC KS3bitesize	Game page	2	DSI
63	The website displays unavailable lessons (e.g. grammar and spelling).	BBC KS3bitesize	Course page	2	DSI
64	The search engine does not work correctly because it is general for BBC and is not specific to this website.	BBC KS3bitesize	Whole website	2	DSI
65	There is no obvious link for registration or login on the homepage.	BBC KS3bitesize	Whole website	2	DSI
66	The pop-up window asking users to partake in a survey annoys users because it appears many times.	BBC KS3bitesize	Whole website	2	DSI
67	If you want to send a comment by using the 'contact BBC' link, the user needs to visit many windows to submit his/her comments. Also, the link to return to, for example, Explorer or Firefox, is stuck on sending the user to the homepage of BBC KS3bitesize.	BBC KS3bitesize	Contact the BBC page	2	DSI
68	Users do not understand the purpose of the blue footwear mark and green hand pictures that are positioned before the lesson titles (misleading).	Skoool	Lessons page	2	UT

69	There are two languages on the same page (Arabic and English) of the titles index in all modules which caused confusion.	Skoool	Lessons page	1	UT
70	The font size is small.	Skoool	Whole website	1	UT
71	The homepage icon on all the videos is not a direction to the same website (it sends the user to another Skoool website).	Skoool	Course page	1	UT
72	There are different design types depending on the country. This means poor design and style for some countries.	Skoool	Whole website	1	UT
73	The 'Intel Company' icon in the main menu does not stand out from its background (blue).	Skoool	Whole website	1	UT
74	The selected icon is not larger than other icons (or does not appear in distinctive way).	Skoool	Whole website	1	UT
75	The 'logout' link in the main menu is named incorrectly because it works as the link to the homepage.	Skoool	Whole website	2	UT
76	The radio button in the testing page is without a distinctive shape when selected.	Skoool	Test page		UT
77	There is no highlight on the error choices on the test pages, and the error messages are grouped at the end of a page, which means scrolling is often needed to find the errors.	BBC KS3bitesize	Test page	2	UT
78	It is not clear that user needs to register with the discussion board.	BBC KS3bitesize	Whole website	2	UT
79	There are a lot of advertisements on the homepage, which hinder the user's task.	BBC KS3bitesize	Homepage page	1	UT
80	The test link in the revise page is not clearly positioned.	BBC KS3bitesize	Test page	1	UT
81	The check score link in the test page is not clearly positioned.	BBC KS3bitesize	Test page	1	UT
82	The email link is not clearly positioned.	BBC KS3bitesize	Test page	1	UT
83	The registration and login links are not clearly positioned on their pages, and they are not in a distinctive form.	BBC KS3bitesize	Registration and Login	2	UT
84	There is too much content on some pages.	BBC KS3bitesize	Whole website	1	UT
85	There is no visual feedback when questions are selected in the test pages.	BBC KS3bitesize	Test page	1	UT
86	The selected icon is not larger than other icons or does not appear in a distinctive way.	BBC KS3bitesize	Whole website	1	UT
87	The above and bottom menus are not specific for this website. This creates confusion for users.	BBC KS3bitesize	Whole website	1	UT

88	The link to reach more subjects is positioned in an invisible place, and it is labelled with an unclear term (i.e. More Bitesize).	BBC KS3bitesize	Whole website	1	UT
89	The subject menu shows some subjects and hides others, which gives the impression that the website provides six subjects only; the truth is otherwise.	BBC KS3bitesize	Homepage	1	UT

Appendix O: Sets of tasks for Falsification Test for the educational domain

Task 1 < Skoool >	
Task Goal	To examine the reality of the DSI problems in order to identify whether they are real or false problems
User's Tasks	• You are watching a video and you want close it and return to the homepage; what do you do? Did you enjoy the lessons? (Yes, No) Also, you want to use the link to return to the homepage; what do you do? • You have listened to the "القوى الصحيحه الغير سالبه ١" lesson and you want to conduct a test to measure the extent to which you have absorbed the lesson; what do you do? • You are looking for a particular lesson and you want to use the search engine; what do you do? • You want to access a wide range of resources to understand a lesson; what do you do? Do you like this feature? (Yes, No) • You are on holiday and you want to use your mobile phone to access some lessons; what do you do? Did you enjoy the lessons? (Yes, No) • You want to visit the FAQ page to find a solution; what do you do? • You want to visit the Saudi Arabia website by selecting it from the world map on the homepage; what do you do? Did you face any problems? (Yes, No) • After conducting the test, you want to know how many answers you got correct; what do you do? Did you like the difficulty levels of questions? • You want to watch three lesson videos at the same time; what do you do? Did you face any problems in toggling between them? (Yes, No) • You are on the homepage and you find 'Grade 10' terminology; what do you feel about the use of this terminology? • You are in a lesson video and you find these words "التعلم" and "المراجعه" on links; what do you feel about the use of this terminology? • You want to conduct a test after finishing the "مملكة الفطريات - نشاط تفاعلي" lesson; what do you do? • Try to visit the website again after three weeks. Is any updating evident in the content of the website? (Yes, No)
Task 2: < AcademicEarth >	
Task Goal	To examine the reality of the DSI problems in order to identify whether they are real or false problems
User's Tasks	• You are on the homepage and you want to register or login; what do you do? Did you face any problems? (Yes, No) • You want to register and fill in the fields; what do you do? Is it supported by SSL protection? (Yes, No) Did you face any problems? (Yes, No). • You want to use the search engine; what do you do? Did you find it accurate? (Yes, No) • You visit different pages and you want to know your current position in the website; what do you do? Did you face any problems? (Yes, No) • You want to receive emails from the website; what do you do? • You want to visit the FAQ page to find a solution to a problem; what do you do, and what do you feel about having to scroll down? • You visit certain courses that are provided by universities and you want to return to the homepage; what do you do? Did you face any problems in the list of courses or in the university names? (Yes, No) • You want to visit the Massachusetts Institute of Technology; what do you do? Did you face any problems? (Yes, No) • Try to access an audio file for any lesson; what do you do? Did you face any problems? (Yes, No)

	• Try to visit the website again after three weeks. Is there any updating evident in the content of the website? (Yes, No) • You want to enjoy your holiday by watching different videos lessons; what do you feel about the videos in terms of imagination, surprises, paradoxes, etc.?
Task 3: < BBC KS3bitesize >	
Task Goal	To examine the reality of the DSI problems in order to identify whether they are real or false problems
User's Tasks	• You are in the homepage and you want sign in or register; what do you do? Did you face any problems? (Yes, No) • You want to register but you have changed your mind and you want to leave the registration page; what do you do? Did you face any problems? (Yes, No) • You have visited the wrong page, and you want to return to the homepage. You have two options, which are using the home link or back button; what do you do? Did you face any problems? (Yes, No) • You feel you want to join some lessons on a grammar and spelling course; what do you do? Did you face any problems? (Yes, No) • You want to play some games and enjoy your leisure time. Try to customize a video game; what do you do? Did you face any problems? (Yes, No) • Try to use the search engine to find what you are looking for; did you face any problems? (Yes, No) • You want to send a comment by using the "Contact BBC" link and return to the homepage; what do you do? Did you face any problems? (Yes, No)

Appendix P: Sets of tasks for ’Step Two’ in the adaptive framework for the educational domain

❖ Task Scenarios for Social Network Websites

Someone has asked that, over the weekend, you visit any social network website you like; it should offer many services and activities. Kindly surf this website and take 10 minutes to explore and learn more about it, and to practice the Think Aloud protocol.

Task 1: < User Input >	
Task Goal	Mini-user testing to elicit feedback on educational websites from the real users
User’s Task	• Browse the homepage • Click on Register on the homepage, and fill in the fields (if there are any) • Log out • Log in again using your username and password • Click on any link you like (e.g. site map, etc.) • Scroll down any page you are interested in • Go to the homepage • Go to a discussion forum and add, edit and delete your question (if you can) • Go to "Contact us"; fill in a form to contact the Admin and send it • Type key words into "Search" and, using different search criteria, check the results • Try to enjoy some services that you are interested in • Try to customise the setting • Do you have a clear idea about the website? (Yes, No) • Did you face any problems? (Yes, No) • Do you have any recommendations or features that you think should be added to improve the website?

Appendix Q1: The results of Step one ‘Familiarization‘ on the adaptive framework for social network domain

Table 1: Privacy heuristics (Jamal and Cole, 2009)

No.	Heuristics	Explanation
1	Available, Accessible and clear	Make information about the systems activities always available to users and simple to access and understand
2	Correct, Complete and consistent	Ensure that disclosures are complete, correct and consistent for users to make informed decisions
3	Presented in context	Relevant information should be presented for each transaction to minimize memory load and ensure users are aware of consequences of actions
4	Not overburdening	Disclosure must take into account human limitations in memory, ability and interest. Provide succinct and relevant information
5	Meaningful Options	Users need to be given real options rather than opt-in/opt-out when possible to avoid coercion and maximize benefits
6	Appropriate defaults	Default settings should reflect most users’ concerns and expectations about privacy
7	Explicit consent	Avoid assuming consent whenever possible.
8	Awareness of security mechanisms	Users should be provided with enough information to judge security of system and their information
9	Transparency of transactions	Systems should provide transparency of transactions and data use to build user confidence and trust
10	Access to own record	Users should have access to all information the system has collected about them, regardless of source
11	Ability to revoke consent	Consent should be retractable

Table 2: A set of principles concentrating on social and motivational aspects for online communities (Malinen and Ojala, 2011)

No.	Heuristics
1	Facilitate self-presentation and creativity in the service
2	Let the users define the limits of their privacy
3	Create a sense of social presence
4	Facilitate easy participation and content creation
5	Support users’ networking
6	Support different user roles
7	Reward and give recognition
8	Offer the content in a motivating way

Table 3: A set of five heuristics developed by Gallant et al. (2007)

No.	Heuristics
1	Interactive creativity
2	Selection hierarchy
3	Identity construction
4	Rewards and costs
5	Artistic forms

Table 4:A list of recommendations on how to improve the usability of Twitter for first-time users (Owens et al., 2009)

No.	Recommendation	Explanation
1	Make the site language more intuitive	For instance, users were often confused with the meaning of the profile, settings, @username, username links, followers, following, and tweets.
2	Have more distinct visual changes between messaging modes	Currently, the only difference between posting an update, posting a reply, and sending a public message to another user is a label change on the web form.
3	Provide additional hints to first-time users	There are features and terms in Twitter that do not make their function or purpose clear. Hints should be applied to the messaging form to assist users in determining the current mode of communication. They should also be applied to usage of @username terminology and links such as Profile, Settings, etc
4	Replace or use simpler Captchas to ward off spammers	Many participants had difficulty with the Captcha on the sign-up form. It could be improved upon by using one word instead of two, eliminating potentially illegible words, or switching to another type of Captcha

Table5: A list of recommendations on how to improve the usability of MySpace, Facebook, and Orkut (Fox and Naidu, 2009)

No.	Recommendation
1	Use consistent and familiar terminology
2	Provide a brief explanation for terms that are unique to the site (e.g. PhotoCube on MySpace, Testimonials on Orkut, Boxes on Facebook).
3	Provide sufficient feedback to the users. Too often the users repeated failed actions simply because they were not sure if the system had performed their initial task.
4	Improve link placement. Uploading a profile picture, finding the chat link and looking for the Settings option should be easy tasks to perform and should be placed within easy view of the user on the profile homepage.

Appendix Q2: The result of Step two ' User Input' on the adaptive framework for the social network domain

Table1: Results of the context meeting on the social network domain

❖ Interview questions and answers:

- Why was the website developed?

  It was developed to find a sociable environment for users who share similar interests and hobbies. Also, to exchange experiences between them and other experts through email, posts, video calls and chat.

- What are the overall objectives for this website?

  - *Communication*
  - *Business*
  - *Seeking a job*
  - *Consultations*
  - *Games*
  - *Dating*

- How will it be judged that this website is a success?

  1- *Based on the number of members*
  2- *Based on the number of advertisements*

- What are the usability problems in the website that lead to low levels of user satisfaction, efficacy and effectiveness?

  1- *Difficult to join*
  2- *Slow help*
  3- *Unfriendly design*
  4- *Paid services*
  5- *Quality of content*
  6- *Security and privacy*
  7- *Low quality of navigation, structure, and layout*
  8- *Few of services*
  9- *Unlinking between different accounts in different social websites*

- What is the solution for each problem?

  1- *Protect all areas in the website particularly personal information and bank detail.*
  2- *The existing Help Center works constantly and answers any inquiry within 12 hours.*
  3- *Applying the rules of the International Organization for Standardization (ISO) and World Wide Web Consortium (W3C) in terms of the navigation, structure and layout*

- Who are the intended users for this website?

  All kinds of users

- What are users' expected experience and expertise in using the website's main functions?

  The ability to use the Internet, and possession of a computer, laptop, mobile, iPad and their software.

- What tasks do users generally perform when they use the website?

  1- *Updating the personal profile*
  2- *Changing the privacy settings*
  3- *Add friend, accept invitation, revoke invitation*
  4- *Searching for groups and jobs*
  5- *Send post, upload images and videos, and report unwanted posts*
  6- *Start chatting and make video calls*

- What are the users' requirements?
    1- *Provides interactive tools such as videos, chat and games*
    2- *Simple design*
    3- *Website is accessible by mobile devices*
    4- *Secure website*
    5- *Provides privacy feature*
    6- *Provides help centre and FAQ*
    7- *Easy join and use*
    8- *Free services*

Table2: List of usability problems discovered in mini-user testing

No	Problems discovered
1	I found some terminologies are new, and some are not understandable or unfamiliar the first time
2	It is not easy to search for someone because there is not an automatic suggestion for names
3	Font text is small and I cannot change the background colour and font colour
4	I cannot find settings link on the homepage
5	It is difficult to report unsuitable content
6	Some links or buttons are not logically grouped
7	I have no idea how to get to the notification page
8	Help Center link does not work
9	There is no FAQ tool
10	The links' colour or button colour does not change after they have been visited
11	There is an advertisement which covers the page
12	I received the post which sent me to another page and after that I cannot see my wall, I just see porn pictures and videos
13	I received a call from an unknown person and when I answered there was nobody. How he/she contacted me even though he/she is not in my contact list, I do not know
14	Search engine is not accurate
15	I cannot delete the post I have posted wrongly
16	I cannot find the link to create an event
17	I cannot upload some videos because they are long, so I need to use another website to upload and share them
18	Scrolling does not work or gets stuck sometimes; it takes too long to download the older posts
19	Page length is too long
20	I cannot distinguish between online and offline friends
21	I cannot customise my profile
22	It is not clear how to tag videos or images
23	I do not know how to send a private post to someone
24	I cannot upload mp4 video

Table 3:Features that should be added to improve the the social network websites

No	Features	Frequency out of 10 users
1	Easy to post, delete or edite the post, and share	10
2	Easy to follow, send invitation, and revoke invitation	10
3	Simple layout and constancy navigation	7
4	Frequent updating the feeds	5
5	Able to access by using different devices	9
6	Good and accurate search engine	10
7	Support different luguages	3
8	Easy to reporting any problem and quickly get help	6
9	All links are labeled and positioned clearly	10

Table 4:Developed heuristics based on results of mini- user testing

No	Heuristics
1	The vocabulary and terminology should be familiar to the users. Also, the content should be readable, scannable and easy to understand, and free from errors.
2	The search button and input field should be clearly visible and consistently placed across all pages. Also, the features of "SafeSearch" should be switched on.
3	The results of searches should be clear, visible, informative, advisable, relevant, and accurate.
4	The font should be easy to read and the user should be able to edit its colour.
5	The links should be labelled for easy identification, positioned correctly, and have a 'mouseover'.
6	Unlawful, harmful, pornographic and racial content should be easy to report to customer service.
7	All functions should be visible and work effectively.
8	FAQ page should be designed so that it is easy to find.
9	Advertisements should do not disturb the user's primary actions.
10	It should be easy to create events.
11	It should be easy to modify, update and remove posts.
12	These websites should offer an appropriate amount of information for the page length.
13	These websites should provide the minimum number of clickable actions, selections and scrolling to complete one main task
14	These websites should allow tagging, uploading and downloading easily for any videos and photos.
15	Users should be able to use direct chat and messages in a private conversation, and they should be able to broadcast and share messages with other users with whom they are directly connected.
16	Users should be able to customise their profile easily.

Table 5: The result of content analysis regarding to identifying usability problem areas

No.	Areas from users	Strongly agree	Agree	Neither	Disagree	Strongly disagree
1	Terminologies are new, and some are not understandable or unfamiliar	100%				
2	Not easy to search	60%	40%			
3	Font text is small	100%				
4	Cannot change the background colour	20%	60%	20%		
5	Cannot change the font colour			20%	40%	40%
6	Cannot find settings link	100%				
7	Unsuitable content	20%	60%		20%	
8	Difficult to report	60%	40%			
9	Not logically grouped	100%				
10	How to get to the notification page	100%				
11	Link does not work	100%				
12	No FAQ tool	100%				
13	Button colour does not change after they have been visited		40%	40%	20%	
14	An advertisement which covers the page	100%				
15	The post which sent me to another page	100%				
16	Received a call from an unknown person	100%				
17	Search engine is not accurate	100%				
18	Cannot delete the post	100%				
19	Cannot find the link	100%				
20	Cannot upload some videos	100%				
21	Scrolling does not work	100%				
22	Cannot distinguish between online and offline friends	100%				
23	Customise my profile	100%				
24	How to tag videos or images	100%				
25	How to send a private post	100%				
26	Too long to download the older posts	100%				
27	Difficult to join	100%				
28	Slow help	100%				
29	Unfriendly design	100%				
30	Paid services	100%				
31	Security and privacy	100%				
32	Few of services	100%				
33	Low quality of navigation, structure, and layout	100%				
No.	Areas from experts					
34	Consistent design style	100%				
35	Help option for search	40%	40%	20%		
36	Fewer clicks		60%	20%	20%	
37	Pop-ups window	20%	40%		20%	20%
38	Hierarchal layout	20%	405		40%	
39	Response time	100%				

40	Privacy policies	100%				
41	Terms and conditions	100%				
42	Transparency of transaction	100%				
43	Marketing communication		100%			
44	Advertising experience of user	20%	60%	20%		
45	Comment option	100%				
46	Hot offer	20%	60%		20%	
47	Classifying advertisements		40%	20%	40%	
48	Tag people	100%				
49	Access each otgher's profile	40%	60%			
50	User's engagement	20%	20%	20%	40%	
51	Blogs and polls	40%	40%	20%		
52	User's wall	60%	40%			
53	RSS feeds		20%	60%	20%	
54	Multiple chat		40%	20%	40%	
55	Managing the personal profile	100%				
56	Password recovery	100%				
57	Public and privet message	60%	20%		20%	
58	Blocking friend	20%	60%	20%		
69	e-mail notification	20%	40%	20%		20%
60	Universal design	100%				
61	Safe search option	100%				

Appendix Q3: The result of Step three ‘ Expert Input’ on the adaptive framework for the social network domain

Table 1: Summary of focus group results

No	Advice
1	These websites target people from different cultures, so they should use a good colour scheme, appropriate font size with the facility to edit its colour and style, minimum scrolling, effective navigation, and more interactive and sociable tools
2	The critical tool that helps these websites to be successful is the effectiveness and accuracy of their search result. This is because these websites were created to find friends from different cultures and countries who share similar interests
3	The chat feature should be improved by including high-quality video and voice chat. Also, it should be possible to send and receive messages from members even if their status is offline
4	The dating feature in these websites should be visible, effective, safe and secure, safeguarding privacy, and free
5	There should be easy sharing, editing, copy, paste, and deleting in the wall
6	The tool of uploading videos and photos should be simple and users should be able to use different formats of videos and photos without a limit on size
7	The other feature that should be considered in the searching tool is response time, and showing results that are relevant to the search key words.
8	The security and privacy features should be strengthened with the use of different protocols and tools such as password and SSL, particularly for important pages such as profile
9	Familiar terminology should be used
10	Appropriate page length should be used, particularly for showing older posts and the registration page
11	The invitation feature should be a mechanism to avoid unwanted and unsecured invitations, and should include the ability to revoke invitations
12	There should be an easy and available tool to retrieve any deleted messages
13	Customising user profiles is an important feature and these websites should provide a mechanism to their members to carry out this task and avoid users' frustrations. Also, it should allow users to comment in public on other users' profiles
14	These websites should be rich to attract users; they should provide high-quality content and rich services, a simple design that is easy to remember after leaving for period of time and easy to learn. There should be an efficient help centre
15	It should be easy to create groups, join groups, and search for groups
16	Each website should be able to be incorporated with other social media services, for example, when a user posts on his Facebook wall he should be able to see his posts in his Twitter feeds
17	These website should support different languages
18	In the registration page, the required fields should be visible, limited and use a CAPTCHA image with an audio option available. Also, the policy and conditions link should be clearly positioned
19	It should be easy to report any inappropriate content
20	Members should be kept informed by notification through email about new features, and new content feeds, and new polices to increase the content production in these websites
21	It should be easy to access these websites through different platforms such as a mobile phone
22	It is important to make the navigation, all links, and button styles consistent throughout these websites. Also, all the pages should be organised and structured in a similar style to avoid confusing the users.
23	The most important tools and links should be in a list placed at the top of the page.
24	Because these websites produce huge content and feeds, it is important to highlight the important changes to help their users, for example, most viewed, most discussed, favourite feeds and recent updates.
25	With regard to the huge content and feeds, these websites should categorise the content based on users’ preference into primary that is absolutely necessary to show and secondary that can be hidden. In this way, secondary content is only shown on user demand.

26	Because there are a lot of social websites, the content of these websites should be reliable, stable and secure, with guaranteed continuity.
27	Because the content of these websites is updated frequently, the last update statement should be displayed in a prominent place.
28	The prompt messages and notifications should be displayed consistently on the top bar throughout the site.
29	These websites should provide hints to help the user and undo and redo features to avoid errors. Also, the error messages should be displayed in the clear place with different colours.
30	Users should be able to access easily the pages of Help Center, FAQ and Contact Us.
31	Two problems can occur on these websites due to their huge content: response time and loaded memory. Both of these problems cause annoyance to the users. Therefore, the upload-time and the response period for each task should be reasonable and suitable.
32	The protected areas should be wholly inaccessible and the websites should not be allowed to display any personal information outside the site such as 3rd party websites.
33	With regard to the huge content on these websites, adult content should not be allowed to be accessible for users without asking them to declare whether or not they are over 18. In addition, users who are over 18 should not be allowed to solicit personal information from individuals under 18 years old.
34	Advertising is one of the aims of these websites as a source of income. These advertisements should be clear in terms of purpose, also attractive, readable, able to be classified, and users should be able to comment on them.
35	These advertisements should not disturb the users and not cover the content of these websites.
36	It is best to use pop-up windows for advertisements when users want go into them, as this is better than taking them outside of the website page without their permission. Also, avoiding pop-up windows it is very important because they disturb users and cause annoyance.
37	As these websites have a positive impact on business, it is important to allow chat and video calls to take place between multi users.
38	The search tool should be able to be used for groups, people, interests, content, suggestions, and companies. Also, should be easy to edit and to resubmit the search key word.

Appendix Q4: Establishing the DSI method for social network domain

Usability problem area	The adaptive Domain Specific Inspection (DSI)
Layout and formatting (LF)	Design consistency Explanation: -The social network should have consistent and familiar design style for every page, link and button. -Better and consistent navigation among all the pages helps user to work quickly and easily.
	Simple user interface Explanation: -The overall navigation and top-list information (with search and help options and easy bookmarks) facilitate the user in performing any actions in networks. -Minimalistic design (GUI - font sizes, colour) with fewer clicks, scrolling and pop-ups as well as highlighting important features helps users to minimise their memory load. -The social network should have content categorisation and hierarchal information layout (such as primary and secondary) with hidden on-demand content.
Content quality (CQ)	Correct, relevant, up to date and reliable information Explanation: -The social network should only provide available, concise, relevant, reliable, non-repetitive and frequently updated information that is suitable to the page length.
	Error free Explanation: -Error-free environment with consistently prompt messages helps in minimising errors and in preventing users making errors. -Easy and corrective actions (like undo, redo options) help users to rectify errors.
	Representation with familiar terminology & understandable content Explanation: -The social network should provide easy readable and understandable content that is placed in separated blocks. -Content used with familiar vocabulary, terminology and graphical symbols facilitate and ease the tasks of users.
	Appropriate & approachable content Explanation: -Approachable and appropriate amount of information provisioned with FAQ will help users to achieve their primary goal.

	-Content blocks, icons and different colours will help users to take further actions in the social network.
	Site upload time & memory utilization Explanation: -The social network should have negligible upload and response time for performing tasks (suitable to users' cognitive processing). -Heavy and unwanted coding can consume inordinate time and memory.
Security and privacy (SP)	Awareness of security mechanisms/settings & protection Explanation: -Social network systems should have completed all the necessary security tests, and should support all security mechanisms and defined standards such as OWSAP, W3C, etc. -The user's website should protect his/her personal data by using privacy and security settings (SSL). All data should be protected, fully inaccessible or accessible as per authentication.
	Transparency of transactions Explanation: -'Privacy policies' and 'terms and conditions' should be displayed clearly (and be clear) to the user. -Transparency of transactions helps in building and maintaining users' trust (e.g. personal information and uploaded data will not be used or displayed without the user's permission). -Users should be informed for any promotional or marketing communication. -Users should have the facility to report (to customer service or the site manager) any suspicious activity or inappropriate data posted by others. -Users should be aware of the information they have stored within the system.
Business support (BS)	Advertising or sales pitches mechanism Explanation: -How is the overall advertising experience of the user? Is it enjoyable, disturbing or undesirable (like pop-up ads.)? Can the user take part in it (e.g. 'like' or comment option). -Users must be aware of paid membership features, benefits and available hot offers if any. The website should help users in easily classifying advertisements.
	Trust & credibility of information sources and company advertising Explanation: -The social network should maintain users' trust and confidence whilst displaying an advertisement. -An advertisement must lead the user to a trusted site in a separate window.

	-The company must have legal rights to publish their product advertisement.
	Easy to follow & share Explanation: -Social content (text, links, media, etc.) should be easy to upload and organise. -The user should be able to tag people and access each other's profile information as well as share their content with other SNS services.
	Forum/blog facilities and connectivity with different groups/businesses Explanation: -The social network should be designed in a manner that increases the user's engagement. It should provide facilities to take part in forums, blogs, polls and other activities to gather ideas about markets, customers and strategies. -Multimode communication, like mail, free calls, etc., increases users' involvement. -The user must have the facility to create events and different network groups. -Information posted by the user's wall must appear on a fan's wall. The social network website should use a 'crowd- sourcing' approach to stimulate innovation, solve problems and share knowledge.
	Syndication of Web content (such as RSS tools) Explanation: -The social network should provide easy updates on the home page (like friends/family updates). -The social network must support RSS feeds (Web 2.0 features).
	Frequent posting & updating Explanation: -Users must have authentication (modify, update and remover own post and group). -The social network assists the user in participating in various facilities (e.g. posting text, or single or multiple chat). -The website should facilitate the user in participating in posts and in posting as frequently (and as much) as they want.
User usability, sociability and management activities (USM)	Manageable personal profile & user-driven content Explanation: -The social network should facilitate the user with easy registration, managing the personal profile (create, modify) and password recovery options. -Users must have overall control and ease to perform any activity. Users' complaints and reports should also be taken care of. -The user-driven content management website (such as edit/ delete, or liked/marked content) should facilitate the user.

	Easy functionality, participation & user privileges, such as revoking & accepting friends/ connections Explanation: -Easy functionality of social network privileges for users to perform various activities (such as public or private messaging, adding/blocking friends or their connections etc.). -User should have complete freedom to create groups, fan clubs, bands, etc. & to choose the friends, groups, etc. they want.
	Supporter of users' skills & freedom, such as the customization of users' content/messaging and notifications Explanation: -The social network should provide complete freedom to the user to make customizations based on user choice (like creating one's own template or page layout). -The social network should facilitate users in initiating actions (messaging, contents, notifications, etc.) on their profile page. -The social network should use e-mail notifications to encourage members.
	Offers of informative feedback - action & reaction: Explanation: -The social network should provide timely, meaningful, easy to understand and informative overviews (such as current level of achievement or profile status) with action confirmation. -The social network must provide the user's current task-related feedback (e.g. error messages) in an appropriate manner (not too long not or too short). -Users should be provided with the opportunity to access extended feedback from instructors through email and internet communication as well as FAQ. -The website should support the performance tools provided to mimic users' real-world counterparts.
	Appropriate multimedia with complete user control Explanation: -The social network should provide high-quality multimedia with alternative text for visually impaired people. -The social network should provide the user with complete freedom and control over multimedia, which should facilitate the user in performing any action (e.g. edit/post/embedded, view audio/video, or set up their own YouTube or other channel, etc.). -The social network should sometimes provide permission to the user to play videos outside the site (e.g. YouTube), and also to view ratings/comments of video.
	Accessibility and compatibility of hardware devices Explanation: -The social network website must have various platform and hardware

Accessibility and compatibility (AC)	compatibility. Do users required any special computer skills to use the system? -The social network should disable inputs when required. -The social network website must be properly load-tested (allow multiple users at a time) and have a proper Disaster Recovery system. -Easily accessible, multilingual lessons should help users who have physical impairments.
	Accessible path-contact details, help and support Explanation: -The social network website must provide easy accessibility to various assistance options (e.g. FAQ page, help and other additional guidance) as and when the user needs. -The user should be assisted with clear contact details (using multiple contact formats, like email, forms, etc.) and it should resume incomplete work left off. -The social network must have satisfactory performance and be able to load content quickly.
	Easy access through universal design Explanation: -The social network must have a universal design and structure (not too tight, not too loose) to facilitate diversified user groups.
Navigation site and search quality (NS)	Correct & reliable navigation/directions Explanation: -The social network should provide breadcrumbs, and correct and reliable navigation options to facilitate users.
	Easy identification of links and menus Explanation: -The social network website should have well positioned navigation objects, links and menus for ease of work and understanding. -The social network content should have an option to hide or view any content when required.
	Search support & functionality Explanation: -The social network website should facilitate users with clear functions that allow them to conduct any related work without having to leave the current working environment. -Users should have an accurate search engine for basic and advance search support (e.g. groups, people, interests, content, suggestions and companies) with clear and relevant search result pages that allow them to view, edit and resubmit their search. -Moderated or restricted content can be viewed by members with "Safe Search" switched on or off.

Appendix R1: Example on how to develop DSI checklist for social network domain

Usability problem area	Resource	From	The Adaptive Domain Specific Inspection (DSI) Checklist
Navigation site and search quality	Facebook searching	((Alam and Ali (2010))	<u>Heuristic1:</u> Search support & functionality o Are the functionality of buttons and controls obvious from their labels or from their design? o Are there clearly visible search buttons and search input fields consistently placed across all pages? o Are there live search results and filtering? o Does site help to auto fill the search query? o Does the search response are fast enough? o Are the results of searches clear, visible, informative, advisable and relevant? o Does the site support different search criteria (e.g. groups, people, interests, content, suggestions, and companies)? o Does the results page show the user what was searched for, and is it easy to edit and resubmit the search? o Are all the necessary functions of the site available without having to leave the site, and do they work correctly? o Are all the functions clearly labelled, thus facilitating successful completion of the task? Is the status of each task made clear on every page? o Is the search engine accurate? o Does the site support onsite searches within country/region, language, interests, industry, keyword videos, channels, play lists, and groups? Can the moderated or restricted content be viewed by members with "SafeSearch" switched on?
	Quality of navigation, structure, quality and layout	Context meeting interview	
	It is not easy to search for someone because there is not an automatic suggestion for names	Mini-user testing Problems	
	Search engine is not accurate		
	Good and accurate search engine	User testing post questionnaire	
	All functions should be visible and work effectively.	Developed heuristics based on results of mini- user testing	
	The search button and input field should be clearly visible and consistently placed across all pages. Also, the features of "SafeSearch" should be switched on.		
	The results of searches should be clear, visible, informative, advisable, relevant, and accurate.		
	The critical tool that helps these websites to be successful is the effectiveness and accuracy of their search result. This is because these websites were created to find friends from different cultures and countries who share similar interests	Summary of focus group results	
	The other feature that should be considered in the searching tool is response time, and showing results that are relevant to the search key words.		
	It should be easy to create groups, join groups, and search for groups		
	It is important to make the navigation, all links, and button styles consistent throughout these websites. Also, all the pages should be organised and structured in a similar style to avoid confusing the users.		
	The search tool should be able to be used for groups, people, interests, content, suggestions, and companies. Also, should be easy to edit and to resubmit the search key word.		

Appendix R2: Establishing the DSI checklist for social network domain

Usability problem area	The adaptive Domain Specific Inspection (DSI) checklist
Layout and formatting (LF)	Design consistency: o Are all links and button styles throughout the site consistent? o Are all the pages organized/structured in a similar style? o Are the font choices, colours and sizes consistent with good user screen design? o Is the navigation of the site consistent? o Does site has access to Home, Contact Us and other relevant information link on all the pages?
	Simple user interface: o Does the site provide brief, constructive, unambiguous descriptions of the task? o Are the most important items in a list placed at the top? o Does site have search & help option? o Does the site use minimal page scrolling (i.e. the pages are not too long)? o Does the site highlight important changes (i.e. most viewed, most discussed, favourite feeds and recent updates)? o Does the site use glyphs and icons (metaphors) for representation and recognition in a context that is relevant, and not just for decoration? o Does the site use alternative text for the graphics/images? o Does the site categorize content into primary (absolutely necessary to show) and secondary (can be hidden), and show secondary information only on user demand? o Is the site layout, and architecture logical and hierarchical o Does the colour scheme override the content (undesirable)? o Is the site easily readable? o Does the site make important keys larger than other keys? o Are pages easy to bookmark? Is it possible to bookmark a person? o Is a casual user able to return to using the site after some period without having to learn everything all over again? Are all functions and information well-presented and easy to remember? o Is the screen layout efficient and visually pleasing? o Does the site provide the minimum number of clickable actions, selections and scrolling to complete one main task? o Is the site constantly used pop-up windows? o Can users switch between windows during overlapping windows? o Are users allowed to move backward, forward and skip data entry screens among all the pages? o Do all pages have a title? o Does the site helps user to pre populate data during registration, search etc.?
Content quality (CQ)	Correct, relevant, up to date and reliable information: o Is the content updated frequently, the last update statement being displayed in a prominent place? o Does the site display only information that is relevant for its purposes? o Does the site display only the available content, and is the content suitable to the page length? o Does the site provide concise and non-repetitive information? o Is there a link provided to the homepage? Was the site built by a reliable institution? o Are the reliability, stability and continuity of the site content guaranteed?

	Error-free: o Are errors, confirmation, and prompt messages displayed consistently throughout the site? o Is the site free of typographical errors and spelling mistakes? o Do error messages prevent potential errors from happening? o Does the site provide solutions that help the user avoid errors, such as providing 'undo' and 'redo' features? o Can errors be averted or minimized when possible? o Can corrective action be taken to rectify errors? o Are the details of the error messages available with indication to what actions are that users need to take to correct the error?
	Representation with familiar terminology & understandable content: o Is the content readable, scannable and easy to understand? o Do the content blocks need to be visually separated? o Are the vocabulary and terminology used familiar to users? o Does the site provide correct spelling and grammar, and understandable graphic symbols?
	Appropriate & approachable content: o Is the organization of the content suitable for achieving the primary goals of the site? o Are users provisioned with FAQ? o Does the site offer an appropriate amount of information for the page length, and is all the text of a viewable/readable size? o Does the site provide an icon for help next to a field? o Does the site show error in different colour and layout to read easily?
	Site upload time & memory utilization: o Is the site upload-time reasonable? o Is the site free from heavy coding /unwanted scripting which could consume more time/memory? o Are response period suitable to the member's cognitive processing? o Are response period suitable to each task?
Security and privacy (SP)	Awareness of security mechanism/settings & protection: o Are sensitive areas of the site protected against hackers by credentials and SSL security (e.g.,VeriSign™)? o Is it easy to change privacy and security settings? o Does the site protect customers' personal data adequately? o Can the uploaded content still be displayed outside the site if the user decides not to permit it (undesirable)?? o Is the adult content accessible to anyone without asking them to declare whether or not they are over 18 (undesirable)?? o Are users who are over 18 allowed to solicit personal information from under 18s (undesirable)? o Are all protected areas wholly inaccessible? o Does site has taken adequate measure of penetration testing to improve the security? o Does site displays what are the security measure has been taken care to the user? o Does site support industry defined standards like OWSAP, W3C.

	Transparency of transactions: o Is the adopted security mechanism and policy clearly displayed? o Does the site provide transparency of transactions and data use to build user confidence and trust, unless the user gives a clear indication not to expose it? o Are links to 'privacy policies' and 'terms & conditions' clearly displayed? o Is it clearly stated that any data submitted will not be used for other purposes, in order to build user confidence and trust? o Are there processes in place to check the number of memberships or access statistic data? o Should users upload, post, email, transmit or otherwise make available any content that is unlawful, harmful, pornographic and racial, do other users have the option to report any suspicious activity or inappropriate content that breaches the terms of service directly to the customer service or site manager? o Does site provide details like what is the user's information are going to be stored? o Does site declares about sharing the user's information to 3rd party for any purpose? o Does site informs user to contact for promotion, marketing and others such communication?
Business support (BS)	Advertising or sales pitches mechanism: o Is the advertising experience on the site too intrusive, disturbing the user's primary actions? o Does the site have pop-up advertisements (undesirable)? o Does 'multimedia help' make advertising enjoyable/attractive? o Can users leave comments and "likes" (these are social media terms)? o Can users classify advertisements easily? o Do the features of the paid membership are clearly described with giving hot offers?
	Trust & credibility of information sources and company advertising: o Is the user interested in the advertisement characters because they are drawn from the user's own culture? o Does the user have confidence that the site is operating in the way it was designed to?
	Easy to follow & share: o Can users share the content easily (text and links)? o Are the videos and photos easy to upload, download, share, retrieve and organise? o Can users share (i.e. post to friends' profiles) and tag other members in photographs and videos. o Are users able to access each other's profile information? o Are users allowed to share their content with other SNS services?
	Forum/blog facilities and connectivity with different groups/businesses: o Do users become engaged with the site through a set of facilities that are designed to promote engagement (e.g. by creating a group, blog, business)? o Will information posted on users' walls appear on their fans' walls? o Is it easy to create polls, pages and forums? o Are blogs and forums used to get ideas about markets, customers, and strategies? o Is it easy to use site mail to communicate with friends? o Is it allowed to make free calls between computers and/or phones? o Is it easy to create events or select widgets using a calendar? o Can users join regional, educational or workplace networks? o Do websites use 'crowdsourcing' approach to stimulate innovation, solving problem and sharing knowledge?
	Syndication of Web content (such as RSS tools):

	○ Is there a news feed on users' home pages that provides them with friends/ company activity updates? ○ Can users publish RSS feeds to their profiles? ○ Are RSS filters used to create content streams to improve customer relationship management?
	Frequent posting & updating: ○ Are interactive tools such as post text, single chat and multiple chat provided? ○ Is it easy to modify, update and remove posts? ○ Can the users participate as much as they want?
User usability, sociability and management activities	Manageable personal profile & user-driven content: ○ Is it easy to register on the site? ○ In case of theft and/or a forgotten password, is recovery option available? ○ Can customers personalise (customise) their online workplace? ○ Can users edit/delete the content that they have posted? ○ Can users easily collect and access the content that they have found and liked/marked as favourite? ○ Can users create and modify their personal profile, and delete it if necessary? ○ Reporting mechanism: Can users report content that they may have a problem with (such as sexual, religious, illegal, etc.) easily? ○ Can the network delete a content that has received a lot of complaints? ○ Can the user manage all the activities pertaining to the site with ease, and have overall control? ○ Are items logically labelled and grouped in a control panel?
	Easy functionality, participation & user privileges, such as revoking & accepting friends/connections: ○ Private messaging: Can users who are directly connected chat/ message each other in a private conversation? ○ Public messaging: Can users broadcast and share messages with other users with whom they are directly connected? ○ Is it easy to accept new friends and blocking unwanted friends/connections? ○ Can users choose who they want to be directly connected to? This should be a two way agreement - where both users approve of the connection. ○ Can a conversation take place between more than just 2 users? ○ Can users register a group or book or band? Can they create a fan club for a band?
	Supporter of users' skills & freedom, such as the customization of users' content/messaging and notifications: ○ Does the site allow the user to initiate actions? ○ Can users create their own templates or page graphics? ○ Are there enough options for organising page layout or templates? ○ Can users choose a number of applications to be displayed on their profile page? ○ Does a website use e-mail notifications to encourage members? ○ Does site provides customisation based on users choice?
	Offers of informative feedback - action & reaction: ○ Is there confirmation for each action? ○ Is feedback given in proportion to the action performed (not too much and not too little)? ○ Are errors conveyed in context and written in a way that users will understand? ○ Does the site provide an overview of the work process that has been completed by the user (e.g. completing a user's profile)?

	○ Is the feedback given at any specific time tailored to the content or problem being studied by the user? ○ Does the site feedback provide the user with meaningful information concerning their current level of achievement within the program? ○ Is the message of current status related to the user's task? ○ Does the site program provide the user with opportunities to access extended feedback from instructors through email and internet communication, and are adequate FAQs also offered? ○ Does the performance support tools provided mimic their real–world counterparts?
	Appropriate multimedia with complete user control: ○ Are the videos and images on the site of high quality, with the inclusion of alternative text for visually impaired people? ○ Can users change video, audio and image settings easily? ○ Is a mechanism provided to skip/stop animation and video without disruption? ○ Does the site include sound and visual effects, these effects providing meaningful feedback or hints, designed perhaps to stir particular emotions? ○ Does the site include surprises, humour and interesting representations for the user, while avoiding unnecessary multimedia representations that could confuse a user who has just started to work with the site? ○ Is there unnecessary animation and 'flash' on the site (undesirable)? ○ Is it easy for users to set up their own channels (e.g. YouTube channels)? ○ Are video ratings and comments available on the site? ○ Can users modify photo, audio and video submissions? ○ Are users allowed to play videos outside the site (e.g. YouTube) which would mean that they could be 'embedded' into other websites?
Accessibility and compatibility	Accessibility and compatibility of hardware devices: ○ Is the site compatible with various platforms and hardware, and can its features be adapted to individual user preferences? ○ Do potential users have to have special computer skills to be able to use site? ○ Are all the input devices/buttons that have no function disabled to prevent user-input errors? ○ Are the lessons accessible to users with physical impairments, and their contents available in various languages? ○ Does the site is properly load tested and support agreed number of users at a time. ○ Does the site have proper Disaster Recovery in place? ○ Does the site is supported by text reader or other such devices?
	Accessible path-contact details, help and support: ○ Is a site map and /or table of contents available, as well as a calendar? ○ Is there accessible and appropriate help available on demand? ○ Does the site provide clear contact details, using multiple contact formats (email, forms, etc.)? ○ Is the FAQ page easy to find? ○ Is everything on the site clearly understandable by the user, including how to access options for additional guidance (chatting, editing, adding, seeking instruction or other forms of assistance) when needed? ○ Does user allowed to resume work where they left off after getting help? ○ Does the performance of the site is satisfactory and it loads most of the content in less than a second?
	Easy access through universal design: ○ Has a universal design been implemented to cater for diversified user groups? ○ Is the structure too tight (strangling) or too loose (lacking cohesion), both of which are undesirable?

Navigation site and search quality	Correct & reliable navigation/directions: o Do all links and buttons lead to the correct location? o Does the site provide a breadcrumb (cookie crumb trail) to identify the path to the current location? o Does the site match the menu structure to the task structure, and can the user distinguish between options and content on the pages?
	Easy identification of links and menus: o Are the navigation objects and tools placed in consistent, clearly defined positions, and are they of an adequate size? o Are icons and links labelled? o Is an item still visible when it should be hidden from view, and vice versa? o Are the menus straightforward and easy to understand, the items being logically grouped and labelled? Do buttons, links and features have a 'mouseover' or pop-up window that provides meaningful feedback?
	Search support & functionality: o Are the functionality of buttons and controls obvious from their labels or from their design? o Are there clearly visible search buttons and search input fields consistently placed across all pages? o Are there live search results and filtering? o Does site help to auto fill the search query? o Does the search response are fast enough? o Are the results of searches clear, visible, informative, advisable and relevant? o Does the site support different search criteria (e.g. groups, people, interests, content, suggestions, and companies)? o Does the results page show the user what was searched for, and is it easy to edit and resubmit the search? o Are all the necessary functions of the site available without having to leave the site, and do they work correctly? o Are all the functions clearly labelled, thus facilitating successful completion of the task? Is the status of each task made clear on every page? o Is the search engine accurate? o Does the site support onsite searches within country/region, language, interests, industry, keyword videos, channels, play lists, and groups? o Can the moderated or restricted content be viewed by members with “SafeSearch” switched on?

Appendix S: The three methods' performances in discovering usability problems for the social network domain

No:	Usability problems discovered	Website	Area	Severity rating	Method
1	It uses unfamiliar words that do not explain what is meant, such as "limited", which is the link next to the post link (also, Hangouts, Poll).	Google+	Homepage	2	HE
2	It is not obvious how to cancel, edit… etc. the post.	Google+	Post window	2	HE
3	Adding people to your circles from the right rail does not offer an undo.	Google+	People page	2	HE
4	Lack of ability to invite a big group of people to an event; for example, to invite 30 people, the user needs to type them all in manually.	Google+	People page	2	HE
5	'Add people' icon on the homepage looks like a chatting service.	Google+	Homepage	1	HE
6	The problem that I faced at the time was the 'upload image' option; I selected an image and started the creative toolkit but in the toolkit window that image did not appear and there was no option available to upload it again.	Google+	Photo page	2	HE
7	On 'Photos', I expected to be able to click on the CTAs.	Google+	Photo page	1	HE
8	Not completely apparent on HOME, as the 'posts' widget is overpowering.	Google+	Homepage	1	HE
9	Using the Chat link is better than the unclear icon.	Google+	Homepage	1	HE
10	The breadcrumb-style heading looks clickable but it is not clickable.	Google+	Homepage	1	HE
11	The menu is not associated to the context of the breadcrumb heading; it is a universal menu for Google+.	Google+	Homepage	3	HE
12	When the user is selecting any other language, for example Arabic, half of the information comes in English by default.	LinkedIn	Homepage	3	HE
13	It cannot see the list of my profile visitors.	LinkedIn	Profile	2	HE
14	Colour of visited links does not change,	LinkedIn	Whole website	1	HE
15	'Imported contacts' link is a foreign term to me. I did not import these contacts from anywhere, but LinkedIn has classified these contacts as such. This is confusing; I thought all of these folks were already connected to me.	LinkedIn	Hangout	2	HE
16	'Grow your network' and 'My Connections' options need to be qualified. 'Grow your network' should include a description list: 'your contacts and their contacts'.	LinkedIn	Homepage	1	HE

17	The presence of the link 'Remove Connections' is not obviously positioned and is unclear in meaning. Does it remove selected connections or all connections?	LinkedIn	Connections	2	HE
18	There seems to be a disconnection. Perhaps this area should have just one link: 'Edit Connections'. The 'Add Contacts' screen can be combined with 'remove connections' to reduce cognitive friction.	LinkedIn	Connections	1	HE
19	Too many functions and links in the homepage which cannot be remembered.	LinkedIn	Homepage	2	HE
20	Continues to ask to accept the same contacts even after clicking the 'accept' button and refreshing the page.	LinkedIn	Homepage	2	HE
21	Not all icons are labelled clearly, e.g. 'Help Centre'.	Ecademy	Whole website	2	HE
22	The website uses abbreviated words which are not familiar to users, such as 'Msgs'.	Ecademy	Homepage	3	HE
23	The zones are separated but not clear.	Ecademy	Whole website	3	HE
24	The most important items in a list are placed at the top.	Ecademy	Homepage	1	HE
25	The most important features are not positioned on the top bar, e.g. 'settings' and 'help'.	Google+	Homepage	3	HE & DSI & UT
26	When the user chooses 'Post visibility' from the window for creating a post, and after the post is created, there is no option to change it (e.g. from public to specific people).	Google+	Post window	3	HE & DSI & UT
27	Page is too long and should be linked by the top icon after each paragraph.	Google+	Homepage	1	HE & DSI & UT
28	Breadcrumbs to identify the path to the current location are not provided.	Google+	Whole website	3	HE & DSI & UT
29	No 'Site map' and/or 'table of contents' available	Google+	Homepage	1	HE & DSI & UT
30	It is not clear how to report unwanted posts or content.	Google+	Homepage	2	HE & DSI & UT
31	Sharing content (text, video, photos and links) is available but not clearly marked, particularly for novice users.	Google+	Homepage	1	HE & DSI & UT
32	It is difficult to change privacy and security settings (if this service exists).	Google+	Homepage	2	HE & DSI & UT
33	It is not clear how to upload a CV.	LinkedIn	Job	2	HE & DSI & UT
34	It is not clear how to add a tag when creating a post.	LinkedIn	Homepage	2	HE & DSI & UT
35	The terms of a basic account say that you cannot see who has viewed your profile. So, this link should be disabled if the user has a basic account, but the website does not do so.	LinkedIn	Profile	1	HE & DSI & UT

36	The 'Recent Activity' link is not clearly positioned. It should be positioned on the homepage or it should be listed under the profile tab (drop-down menu).	LinkedIn	Homepage	2	HE & DSI & UT
37	The 'Ask to be recommended' link is not clearly positioned.	LinkedIn	Profile	2	HE & DSI & UT
38	There is no site map for this website.	LinkedIn	Homepage	2	HE & DSI & UT
39	It is not clear how to delete a comment from someone's post.	LinkedIn	Homepage	1	HE & DSI & UT
40	It is not clear how to stop endorsements for skills.	LinkedIn	Profile	1	HE & DSI & UT
41	There is no button to remove the summary in the profile page.	LinkedIn	profile	1	HE & DSI & UT
42	The 'Help Forum' page displays this message: 'Oops! We weren't able to retrieve Help Forum results. Please try again later.'	LinkedIn	Help Forum	2	HE & DSI & UT
43	'Add Media' button/icon does not work on the profile page.	LinkedIn	Profile	3	HE & DSI & UT
44	Borders are not used to identify meaningful groups.	Ecademy	Whole website	1	HE & DSI & UT
45	There is no identified link for going back from page to page.	Ecademy	Whole website	2	HE & DSI & UT
46	Links are not eye-catching.	Ecademy	Whole website	2	HE & DSI & UT
47	Many pages start without headers, e.g. Inbox and Settings.	Ecademy	Inbox	1	HE & DSI & UT
48	It is not clear that this icon '+' on the photo page means create an album (until I move the mouse over it).	Google+	Homepage	1	UT
49	It is not clear that if you click on a photo, this icon '⋮' does not mean 'download' until I move the mouse over it and see the text 'download'.	Google+	Homepage	1	UT
50	The 'Albums' link is positioned in an unobvious area; it is hidden under 'More' when you click on 'Photos' in the main menu.	Google+	Photos	2	UT
51	The colours of all the links are not clear; they are similar to the background colour, and thus they are not eye-catching.	Google+	Whole website	1	UT
52	Font size is too small.	Google+	Whole website	1	UT
53	The error messages on the registration form should appear above of the text field.	Google+	Registration	1	UT
54	Required fields are not identified in the registration form.	Google+	Registration	3	UT
55	It is not clear how to send an email to a particular friend.	Google+	Connections	3	UT
56	It is not clear how to save a backup of your photos or profile information.	Google+	Setting	3	UT

57	To create an account, users have to receive a verification code from website. However, I did not receive a verification code.	Google+	Registration	4	UT
58	The hyperlink on the Google page to visit Google+ is not clearly positioned.	Google+	Homepage	3	UT
59	It is not clear what '+1' means; is it the same as 'like'?	Google+	Homepage	2	UT
60	The search field is not clearly positioned. It looks like the search field for the Google search engine. It should be positioned in the top bar with other links.	Google+	Homepage	1	UT
61	Location support is needed to show the city location of Google+ friends.	Google+	People/ search	2	UT
62	The 'Follow/Unfollow' link is not clearly positioned in the collection page. Also, it should be of a different colour to the background.	Google+	Collection	1	UT
63	The search field is not specific for the collection page. So, users would not go through collections to find something interesting.	Google+	Collection	2	UT
64	The 'suggested people' link is not clearly positioned.	Google+	People	2	UT
65	There is no refresh button on the profile page because when a link is shared, it should refresh the screen completely to see the post in the profile page.	Google+	Profile	2	UT
66	The 'Start a Video Hangout' link is not clearly positioned.	Google+	Hangout menu	3	UT
67	There is no exit button on the Tour window.	Google+	Tour window	1	UT
68	There are no options in the 'Relationship status' menu to choose Girlfriend, Boyfriend, Engaged, Married partner.	Google+	Profile	1	UT
69	In the 'Chat' and 'Video Hangout' windows, there is no information to identify who is online or offline.	Google+	chat / Hangout windows	2	UT
70	There are two Home clickable icons on the top bar which have the same functionality, so they confuse users ().	Google+	Homepage	1	UT
71	It is difficult to reach the 'My Account' page to manage your account access and security settings.	Google+	My Account	4	UT
72	There is no back link to return from the photo page to the homepage.	Google+	Photo	2	UT
73	There is no search box in Google+ Pages allowing users to search for specific words from the all the previous posts in that Google+ page.	Google+	Homepage	3	UT
74	The drop down menu that is next to the search field is not visible.	LinkedIn	Homepage	2	UT

75	The 'Jobs' tab does not work.	LinkedIn	Homepage	4	UT
76	Required fields are not identified in the 'Join' form.	LinkedIn	Join	4	UT
77	The above links for 'Add education' and 'Add publication' do not appear (hiding) until the user moves the mouse over the same section.	LinkedIn	Profile	2	UT
78	The bottom buttons ('Add education', 'Add publication' and 'Add position') in each section are not clear because they do not have borders and their colours are the same as the background colour. Also, they do not have a plus (+) symbol.	LinkedIn	Profile	2	UT
79	Many questions need to be answered for 'Profile Strength'.	LinkedIn	Profile	1	UT
80	The 'managing public profile setting' link is not clearly positioned.	LinkedIn	Profile	2	UT
81	The LinkedIn link image to go back to the main page on the forgotten password page is not clickable (as users thought).	LinkedIn	forgotten password	3	UT
82	The error message, when the user enters an incorrect email in the 'Forgotten password' page, is not clearly positioned.	LinkedIn	forgotten password	2	UT
83	The drop down menu, which has the user's photo on the right of the Homepage on the above bar, confuses the user because it is clickable and it works with the same functionality as the 'Profile' link.	LinkedIn	Homepage	2	UT
84	It uses unfamiliar words, which results in users stopping their task in order to add connections such as using 'Keeping in Touch' (which means your contacts or connections).	LinkedIn	Homepage	2	UT
85	When the user creates an account, LinkedIn asks the user in Step 2 to add contacts. If the user does not have any contacts to add, an error message appears. After LinkedIn has failed to add any contacts, the message is not clear in terms of information and colour.	LinkedIn	Join	2	UT
86	It is not clear how to turn off the 'suggested people' feature.	LinkedIn	Homepage	3	UT
87	When clicking on the 'Privacy & Settings' link, it asks to login again. It is boring.	LinkedIn	Homepage	2	UT
88	It is not clear how to revoke a recent connection.	LinkedIn	Connections	3	UT
89	'Terms and conditions' link is not clearly positioned.	Ecademy	Homepage	1	UT
90	The registration form is too long, which leads to user frustration.	Ecademy	Registration	1	UT
91	The Homepage link is not easy to find.	Ecademy	Homepage	1	UT

92	Settings link is not clearly positioned.	Ecademy	Homepage	1	UT
93	The aim and meaning of the 'SEO' link on the top bar is not clear.	Ecademy	Homepage	3	UT
94	It is not allowed to send messages until the user has upgraded the membership.	Ecademy	Message	3	UT
95	The 'Suggested people' link does not work.	Ecademy	Homepage	4	UT
96	The required fields to create a blog are not identified.	Ecademy	Blog	1	UT
97	The 'Follow' link for a blog or a company is not clearly positioned.	Ecademy	Blog	1	UT
98	Some features do not work properly on mobiles (e.g. Forgotten password).	Ecademy	Mobile	4	UT
99	Some companies appear on the search results without any information.	Ecademy	Homepage	4	UT
100	The links are not grouped; they are just listed on the two above bars.	Ecademy	Whole website	1	UT
101	The Ecademy logo on the homepage is clickable, but when the mouse moves over it, it does not show that it is clickable.	Ecademy	Homepage	1	UT
102	The icon to connect to Twitter does not work.	Ecademy	Homepage	1	UT
103	It does not support searching for people based on their email.	Ecademy	Homepage	1	UT
104	There is a limit to adding people to your circles.	Google+	People	2	DSI
105	There is no way to filter poll posts in communities or anywhere.	Google+	Poll	2	DSI
106	To change your name in Google+, you can be permitted 3 name changes in a 90-day period from the first change.	Google+	Profile	3	DSI
107	There are no links for reaching different pages in the help forums on Google+, such as Business forum and Hangout forum.	Google+	Whole website	4	DSI
108	Google+ does not support any games.	Google+	Whole website	3	DSI
109	There is no sync between the desktop and mobile applications in terms of notifications. So, when a notification is read on the desktop, it still appears as unread on the mobile application; it should be amended on the mobile.	Google+	Notification	3	DSI
110	Google+ claims that it provides strict controls to block spam post. However, many spam posts (e.g. porn) still appear many times, even after reporting them.	Google+	Homepage	4	DSI
111	If you turn off notifications when using a mobile, you still receive emails. So, you have to turn off the notifications from the computer.	Google+	Notification	3	DSI
112	When you get blocked from a community, you can still post, but you cannot comment. A blocked user	Google+	Communities	2	DSI

	should be prevented from both posting and commenting.				
113	The Google+ logo is clickable and is positioned above the 'Home' icon but both are doing the same job. It is confusing.	Google+	Homepage	1	DSI
114	It is not clear how to change the language.	Google+	Setting	3	DSI
115	When you change the language to Arabic, some links still appear in English as a default.	Google+	Homepage	2	DSI
116	It is not clear how to connect Google+ to another social network such as Twitter or Facebook.	Google+	Homepage	3	DSI
117	Invalid URL format is accepted to post.	Google+	Post	2	DSI
118	URL field is not required in 'share a link'.	Google+	Share	2	DSI
119	There is no option to search by using a phone number, because even using email, Google+ does not give the person's profile correctly.	Google+	Profile	2	DSI
120	If you have a collection and you want to delete this collection, but you want to move the posts before deleting the collection, there is no option to move or save these posts.	Google+	Collections	3	DSI
121	If you want to post something and you want to attach photos, and when you select photos, another window appears which asks you to reselect the same photos from the same album. This task is then repeated (twice), which makes this feature inefficient and frustrating.	Google+	Photos	2	DSI
122	In the polls, you cannot add more than one photo in one question. It should be allowed to add multiple images.	Google+	Poll	2	DSI
123	There is no option to get a notification from an individual user as soon as he/she posts something; this notification is only available for Circles.	Google+	Notification	2	DSI
124	Anyone can add you into his/her circles without your permission, and you cannot delete yourself from his/her circles.	Google+	People	4	DSI
125	If your friend adds you on his circles, and he creates a new community, his post in the community will post on your wall. However, you are allowed to re-share it, but you cannot comment on his post until you join his community. So, why is his post added to your wall when you have not joined his community?	Google+	Community	3	DSI

126	It is confusing; there are many help forums, e.g. business forum, Google+ forum, hangout forum, etc.	Google+	Help Forum	3	DSI
127	If you have a problem in hangouts, Google+ cannot help you because a hangout is a separate product and it has its own Help Forum.	Google+	Help Forum	3	DSI
128	It is allowed to post only a limited number of questions (4 times) per day in the Google+ help community. This means that if you have more than four questions, you have to wait until the next day.	Google+	Help community	2	DSI
129	If you have forgotten your password and have tried to recover it but cannot, you should ask for help from Gmail Product Forum, not from Google+. So, what is the benefit of the Google+ help centre?	Google+	Forgot password	4	DSI
130	If anyone gives a post +1, and then cancels his/her +1, it still shows up on the notifications button. This is an incorrect notification.	Google+	Homepage	1	DSI
131	There is no option to stop receiving notifications when commenting on a post. When you do this, you will get a notification for all the persons who comment after you. Google+ should modify this feature to notify only those persons who tagged you by name.	Google+	Notification	2	DSI
132	If you want start a business and you want to create a web page and URL, when I typed in a unique name, a message appears: 'Many people have the same name'; however, when I searched the name, I could not find anyone using my specific name.	Google+	Business	3	DSI
133	The website does not prevent or recover from errors such as adding a friend in two different circles.	Google+	People	2	DSI
134	There is tendency to shuffle the order of the contents on the homepage feed. Every time I visit this website, I see the same posts but in a randomly different order.	Google+	Homepage	2	DSI
135	The single-column view does not work on the communities' page, which does not make very good use of screen space.	Google+	Homepage	2	DSI
136	The option of 'Automatically enhance new photo' is not available at the time of upload. To get it, it needs to go to the settings account.	Google+	Photos	3	DSI
137	The site does not provide clear contact details.	Google+	Contact	4	DSI
138	It is not clear how to change themes.	Google+	Setting	1	DSI

139	There is no option to change the colour of the post.	Google+	Setting	1	DSI
140	The search functionality is not accurate (e.g. if I entered 'events Austin Film festival', it shows everything).	Google+	Homepage	2	DSI
141	This website does not provide a chat service.	LinkedIn	Homepage	3	DSI
142	Receiving fraudulent email from LinkedIn: 'Upgrade to Free Premium Account', but when you set up your credit card information; they deduct the money without permission.	LinkedIn	Email	4	DSI
143	Not all posts on LinkedIn appear on the Twitter wall, and sometimes it takes a long time to appear.	LinkedIn	Sharing	2	DSI
144	There is no option to allow the user to get notifications about recent discussions on the help forum.	LinkedIn	Notifications	3	DSI
145	It is difficult to remove credit card details from an account.	LinkedIn	Setting	3	DSI
146	It is not possible to undo mistakenly flagged posts, and there is no option to see all posts that have been flagged.	LinkedIn	Post	3	DSI
147	Post and discussion are two terminologies but they are functionally the same. So, it can post a discussion in a group forum and also post a comment on the wall. Thus, one term should be used for both functions (inconsistencies).	LinkedIn	Homepage	2	DSI
148	There is a pop-up window that appears on the group page which cannot be closed; it asks you to follow another discussion group.	LinkedIn	Whole website	2	DSI
149	The bookmarking feature is not provided, e.g. in articles that are published on Pulse.	LinkedIn	Pulse	3	DSI
150	It does not support MP4 video format.	LinkedIn	Post	2	DSI
151	It is not clear how to remove/edit the pay details.	LinkedIn	Whole website	2	DSI
152	There is no link for RSS or a search link for RSS on the homepage.	LinkedIn	Homepage	2	DSI
153	The sharing button does not show when you update a post and want to share it with your group.	LinkedIn	Sharing	2	DSI
154	It is not possible to return to your answer on the Forum if it is more than 18 months from the date of publication.	LinkedIn	Forum page	2	DSI
155	The link 'Help Forum' is not positioned visibly in the top line of the menu.	LinkedIn	Homepage	3	DSI
156	The time for posting an advertisement by moderators is longer than expected. It takes five days, and it should take only two days.	LinkedIn	advertising	1	DSI

157	It is not allowed to advertise a picture if its size is more than 50 x 50.	LinkedIn	advertising	1	DSI
158	There is no option to post advertising like video or animation.	LinkedIn	advertising	1	DSI
159	It is not obvious how to remove an advertisement from the homepage.	LinkedIn	Homepage	2	DSI
160	There is no link to report spam on the wall.	LinkedIn	Homepage	3	DSI
161	Important features need to be paid for to get them to work, such as sending a message to someone not in your connection.	LinkedIn	Message	2	DSI
162	There is no option to move a message in the archive folder to the inbox folder.	LinkedIn	Inbox	1	DSI
163	The admin of a company cannot use the 'like' link or the 'share' link on any post or news for the company.	LinkedIn	Homepage	1	DSI
164	There is no option to search by date for finding a company that is a recently founded business.	LinkedIn	Search	1	DSI
165	Cannot upload photos, which means that there is something prohibiting this functionality.	LinkedIn	Profile	3	DSI
166	The FAQ is not available on the homepage, and there are some questions on the Help Forum link.	LinkedIn	Whole website	3	DSI
167	There is no option to stop sending 'invite' messages to contacts during initial registration.	LinkedIn	Registration	2	DSI
168	The 'Contact us' link in the Help Centre page (which has no functionality) is disabled.	LinkedIn	Homepage	3	DSI
169	It is difficult to find the 'Account Setting' link to customise settings.	LinkedIn	Homepage	2	DSI
170	Some menus are not understandable or straightforward (e.g. the Pulse link).	LinkedIn	Homepage	1	DSI
171	Home page is too long, which needs a great deal of scrolling.	LinkedIn	Homepage	1	DSI
172	Some pages require much scrolling (e.g. 'New messages' page contains too many messages, need a lot of scrolling).	Ecademy	Messages	1	DSI
173	The FAQ is not available.	Ecademy	Whole website	2	DSI
174	Registration error messages appear at the top of the form rather than at the top of the text box.	Ecademy	Homepage	1	DSI
175	Search results are not accurate, e.g. searching for UEA.	Ecademy	Search	1	DSI
176	Breadcrumbs are not clearly available; this is especially important for novice users.	Ecademy	Whole website	1	DSI
177	Reporting mechanism (e.g. deleting content that has a lot of flags reported)	Ecademy	Homepage	2	DSI

	is not available or if it is, it is not clear or positioned clearly.				
178	Site map is not available.	Ecademy	Homepage	2	DSI
179	There are too many pop-ups.	Ecademy	Whole website	1	DSI
180	Unable to find language option on home page for diversified users following different languages.	Ecademy	Homepage	1	DSI
181	Too many advertisements on the homepage.	Ecademy	Homepage	1	DSI
182	'Lost password?' link should be called 'Forgotten password'.	Ecademy	Join	1	DSI
183	Too many links and functions on the Homepage.	Ecademy	Homepage	1	DSI
184	E-news section is not well-organised.	Ecademy	Homepage	1	DSI
185	The distance between 'members online' and the drop down list is big, which confuses users.	Ecademy	Homepage	1	DSI
186	The links 'Hide message', 'See all', and 'More' should be grouped.	Ecademy	Homepage	1	DSI
187	Too much content on some pages.	Ecademy	Whole website	1	DSI
188	Many buttons for 'Join now' on the page.	Ecademy	Join	1	DSI
189	It does not clear what is meant by 'Blackstar' on the group page.	Ecademy	Search	2	DSI
190	The required fields in the registration form are not identified.	Ecademy	Registration	3	DSI
191	There are limited features for users who have created an account: create a profile, search for contacts and respond to messages on the website.	Ecademy	Whole website	1	DSI
192	It is not clear that those users who have a basic account can only respond to messages; they cannot send messages until they upgrade their membership.	Ecademy	Whole website	2	DSI
193	The website does not support properly surfing by mobile.	Ecademy	Mobile	3	DSI
194	The calendar link on the 'Events' page is in a small font.	Ecademy	Events	1	DSI
195	Terms and conditions for creating a blog are not found on the blog page.	Ecademy	blog	2	DSI
196	Too many confusing fields to complete in the registration form for a company.	Ecademy	Company	1	DSI
197	Link colours should be standardized.	Ecademy	Whole website	1	DSI
198	It is not clear how to close an account.	Ecademy	Setting	2	DSI
199	Not all icons are labelled clearly, e.g. 'Help Centre'.	Ecademy	Whole website	2	DSI
200	The website uses abbreviated words which are not familiar to users, such as 'Msgs'.	Ecademy	Homepage	2	DSI

Appendix T: Sets of tasks for Falsification Test for the social network domain

Task 1 < Google+ >	
Task Goal	To examine the reality of the DSI problems in order to identify whether they are real or false problems
User's Tasks	• You want make a video call; what do you do? • You want create a vote to choose the best place for a barbecue; thus, you want upload two images for two places. What do you do? • You want send a post to a specific friend; what do you do? • You find a mistake in the post, and want edit it; what do you do? • You want send a post by using an invalid URL; what do you do? • You want to add someone to your circles. Once you have added him, you want to undo this action; what do you do? Also, try to add him again into two groups of circles; how does his process make you feel? • You want to create a family circle and to add 50 people to this circle; what do you do? • Someone has added you to their circles but you do not want to be there; what do you do? • You want to create an event in order to invite 40 people; what do you do? • You want to upload a photo and start using the creative toolkit; what do you do? • You want to choose another language; what do you do? • You want to know who has visited your profile page; what do you do? • You want to filter poll posts in communities; what do you do? • You want to change your name in Google+ and you have three names, so you want try each name and see how your profile page looks; what do you do? • You have five problems: one in your Business page, one in the Hangout page and three in Google+ page. So, you want to find solutions for these problems; what do you do? • You want to play your favourite game (such as Clash of Clans) in Google+; what do you do? • You have received a post from a friend. Try reading it from your desktop, and after that try to use your mobile to check the homepage icon for a notification; does it still show that there is one post marked as unread? • Try to change the notifications from your mobile so as not to allow receiving an email notification. After that, we will send a post to you; have you nevertheless received an email notification? • You want to go back to the homepage. Try to use the Google+ logo and the Home button on the drop down menu; what do you feel? • You want to connect your Google+ account to another social network account; what do you do? • You want to send a post link for a new movie. In the post's window, you click on the 'share' button without adding that link; what do you feel? • You want to search for a friend so as to add him. Use his email and his phone number to find him; what do you do? • Create a collection and sent many posts. After that, you decide to close this collection but you want move or save these posts before closing it; what do you do? • Someone from your circles has created a new community. You see one post in his community on your wall, and thus you want to comment on this post; what do you do? • You want to create your own business. You want choose a name for your business, so what do you do? • You want to turn on the 'Automatically enhance new photo' option; what do you do? • You want to change the page theme and colour of your post; what do you do? • You answered a question on the Google+ forum. How can you stop receiving notifications from people who comment after you; what do you do?

Task 2: < LinkedIn >	
Task Goal	To examine the reality of the DSI problems in order to identify whether they are real or false problems
User's Tasks	• You want create an account but you do not want to send invite messages to your contacts; what do you do? • You want to import some contacts, so you use the 'Imported contacts' link; what do you do? • You want to add connections and you have these links: 'Grow your network' and 'My Connections'. So, which one you will use, and what do you feel? • You want to edit the connection, so use the 'Edit Connections' link. What do you feel? • You want to remove a connection, so use the 'Remove Connections' link to do it; what do you feel? • On the homepage, you find notifications for accepting some invitations. You have time to do this, so what do you do? • You have a problem and you want to chat with a friend who may be able to help. What do you do, and what do you feel? • You have connected your account to, for example, Twitter. Thus, you want to post in LinkedIn and for it to appear in Twitter. What do you do, and what do you feel? • You want to receive notifications about the recent discussions on the Help forum; what do you do? • You want to edit your payment card information; what do you do? After that, you want to remove that card information; what do you do? • You have mistakenly flagged a post, so you want to undo this action. Also, you want to see all posts that have been flagged; what do you do? • You are in the Pulse page. You find an interesting article, and you want to use the bookmarking facility to read it later; what do you do? • You want to send a post as an MP4 video; what do you do, and what do you feel? • You want to receive RSS news from an interesting company; what do you do? • You want to update an image and share it with your group; what do you do? • You have a problem, and you got an answer to this problem one year ago. Now try to return to your answer in the Help Forum; what do you do? • You have an image as advertisement (50 x 50) and you want post it; what do you do? After that, you want to remove this image and post a video or animation file; what do you do? • You have received a post on your wall but you want to report it; what do you do? • You find an old friend and you want to send a message before asking for a connection; what do you do? • You want to move a message from the archive folder to the inbox folder; what do you do? • You want to search for a company that is a recently founded business; what do you do? • You have a problem and you want to search for an answer in the FAQ; what do you do? • You want to customise your settings; what do you do?
Task 3: < Ecademy >	
Task Goal	To examine the reality of the DSI problems in order to identify whether they are real or false problems
User's Tasks	• You want to create an account in this website but, during the registration process, try to include an incorrect or invalid input (such as leaving a required field empty); what do you feel? • You have a problem and you need help, so you want to visit the Help page; what do you do? • You want to send a message; what do you do?

	• You have a problem and you want to search for an answer in the FAQ; what do you do? • You want to report a post; what do you do? • You want to visit pages that are interesting to you, so you are looking for a quick and easy way to navigate between these pages (e.g. site map); what do you do, and what do you feel? • You want to change the website language; what do you do? • You have forgotten your password; what do you do? • You want to send a message to your friends but you want to know who is online; what do you do? • You want to create your own blog but you want to know the terms and conditions; what do you do? • You want to join the latest blogs who have 'Blackstar'; what do you do, and what do you feel? • You want use your mobile to surf this website; what do you do, and what do you feel? • You want to register your company; what do you do, and what do you feel? • You want to create your own event; what do you do, and what do you feel? • You want to close your account; what do you do?

Appendix U1: Response from BBC website on the problem report

Dear Audience Member Reference CAS-1113745-6DLGHX

Thanks for your enquiry.

As the BBC is a public corporation, financed by the licence fee, its income must be used for broadcasting or closely associated purposes. We therefore need to place sensible limits on the type and quantity of information we can provide as answers to the enquiries we receive.

If we were to oblige any one of the huge number of requests similar to yours that we receive each week, it would set a precedent which we wouldn't be able to maintain. Unfortunately, for this reason, we must turn down such requests and we're regrettably unable to help you on this occasion.

Thanks again for contacting the BBC.

Finally, I have attached an invitation from the Head of BBC Audience Services, asking you to participate in our customer survey. We would welcome your views on our service.

Kind Regards
Deborah McEntee
BBC Audience Services
www.bbc.co.uk/faq

Dear Audience Member

Thank you for your recent email to the BBC. It is our aim to provide the highest standard of responses to emails we receive. To help us do this, I am writing to ask you to complete a customer satisfaction survey which is being conducted by the research company Ipsos MORI.

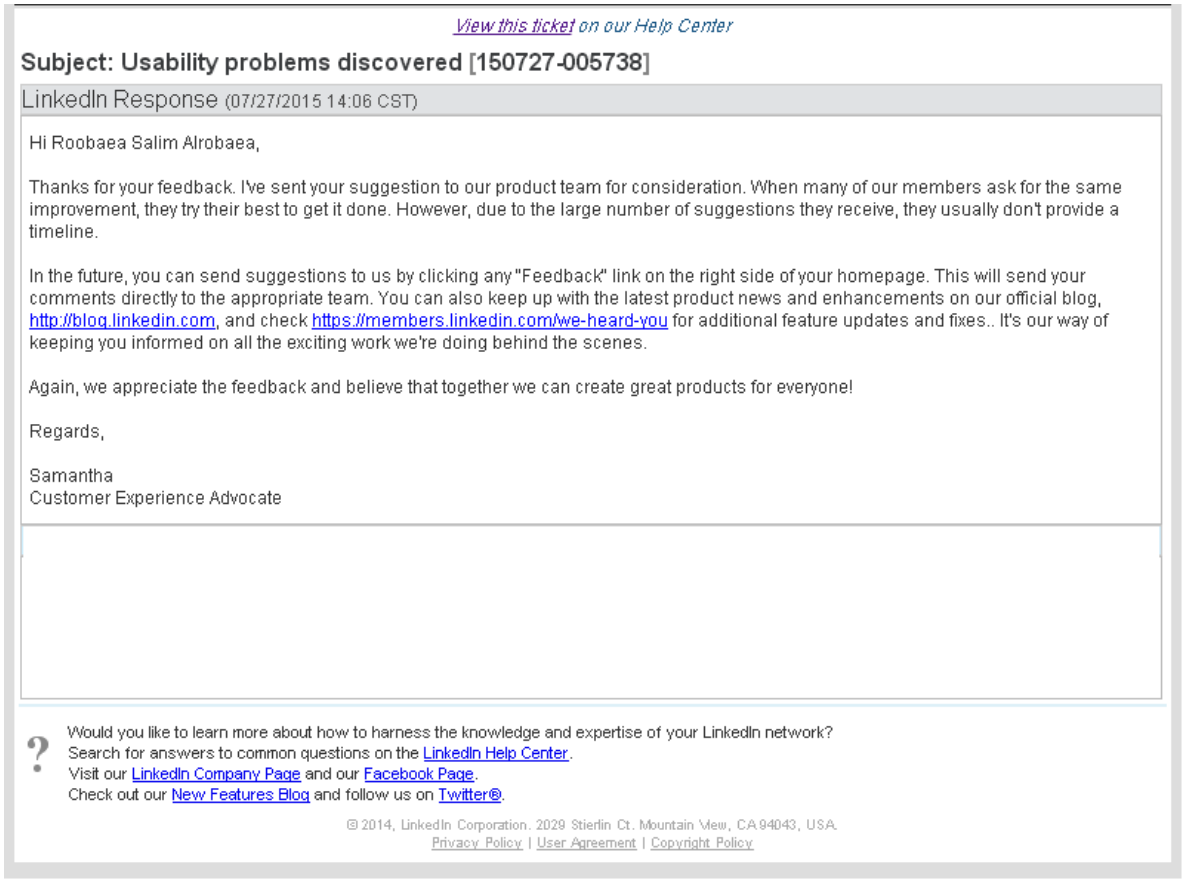
Your response will help us to judge how your most recent email was handled, so we'd be grateful if you could rate the service you received and the quality of our response. We aren't expecting you to rate the BBC, its output, or any previous contacts you may have had with us. The purpose of this survey is to help us understand the level of service we currently provide when responding to emails and to ensure we achieve the highest possible standards. We may use your individual responses to help in our staff training and performance management.

The questionnaire is very straightforward to complete and should take no longer than 5-10 minutes. You can log onto it in two ways, depending upon the email system you are using. You can either click onto the following website address, or paste it into your address bar (where you normally type in a website address). You will be guided through the questionnaire automatically once you have logged into it.

Website Address: http://survey2.infocorp.co.uk/webprod/resources/bbcinfo/start.asp?st=1&pass=CAS-1113745-6DLGHX

Thank you very much for taking the time to assist us.
Sam Smith
Head of BBC Audience Services

Appendix U2: Response from LinkedIn website on the problem report



View this ticket on our Help Center

Subject: Usability problems discovered [150727-005738]

LinkedIn Response (07/27/2015 14:06 CST)

Hi Roobaea Salim Alrobaea,

Thanks for your feedback. I've sent your suggestion to our product team for consideration. When many of our members ask for the same improvement, they try their best to get it done. However, due to the large number of suggestions they receive, they usually don't provide a timeline.

In the future, you can send suggestions to us by clicking any "Feedback" link on the right side of your homepage. This will send your comments directly to the appropriate team. You can also keep up with the latest product news and enhancements on our official blog, http://blog.linkedin.com, and check https://members.linkedin.com/we-heard-you for additional feature updates and fixes.. It's our way of keeping you informed on all the exciting work we're doing behind the scenes.

Again, we appreciate the feedback and believe that together we can create great products for everyone!

Regards,

Samantha
Customer Experience Advocate

Would you like to learn more about how to harness the knowledge and expertise of your LinkedIn network?
Search for answers to common questions on the LinkedIn Help Center.
Visit our LinkedIn Company Page and our Facebook Page.
Check out our New Features Blog and follow us on Twitter®.

Appendix W: List of publications

1. AlRoobaea, R. Al-Badi, A. Mayhew P. (2013). A Framework for Generating Domain-Specific Heuristics for Evaluating Online Educational Websites. *International Journal of Information Technology & Computer Science*, Volume 8,page 75 - 84.

2. AlRoobaea, R. Al-Badi, A. Mayhew P. (2013). A Framework for Generating Domain-Specific Heuristics for Evaluating Online Educational Websites- Further Validation. International Journal of Information Technology & Computer Science, volume 8, page 97 - 105.

3. AlRoobaea, R. S., Al-Badi, A. H., & Mayhew, P. J. (2013). A framework for generating a domain specific inspection evaluation method: A comparative study on social networking websites. In Science and Information Conference (SAI), 2013 (pp. 757-767). IEEE.

4. AlRoobaea, R., Al-Badi, A., & Mayhew, P. (2013). Generating a Domain Specific Inspection Evaluation Method through an Adaptive Framework: A Comparative Study on Educational Websites. International Journal of Human Computer Interaction (IJHCI), 4(2), 88.

5. AlRoobaea, R., Al-Badi, A. H., & Mayhew, P. J. (2013). Generating a Domain Specific Inspection Evaluation Method through an Adaptive Framework. International Journal of Advanced Computer Science and Applications, 4(6).

6. Roobaea AlRoobaea, Ali H. Al-Badi and Pam J. Mayhew (2013), " The Impact of the Combination between Task Designs and Think-Aloud Approaches on Website Evaluation" Journal of Software and Systems Development, Vol. 2013 (2013), Article ID 172572, DOI: 10.5171/2013. 172572

7. AlRoobaea, R., Al-Badi, A. H., & Mayhew, P. J. (2013). Generating an Educational Domain Checklist through an Adaptive Framework for Evaluating Educational Systems. International Journal of Advanced Computer Science and Applications, 4(8).

8. 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.

9. 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.

10. Alghamdi, A., Al-Badi, A., Al Roobaea, R., & Mayhew, P. (2013). A Comparative Study of Synchronous and Asynchronous Remote Usability Testing Methods. International Review of Basic and Applied Sciences, 1(3).

11. Roobaea AlRoobaea, Ali H. Al-Badi and Pam J. Mayhew, "Generating a Domain Specific Checklist through an Adaptive Framework for Evaluating Social Networking Websites" International Journal of Advanced Computer Science and Applications(IJACSA), Special Issue on Extended Papers from Science and Information Conference 2013, 2013.

12. Alhadreti, O., AlRoobaea, R., Wnuk, K., & Mayhew, P. J. (2014, May). The Impact of Usability of Online Library Catalogues on the User Performance. In Information Science and Applications (ICISA), 2014 International Conference on (pp. 1-4). IEEE.

13. AlRoobaea, R. S & Mayhew, P. J. (2014). The Impact of Usability on E-Marketing Strategy in International Tourism Industry. In Science and Information Conference (SAI), 2014. IEEE.

14. AlRoobaea, R. S & Mayhew, P. J. (2014). How Many Participants Are Really Enough for Usability Studies. In Science and Information Conference (SAI), 2014. IEEE.

15. Alqahtani, M. AlRoobaea, R. S & Mayhew, P. J. (2014). Building a Conceptual Framework for Mobile Transaction in Saudi Arabia a User's Perspective. In Science and Information Conference (SAI), 2014. IEEE.

16. Alqahtani, M. A., Alhadreti, O., AlRoobaea, R. S., & Mayhew, P. J. (2015). Investigation into the Impact of the Usability Factor on the Acceptance of Mobile Transactions: Empirical Study in Saudi Arabia. International Journal of Human Computer Interaction (IJHCI), 6(1), 1.

Appendix X: Confirmation of attending BCS HCI 2012 conference

This is to certify that Roobaea Salim Al-Robaea attended the BCS HCI 2012 conference in Birmingham, UK from 10th-14th September 2012

Benjamin R Cowan

Signed on behalf of BCS HCI 2012 Conference Chairs

Appendix Y: Confirmation of student volunteer for organising a conference

This is to confirm that Roobaea Salim Al-Robaea held the position of student volunteer at the BCS HCI 2012 in Birmingham, UK from 10th-14th September 2012.

The duties of student volunteers included:

- Helping delegates with any queries (directions, session times etc)
- Manning the welcome desk and completing delegate registrations
- Technical support and question and answer session assistance.

Benjamin R Cowan

Signed on behalf of BCS HCI 2012 Conference Chairs

Appendix Z: Email and questions to measure the usefulness of the adaptive framework

Dear Mr/ Ms,

This email includes two attached files. The first file is a usability problem report that was produced using three different evaluation methods on your website. The usability problem report has been divided into four parts. The first part describes the usability problem that were discovered by HE. The second part describes the overlapped problems that were discovered by the three methods. The third part describes the usability problem that were discovered by UT. The fourth part describes the usability problem that were discovered by DSI. Also, it includes the areas of the discovered problems and their severity. The second file is a set of recommendations that highlights the most common usability problems that were discovered and gives suggestion on how the usability for your website could be improved.

As this research was proposed using the adaptive framework and its generated method, which is called DSI, could you answer the following questions;

1. Do you think the problems that were discovered by DSI are useful? If not, why?
2. Based on the problems discovered by DSI, would you use the DSI method to improve your website in the future?
3. If you have another website in a different domain, would you use the adaptive framework to generate the DSI method to evaluate your website?