Investigating the usability of e-catalogues systems: modified heuristics vs. user testing

Amen Ali Alrobai
King Abdulaziz University,

Roobaea Salim AlRoobaea
Taif University,
University of East Anglia, Norwich, UK

Ali H. Al-Badi
Sultan Qaboos University, Oman
University of East Anglia, Norwich, UK

Pam J. Mayhew
University of East Anglia, Norwich, UK

ABSTRACT

The growing popularity of the World-Wide-Web has motivated many companies to utilize Internet services as means of maximizing their profit by promoting products and services. Electronic commerce, which is also known as e-commerce, can be defined as an economic environment in which business activities, such as purchasing products and advertising goods, are performed by using electronic communications. Web interface design is an important 'pull factor' in e-commerce websites. Various elements associated with this can be determinants of the success or failure of commercial websites, and electronic catalogues represent a fundamental factor in this respect. Electronic catalogues or e-catalogues are mainly used to provide users with information about products and services. Consequently, it is one of the core support information systems of e-commerce websites.

The aim of this study is to investigate the usability of e-catalogues for two e-commerce websites, which are Buy.com and Qvc.com. Moreover, to investigate the efficiency of the two usability evaluation methods which are modified heuristics evaluation and user testing in discovering usability problems of the online catalogue designs 'e-catalogues' for e-commerce websites. Also, to explore how expert evaluators' knowledge and users' experience can be exploited to discover the good and bad practices that can increase or decrease users satisfaction.

The study concluded that the usability of e-catalogues can significantly influence users' overall acceptance of shopping websites. The results suggest that attractive design, organisation, consistency and matching the real world are the most important usability guidelines in any e-catalogue design. Also, product classification has proved to be the backbone of all online catalogues. The results indicate that any e-catalogue designs that suffer from poor compliance with Nielsen's traditional usability heuristics is also more likely to fail in gaining adequate levels of user satisfaction. This proves that this set of heuristics can still be used as a powerful tool for improving interface quality in Web 2.0.

The results also suggest that, in real-life cases, relying on one UEM might provide misleading results. The heuristics evaluation method was more effective in finding a greater number of usability problems at a low cost and with fewer resources, although the usability testing method was better in finding the more serious ones.

Keywords: E-commerce, E-catalogues, Heuristic Evaluation, Usability Testing

INTRODUCTION

The growing popularity of the World-Wide-Web has motivated many companies to utilise Internet services as means of maximising their profit by promoting products and services. Electronic commerce, which is also known as e-commerce, can be defined as an economic environment in which business activities, such as purchasing products and advertising goods, are performed by using electronic communications [Qin, 2009]. E-commerce websites are complex systems that consist of "front-end" technologies, such as the E-catalogue Management System, and "back-end" subsystems, such as reliable Transaction Processing Systems (TPS) [Albers and Still, 2010]. Designing systems for unknown audiences adds further complexity to the e-commerce environment, as it involves dealing with many types of users. Therefore, properties of e-commerce websites in terms of aspects and environment have to be identified and considered in the analysis and design phase. The user's decision-making process is influenced by various factors; these are: web usability and functionality, cost, aesthetics, brand and users' reviews [Lee and Koubek, 2010]. Each one of these factors has a different weight, based on the mental model of the user, who has unique preferences. In the field of e-commerce, it is not an easy task to identify what makes a website successful, as this is influenced by the type of audience and many other attributes, such as the context, the purpose and the type of system as well as the adopted technologies [Lee and Koubek, 2010]. To sum up, online product catalogues (e-catalogues) are one of the key aspects of e-commerce websites, so they must be designed, implemented and tested carefully. However, many website owners outsource these tasks. In principle, this is due to the fact that specialists can do these better and at lower costs. However, those owners may end up with interfaces that do not satisfy users.

As this research focuses on usability issues, it is worth mentioning that many guidelines, such as in [Thatcher et al., 2006] and [Nielsen, 2000a], have been proposed to apply what is called a "Design for all" approach [Porrero, 1998], which is basically about the universal design of user interfaces from different perspectives, such as usability, which can equally influence all types of users [Thatcher et al., 2006], or accessibility, which can impact users with disabilities. Although websites have shown continuous improvement, users still experience many usability problems [Webcredible, 2010a; Webcredible, 2010b; Webcredible, 2009]. The broad aim of this paper is to investigate the influence of different e-catalogue implementations on user satisfaction. The results of this research can be used to help e-commerce website analysts and designers in applying usability guidelines, and determining which aspects should be taken into account, based on various elements such as the area of business and the audience. The aim of this paper is to identify the usability problems of e-catalogues for two e-commerce websites. Also, to investigate which usability evaluation methods (UEMs) out of 'modified heuristics' and user testing' would provide the best results in terms of detecting problems of e-catalogue. the first step is to assess which heuristics out of ten general heuristics do not work, then remove them, to develop and add new heuristics that cover areas not covered by general heuristics, to create focussed modified heuristics.

This paper is organized in the following way. Section 2 starts with a Literature Review to this study that includes a definition of e-catalogues, usability, usability problems, a severity rating, number of evaluators and users, heuristics process and usability testing. Section 3 discusses the methodology that was applied in the current study. Section 4 discusses the actual experiments. Section 5 provides an analysis and discussion of the results. Section 6 presents the conclusion.

Literature Review

One of the core support information systems of e-commerce websites is the online products catalogue. It is mainly used to provide users with information about products and services. The benefits of this tool are many: lowering advertisement and distribution costs, adding more flexibility to browsing, updating information, adapting information based on users' preferences, and extending searches to other catalogues in what is called "joint catalogues". These are just a few examples of what e-catalogues can add to e-commerce websites [Suh, 2005]. Much research has been carried out to analyse different models of e-catalogues and build better ones. In general, e-catalogues, in terms of functionality, consist of the functions illustrated in Figure 1. Some functions vary based on the nature of the market. Negotiation, which is associated with communication, is a good case in point in that while in retail markets, prices and other features are fixed, in other markets such as stocks and auctions, it is a fundamental aspect [Yen and Kong, 2002]. Maes et al. (1999) concluded that there are similar buying process stages in all the theories and models that they analysed. These stages are: Need identification, Product brokering, Merchant brokering, Negotiation, Payment and delivery, and finally Service and evaluation. It is true that the major challenge in the website environment is how to keep users attracted for a sufficiently long period of time. From the users' perspective, while many studies have confirmed that content is the most important element and that users are "goal-driven"("meaning that they focus on only one thing in mind") [Nielsen, 2000b], navigation-associated problems are considered the second main reason for not continuing to shop on a website [Manning et al., 1998].

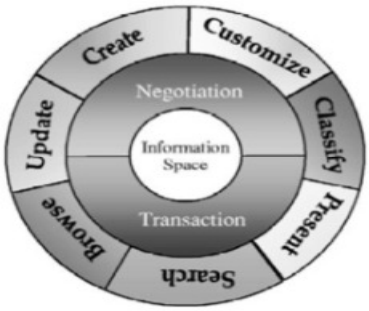

Figure 1: Functional definition of electronic catalogue [Yen and Kong, 2002]

In software development lifecycles, the concept of usability as an engineering activity has become a fundamental element. There are many definitions of usability. [ISO, 1998] has defined usability as, "the extent to which a product can be used by specific users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use". From a practical perspective, it can be defined as a set of attributes and practices that should be considered during the analysis and design phases, aimed to minimize users' frustration. From a user perspective, it is the experience that he or she gets when performing a task "without hindrance, hesitation or question" [Rubin and Chisnell, 2008]. Also, it is a formal method for measuring the quality of any interface design, as perceived by users. The reviewed literature shows that usability is not a single 'one-dimensional' property of a user interface. There are many usability attributes (usability measures) that should be taken into account and measured during experimental sessions. They are the tools that can be used to determine whether or not an interface design is likely to provide users with a satisfactory experience. In usability testing/inspection studies, the aspects of the system that are to be measured (and how they are to be measured) should be clearly specified. Usability measures

can provide both qualitative and quantitative results. Nielsen (2001c) argued that while qualitative studies are more credible, measuring usability through collecting metrics (statistics) is still worth doing. He pointed out that this is because quantitative studies enable researchers to focus on specific aspects rather than dealing with the whole system [Nielsen, 2004]. Shackel, and Richardson (1991) proposed a four-dimensional approach to the definition of usability, in which effectiveness, learnability, flexibility and attitude are the attributes that influence the acceptance of a product. Nielsen (1994) introduced some different attributes, including learnability, efficiency, memorability, and error handling. Learnability can be defined as, "a measure of the degree to which a user interface can be learned quickly and effectively" [Usabilityfirst, 2011b]. Efficiency is the speed with which a task is accomplished accurately [Nielsen, 1994], and can be assessed by measuring the time spent on tasks. Effectiveness is the degree to which an interface helps users to achieve tasks as they were intended [Rubin and Chisnell, 2008]. It can be measured by calculating success rate or number of errors. Satisfaction is the degree to which users' expectations and system performance are matched [Nielsen, 1994]. It can be measured by registering users' subjective responses to a set of questions or statements (e.g. interviews or questionnaires), or by observing users while they interact with the system.

Usability evaluation methods (UEMs) are a set of techniques that are used to measure usability attributes. They can be divided into three categories: inspection, testing and inquiry. Heuristic evaluation is one category of inspection methods. It was developed by Nielsen and Molich (1990a), guided by a set of general usability principles, "heuristics". It can be defined as a process that requires a specific number of experts to use the heuristics in order to find usability problems in a broad range of interface designs in a short time and with ease [Nielsen and and Molich, 1990a]. Magoulas et al. stated that "heuristic evaluation is a widely accepted method for diagnosing potential usability problems and is popular in both academia and industry" [Magoulas et al., 1990]. Also, it can be used early in the development process, and can be used throughout the development process. However, it is a subjective assessment that depends on the evaluator's experience and it produces a large number of false positives which are not usability problems. There are other disadvantages such as Non-involvement of real users and the lack of methodology of identifying if the whole system is evaluated [Nielsen and Molich, 1990b; Holzinger, 2005; Nielsen and Loranger, 2006]

There is no specific procedure for performing heuristic evaluation. Nielsen (1994) provided a model procedure for heuristic evaluation as shown in Figure 2. The pre-evaluation coordination session (a.k.a training session) is very important. Before the expert evaluators evaluate the targeted website, they should take about 10 minutes browsing the site to familiarize themselves with it. Also, they should take note of the actual time taken for familiarisation. If the domain is not familiar to the evaluators, the training session provides a good opportunity to present the domain. It is recommended that in the training session (familiarisation), the evaluators evaluate two websites using the heuristics in order to make sure that all the principles are appropriate for this kind of website. In the actual evaluation, each evaluator is expected to take around one hour to list all usability problems. However, the actual time taken for evaluation should always be noted. The debriefing session would be conducted primarily in a brainstorming mode and would focus on discussions of possible redesigns to address the major usability problems and general problematic aspects of the design. A debriefing is also a good opportunity for discussing the positive aspects of the design, since heuristic evaluation does not otherwise address this important issue. After that, 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 individually estimate the severity of each problem [Chattratichart and Lindgaard, 2008].

T
R

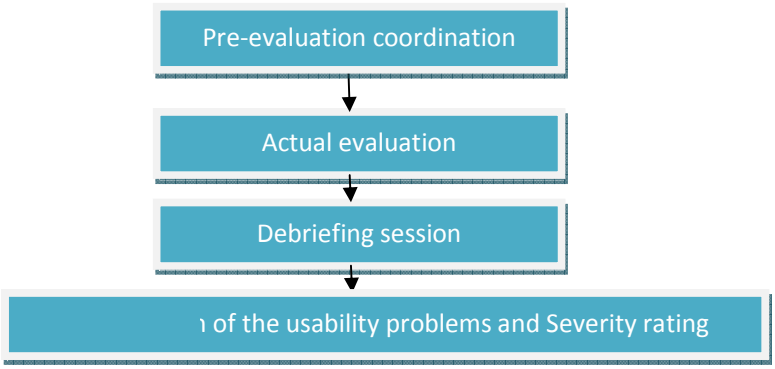

Figure 2: Heuristics Evaluation Process [Nielsen, 1994]

In addition, computer technologies are becoming ever more integrated into everyday life, and new types of human computer interaction are emerging. Consequently, the general ten heuristics are not readily applicable to many new domains with different goals and usability issues. Also, some studies such as [Thompson and Kemp, 2009] suggest that in social networking websites, which classify as Web 2.0 sites, have poor compliance to Nielsen's heuristics. Therefore, e-catalogues in commercial and e-commerce websites that are also classified as Web 2.0 sites might have poor compliance as well. For this reason, Ling and Salvendy (2005) outlined two ways to develop a new set of heuristics which are; 1) development: this step means that the researchers should assess the general ten heuristics to identify those heuristics that do not work, and remove them. Then, develop new heuristics to cover areas not covered by Nielsen's heuristics. Finally, these new heuristics are added to the ones remaining from Nielsen's set. 2) Validation: this step means that the researchers should compare the newly developed heuristics with Nielsen's original set by empirical processes. Then, one must investigate the results to determine which is better.

Usability testing is another method of interface evaluation. It is unlike inspection methods in terms of who conducts the evaluation and how it should be conducted. This method involves having end-user representatives who are observed whilst performing a set of carefully designed tasks that cover the main aspects of an interface design. There are many tools can be used to conduct usability testing. Thinking Aloud Protocol is one tool that involves a specific number of users who interact with the system individually, based on pre-defined tasks. Encouraging the participants to provide verbal descriptions of what they are intending to do and what is happening on the screen is the main aspect of this method [Rubin, and Chisnell, 2008]. It is believed to be one of the best methods for collecting qualitative data, especially when incorporating some usability inquiry methods such as interviews and questionnaires [Nielsen, 1994]. It has been argued that Thinking Aloud Protocol should be avoided in certain circumstances. [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), Thinking Aloud Protocol should be avoided as it might negatively impact on the performance of the users. On the other hand, Tullis, and Albert (2008) questioned 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. The other tool complementary to the first is Tasks Design. The tasks designed for usability testing should be focused on the main functions of the system. The tasks cover the following aspects: 1) Product page; 2) Category page; 3) Display of records; 4) Searching features; 5) Interactivity and participation features; 6) Sorting and refining features. Dumas and Redish (1999) suggested that the tasks could be selected from four different perspectives. These are: 1) Tasks that are expected to

detect usability problems.2) Tasks that are based on the developer's experience.3) Tasks that are designed for specific criteria.4) Tasks that are normally performed on the system. They also recommended that the tasks should be short and clear, in the users' language, and based on the system's goals. In terms of task design, Alshamari (2010) explains that there are three types of task: structured, uncertain and problem-solving. Each type has its own influence on user performance. The same author has found that the problem-solving type is the best in terms of revealing usability problems. In his experiments, however, he found that using both problem-solving and structured tasks help in detecting around 82% of the problems. Of course, this percentage might change slightly according to the system being tested and the testing conditions. In addition, the key point in participant selection is that they should match the real audience of the selected websites, or at least be as close as possible [Rubin, and Chisnell, 2008].

The result of applying heuristics and usability testing are a list of usability problems. These problems are classified into different groups in which a numeric scale is used to identify the severity of each problem. Firstly, this issue is not a usability problem at all. Secondly, this is a cosmetic problem that does not need to be fixed unless extra time is available on the project. Next, this issue is a minor usability problem; fixing this should be given low priority. Then, this is a major usability problem; it is important to fix this, so it should be given high priority. Finally, this issue is a usability catastrophe; it is imperative to fix this before the product can be released [Nielsen, 2005]. Tana et al. (2009) summarized that "both user testing and heuristic evaluation methods provide valuable insight into usability problems all stages of development. User testing relies mainly on the experience and comments of the users and is usually conducted in a scenario-based environment". As a result, user testing would usually evaluate according to what already exists, rather than to what is possible. On the other hand, heuristic analysis relies mainly on the expertise and knowledge of human factors engineers that would evaluate the web site based on a set of heuristics. Both of these methods have their individual strengths and weaknesses, and neither one guarantees an optimal result. Jeffries et al. (1991) reported that heuristic analysis discovered approximately three times more problems than user testing, however, more severe problems were discovered through user testing, as compared to heuristic analysis [Nielsen, 2005].

Therefore, e-catalogues usability is the main area that will be investigated throughout this study. The aim of this study is to investigate the usability of e-catalogues for two e-commerce websites by using two usability evaluation methods which are modified heuristics evaluation and user testing in discovering the usability problems. Also, to explore how expert evaluators' knowledge and users' experience can be exploited to discover the good and bad practices that can increase or decrease users satisfaction.

RESEARCH METHODOLOGY

Prior to discussing the methodologies, the kinds of data (quantitative and qualitative) that need to be collected in order to ensure an appropriate selection of methods will be described. In terms of quantitative data, this study will evaluate usability by measuring three quality attributes: effectiveness, efficiency and satisfaction. These attributes will help to assess the degree to which the selected e-catalogues are easy to use. In order to measure these attributes, there must be a set of metrics, and these are: success rate, error rate and time spent on tasks. On the other hand, users' thoughts and behaviour will be the basis of the qualitative data that will be collected, in order to identify usability problems as well as any bad practices that are likely to exacerbate user frustration.

This study is based on the experimental methods (see Figure 3) that is used together with usability inspection (modified set of heuristics) and usability testing with Think-Aloud protocol and careful task designs.

Modified heuristics evaluation will be used in order to measure the usability of the selected websites in terms of effectiveness [Nielsen, 2001c]. This will be achieved by engaging expert evaluators who will be using a set of modified heuristics to review the interface designs.

Usability testing, on the other hand, will be employed to study the users' behaviour while performing their tasks on the selected websites, and to evaluate the e-catalogues in terms of effectiveness, efficiency and satisfaction [Sauro and Kindlund, 2005]. After conducting both experiments, and in the stages of data gathering and analysis, all the detected problems will be rated based on the severity rating scale.

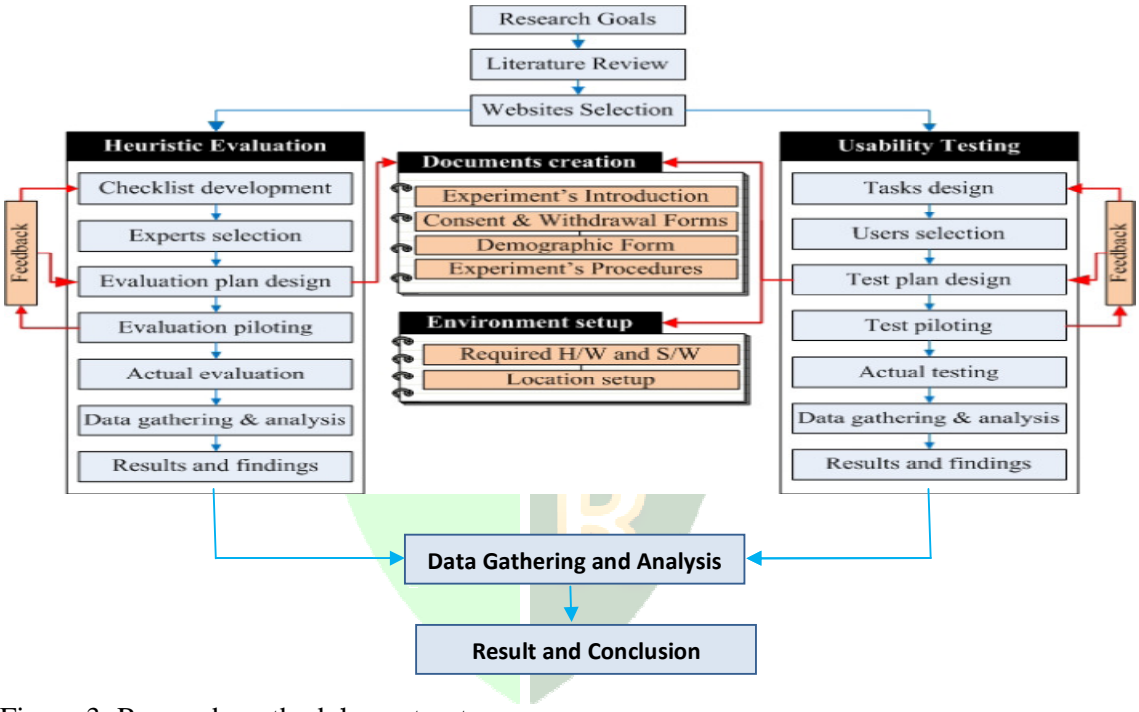

Figure 3: Research methodology structure.

CONDUCTING THE EXPERIMENTS

Before conducting the actual experiment there were a set of procedures that the researchers followed, which are:

Selection of the targeted websites

The researchers want to ensure that the selected websites will support the research goals and objectives. Therefore, the selection process was criteria-based; five aspects have been determined and verified for each website, and these are: 1) Good interface design.2) Rich functionality. 3) Good representatives for Web 2.0 sites. 4) Not familiar to the users. 5) No change will occur before and during the actual evaluation. The researchers also decided to consider two extra aspects. One is that the websites should be complex; a complex website in this context means a website that employs a great many features and technologies. The second aspect is that the researchers will try to find websites that have copied other popular

ones, such as Amazon and e-Bay. This might help to determine why these copied websites have failed to gain their anticipated success, though of course various factors would play a role in this [Zhou et al.,2007]. However, in this study the researchers will examine this in respect of online catalogues. The selected websites are: Buy.com and Qvc.com. Both of these have all the aspects mentioned above, with a couple of exceptions. With regard to the implementation of any changes in the interface design or website functionality during the period of website testing, the researchers did not receive any response from either of the websites' administrators. The other exception is associated with Qvc.com in that the researchers have been unable to confirm whether or not the website is a copy of another popular one. However, as long as most aspects are present in both websites, and as long as the testing sessions are conducted in the shortest possible time (maximum one week), their selection should not pose any problems in this study.

Severity rating

The aim of the severity rating is to help the expert evaluators and observer to rank the usability problems. Consequently, this research has used a rating scale (from 0 to 4) as recommended by [Nielsen, 1994] in literature review; 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.

Heuristic Evaluation

This experiment can be divided into three phases, each of which has a set of processes and procedures. However, they are generally aimed at ensuring the best possible preparation for the actual testing. These phases include the modified heuristics and check-list, which will be rated based on a rating scale. They also include the selection of evaluators, the creation of inspection instructions and procedures, and finally piloting the test in order to make final improvements before starting the actual evaluation.

Modified Guidelines for Heuristics Evaluation

This study will consider the most commonly used general heuristics (Nielsen's heuristics) as table 1 shows. While it has been argued that these heuristics are general, the researchers assess those heuristics that do not work and remove them by scanning the literature review and two independent expert evaluators. Then, the new heuristics were added to cover areas not covered by Nielsen's heuristics to the ones remaining from Nielsen's set as outlined by [Ling and Salvendy, 2005]. Consequently, guidelines number 5 'error prevention' and 9 'Helps users recognise, diagnose, and recover from errors' have been combined in one rule called: Error prevention and correction [Alshamari, 2010]. In fact, in the test piloting, having them separated seemed to create some confusion for the evaluators. Additionally, 'participation', as a main aspect of Web 2.0 sites, is the new heuristic that has been added to the general ones. This is in order to check if the system provides a suitably good interactive environment for users, one in which they can exchange information and share their experiences (e.g. the website that allow users give rating and reviewing). These guidelines have been broken down into more meaningful elements; this should greatly facilitate the inspection process. The evaluators, in this regard, have been advised to use these elements as good examples of the issues in the main set of guidelines. The researchers believe that this

might inspire the evaluators to think more deeply and to suggest other 'examples', which in turn might help in spotting other problems as table 1 shows.

Modified Heuristics	**Nielsen's Heuristics**	**Comparison**
Visibility of system status	Visibility of system status.	Same
Match between system and the real world	Match between system and the real world.	Same
User control and freedom	User control and freedom.	Same
Consistency (within a site) and Standards (between sites)	Consistency and standards.	Same
Recognition rather than recall	Recognition rather than recall.	Same
Flexibility and efficiency of use	Flexibility and efficiency of use.	Same
Aesthetic and minimalist design	Aesthetic and minimalist design.	Same
Error prevention and correction	Help users recognize, diagnose and recover from errors.	Modify
	Error prevention.	
Help and documentation	Help and documentation.	Same
Participation		New

Table 1: Modified heuristics comparing with Nielsen's heuristics

Selection of Evaluators

In this research, three expert evaluators were involved to conduct the heuristics evaluation experiment as it suggested by [Nielsen, 1990]. All the evaluators confirmed their participation and all the experiment materials were provided to them accordingly. Table 2 outlines the demographics of the selected evaluators.

Evaluator No.	Gender	Level of English	Experience in usability	Prior use of e-catalogues	Profession	Education level
1	Male	native	More than 5 websites	Yes	Associate Tutor	PhD
2	Male	Good	1-2 websites	Yes	IT Specialist	MSc
3	Male	Good	1-2 websites	Yes	IT Specialist	MSc

Table 2: Demographic information of the expert evaluators

Inspection preparation and procedures

Nielsen (2005d) recommended that the evaluation sessions should be run separately in order to ensure impartial results; this is because the perception of evaluators towards problems is not equal. Another rule is that the evaluators should not be helped or guided on how to use the system that is to be tested. All the evaluators were provided with an introduction sheet in which the goals and objectives of the evaluation, the participants' roles, the data collection and storing methods were all explained in detail. This sheet was submitted to them with consent and withdrawal forms along with the evaluation procedures that they should follow. The following are the sequential steps of the heuristic evaluation; 1) The evaluators were given a quick introduction to the test, and what he/she is expected to do is

explained; 2)The guidelines checklist, instructions and procedures of the test were provided along with user accounts' information (i.e. username and password) that created for testing purpose; 3)The evaluators were asked to spend five minutes familiarizing themselves with the websites; 4)The evaluators review both websites consecutively and rate all the problems they find.

Piloting the Experiment

The key point in piloting the heuristics evaluation materials is to ensure that the guidelines checklist is sufficiently clear and that it is applicable to the selected websites. This is because most of the checklist elements which are derived from the heuristics of [Nielsen and Molich, 1990b] are, in fact, general ones, albeit well-established ones and commonly used in the literature. The researchers selected an independent evaluator, who then performed a full evaluation in which all the test steps and procedures were carried out. The data of this test are not included in the analysis and results, as the pilot test involved a great many interruptions for dissection and explanation.

Usability Testing

The usability testing is the main method in the second experiment. While it was aimed also at finding problems associated with the design and functionality of the two interfaces. it was used to measure user' performance, satisfaction and willingness to use the selected e-catalogues. This method involves employing representative users to perform some carefully designed tasks. This will help in detecting those problems that could frustrate real users.

Participants recruitment

As shopping websites target a wide range of user types, it was not difficult to find participants who match the real audience. Dumas and Redish stated that 6 to 12 participants are typical numbers of user testing [Dumas and Redish, 1999]. Then researchers decided to employ up to ten users. All the subjects were provided with the test introduction sheet, consent and withdrawal forms, and also the instructions and procedures that they should follow. After obtaining the subjects' consent to participate, a schedule containing the location and timing of the test session was created, and the subjects were notified accordingly

Task selection and design

The tasks were designed based on the main functions that users would normally perform on both websites. Due to the nature of the aspects that were to be examined, there was a mixture of structured and uncertain tasks; problem solving tasks were ignored because the aspects to be investigated are mainly about using the catalogues to find products and related information. In the pilot study (Section 4.3.3), there is more explanation on why this type of tasks was ignored. There are six tasks in total. Both websites are to be tested with the same set of tasks as Figure 4 shows.

Task 1	**Goal: Getting an idea about the website the Thinking Aloud protocol.**
Steps	▪ Go the website www.buy.com ▪ Spend maximum 5 minutes to browse the website and practice the Thinking Aloud Protocol.
Task 2	**Goal: To what extent browsing the catalogue is easy without using the search engine?**
Steps	▪ In the same website, go the home page. ▪ Try to find **a printer** that has got these information: – ***Brand Name:*** HP – ***Model:*** Officejet 6500A ▪ Add the printer to your shopping cart.
Task 3	**Goal: To what extent finding required information is easy?**
Steps	▪ In the same website, go to the home page. ▪ Find any camera you like. ▪ If it has 3 years warranty and available in the stock, add it to your shopping cart. ▪ Find a memory card that can be used with this camera. ▪ Add the memory card to your shopping cart.
Task 4	**Goal: Searching and sorting features.**
Steps	▪ In the same website, go to the home page. ▪ Use the search engine to find the camera that has got these information: (Note: when you get the results, first, **sort them by the lowest price** then look for the camera) – Brand Name: Samsung. – Model: ST700. ▪ Add the camera to your shopping cart.
Task 5	**Goal: Products' options, images and the shopping cart.**
Steps	▪ In the same website, go to the home page. ▪ Find any shoes you like. ▪ Select your appropriate color and size. ▪ Check the main picture of the shoes and try to maximize it then minimize it. ▪ Check also the other pictures (if any). ▪ Add it to your shopping cart ▪ Go to your shopping cart and increase the quantity of this product.
Task 6	**Goal: Interaction and participation.**
Steps	▪ In the same website, go to the home page. ▪ Find any product that you like. ▪ If it has got good rating, add it to your shopping cart. ▪ Write your own review about this product (short sentence).

Figure 4: Usability testing tasks

Piloting the experiment

The researchers selected independent users to perform all the test steps and procedures. The data derived from this test are not included in the analysis and results, as it involved many interruptions for dissection and explanation. The researchers benefited from this test by gaining an idea of the time required to perform the tasks. Also, it was a good resource for further ideas for improving the tasks and the questionnaires. When the independent user was given a problem-solving question on how to find a specific product, the researchers noticed the following: the user, based on his perception, developed his own approach to accomplish the task. Therefore, he did not try to use the various tools on the website, and so the researchers were not able to collect user feedback on these tools. Therefore, the researchers decided to employ structured and uncertain tasks only. Also, the researchers and independent users decided Buy.com to the first website tested in all the testing sessions, followed by Qvc.com. However, after having conducted the pilot session (i.e. first independent user), it was noticeable that the users had become frustrated by using Buy.com. Therefore, the researchers decided to apply a slight change to the sequence of all testing sessions in that he decided to split users into two equal groups. The first group would then start with Buy.com, followed by Qvc.com, and contra-wise for the other group. This was done because the users may have tendency to heavily criticize the first website to prove to the researchers that they understand and are practicing the Thinking Aloud Protocol very well.

Consequently, there may be a slight tendency to criticize the second website less severely. While this could possibly influence the satisfaction questionnaires, the swapping of the websites among the two groups will help the researchers to detect such behaviour; if this were the case, the satisfaction questionnaire's results would be given a very low priority in this study. On the other hand, users' performance (Section 5.2.3) will not be affected, as it does not consider participant feedback

Testing Environment

Arranging an appropriate location where the test sessions can be conducted successfully is an essential part of this experiment. Therefore, the researchers ensured that the selected locations have the following features: 1) Easy to access for participants; 2) Controlled location where no interruptions can occur; 3) Quiet area; 4) Reliable Internet connection. The locations that were selected and that matched the above conditions are as follows: 1) Pre-booked rooms in the main library of the University of East Anglia (morning sessions); 2) The MSc Lab in the School of Computing (night sessions). Figure 5 illustrates the testing room setup. The researchers (observer) sat a couple of feet away from the user in order to observe the testing and also not to stress him/her or distract his/her attention.

Figure 5: Test room setup.

Online satisfaction questionnaire

This questionnaire was implemented through the 'Morae software' to be provided immediately after the test in order to collect user feedback and comments pertaining to the selected website, based on performed tasks.

Data collection

The users' performance was measured by collecting four metrics: "success rate (whether users can perform the task at all), the time a task requires, error rate and user subjective satisfaction" [Nielsen, 2001]. Success rate and error rate are used to measure effectiveness, while time is used to measure efficiency, and satisfaction is measure by the satisfaction questionnaire [Sauro and Kindlund, 2005]. Also, video recordings were reviewed carefully to observe the user's behaviour in order to assess the impact and persistence of each error.

Testing procedures

The second experiment (User Testing) was conducted by giving a quick introduction about the researchers and the purpose of the study for each user. The next step was for explaining the environment and equipment, followed by a quick demonstration on how to 'think aloud' while performing the given tasks. All the above steps took approximately ten

minutes for each test session. The actual test started from this point i.e. when the user was given the task scenarios sheet and asked to read and then perform one task at a time. In fact, the first task was designed merely to familiarize the users with the test environment, equipment, selected websites and how they could naturally verbalize their actions.

ANALYSIS AND DISCUSSIONS

In this section, the data collected from both experiments (HE+UT) that was performed on both websites will be analyzed separately. A comparison will be made between the websites in each experiment separately in order to investigate whether the websites have achieved similar results. For this, the knowledge of the expert evaluators' will be utilized to examine if the heuristics evaluation on its own is sufficient for judging how real users might be affected by usability problems. In other words, the validity of the results in the first experiment will be verified by the second experiment. Also, the users' performance in the usability testing will be examined. Finally, the performance of both usability evaluation methods (UEMs) will be assessed and compared with other studies.

The Analysis of the heuristics evaluation

This section discusses and analyses the number and types of detected usability problems. Moreover, graphs will be introduced for further clarification.

The number of usability problems discovered

Each evaluator reviewed both websites and rated the discovered problems based on the severity rating scale. All these problems in the individual reports were consolidated into one list. This list consists of unique problems only accompanied by the score given by each evaluator. Then the average rate was calculated. If one of the evaluators gave a particular problem severity rating of zero (i.e. it is not a usability problem), he would not be considered in the calculation. Thus, this evaluator does not have an effect on the average rating of that problem. Figure 6 shows the total usability problems found by the heuristic evaluation. The usability problems detected in Buy.com were 18.46% higher than in Qvc.com (59% vs. 41%). The figures in the chart on their own cannot explain the influence of these problems, unless they are classified based on their error type (Section 5.1.2). This might reveal that the figure below is mostly influenced merely by the least important types of problems (i.e. cosmetic and minor).

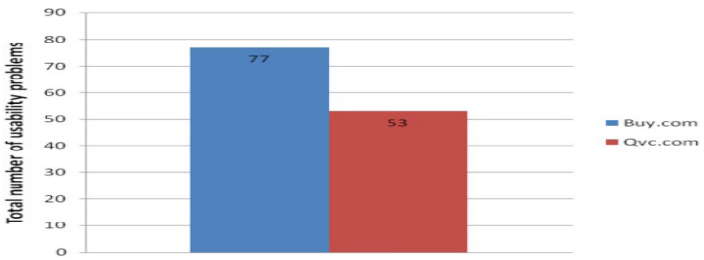

Figure 6: Total usability problems found by the heuristic evaluation.

The types of usability problems discovered

Figure 7 illustrates the number of problems classified by type. It can be seen that Buy.com achieved worse results in terms of cosmetic and minor problems only. The number

of minor problems in Buy.com was almost double the number in Qvc.com. However, and more importantly, Buy.com was less affected by major and catastrophic problems.

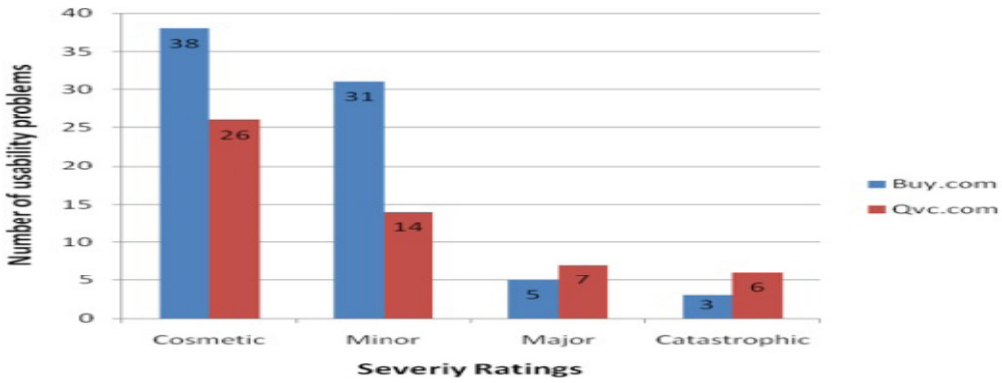

Figure 8: Problems distribution based on severity ratings - Heuristic Evolution.

General observations

Figure 8 shows that Buy.com has a complex and often redundant organization scheme; also there is more than one type of product categorizations. Consequently, in some scenarios users might find it difficult to anticipate where a product will be found. This reflects poor compliance to the heuristic that emphasizes consistency and reduces redundancy. In Figure 8, there are two subcategories, called 'Clothing shoes' and 'Sports Bags'; the 'shoes' section exists in each one of them.

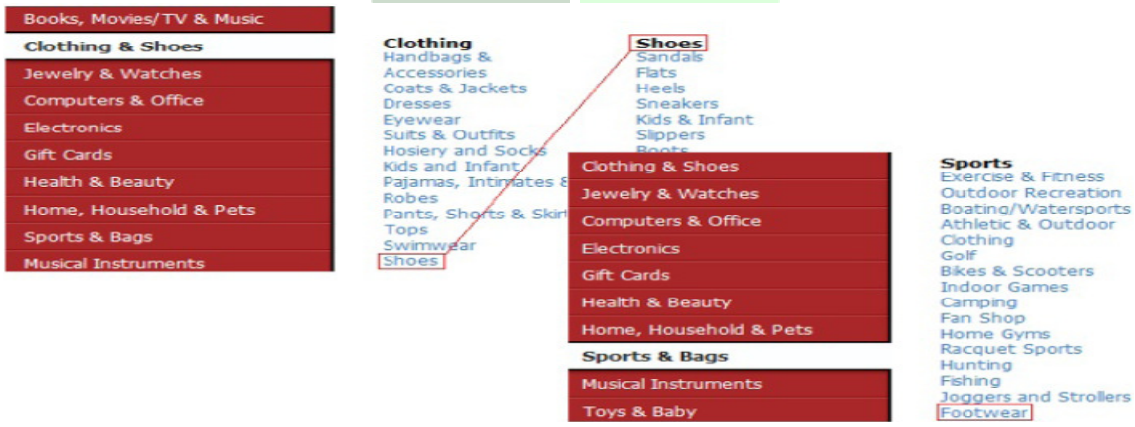

Figure 8: Complex and redundant organization scheme - Buy.com.

In Buy.com, when users login to their accounts, which means that their profiles and personal information are enabled for access, they are likely to click on links pointing to external websites. This reflects poor compliance to the heuristic that recommends good visibility of the website status. Also, this could pose the risk of identity theft, as the user leaves the main website without signing out. Figure 9 shows how external links, under the title 'Sponsored Links', are mixed with the website's products.

Figure 9: Poor visibility of system status - Buy.com**.**

Some evaluators reported that the 'Add to Cart' buttons in the Qvc.com catalogue are confusing, as users might be confused about whether the button belongs to the product above or below. This actually has been verified in the usability testing, and it was confirmed that this is a real usability problem, associated with the heuristic 'recognition rather than recall' see figure 10.

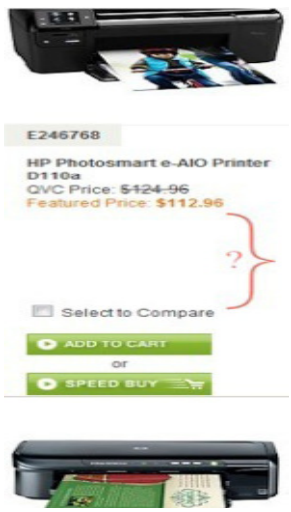

Figure 10: Recognition rather than recall - Qvc.com.

Table 3 outlines the important usability problems discovered by each method in Buy.com.

Problems discovered by HE	Problems discovered by UT
1. The website doesn't provide feedback for every action. 2. No feedback on users' location (e.g. Breadcrumb) 3. No identified link to navigate back to the product's parent category. 4. The search engine doesn't always provide accurate results. 5. The formatting standards aren't consistent in all pages. 6. The website doesn't prevent users from making errors whenever possible. 7. Prompts and messages aren't placed where the eye likely to be looking on the screen. 8. Information isn't grouped into logical zones. 9. Headings aren't used to distinguish between different zones. 10. Users are not enabled to set their own default choices/interests. 11. The website isn't aesthetically pleasing. 12. Too much variety of colours, font sizes and formats. 13. The location of shopping cart is confusing and users' often can't find it easily. 14. On-line instructions aren't visually distinct. 15. No consideration for sequence of user actions. 16. Help information isn't descriptive (what is this thing for?) 17. Help information isn't Interpretive (why did that happen?) 18. There isn't context-sensitive help. 19. Users can't resume their task where they left off after accessing the help. 20. Filtration features are not accurate. 21. Products classification is not accurate. 22. Attention-grabbing techniques/strategies aren't used with care. 23. Error handling is confusing. 24. Links are not always underlined. 25. User can't ask questions about specific products. 26. Using colloquial slang language. 27. Related accessories are not provided. 28. The design is cluttered in places.	1. No link for comparison products 2. There is no link to the "Home" page 3. There is no distinguishing between home page and other pages. 4. Products classification is confusing (e.g. "Printers" section is not in the "Electronics" page. 5. Too much sections and subsection in the main dropdown menu "All Products". 6. Poor consistency in the interface design of the products' pages. (e.g. TV's page completely differ from Printer's page). 7. Overload of advertising and related links. 8. Filtration features are not for all products types. 9. The use of Pop-up window is confusing. 10. Relevant information is not grouped in one distinct area. 11. No use of titles to distinguish between different zones. 12. So difficult to understand the structure of the website. 13. "Product Description" page is cluttered and irrelevant/redundant information is included. 14. Unclear terminologies (e.g. using "Essentials" instead of "Accessories"). 15. Poor prioritising of users' tasks. 16. The search engine sometimes provides irrelevant results. 17. Too many filtration options in some cases. 18. It is not clear how to minimized/maximized products' pictures. 19. In a particular zone, irrelevant function/information is provided (e.g. product picture zone). 20. Locating the drop-down menu called "all products" and fly-over menu called "all categories" in the same row (level) is confusing as each one has different job. 21. It is not clear that printers' manufacturer brand logos are clickable and used to browse by brand. 22. "Sort by" features don't have a title and some users couldn't notice it. 23. Error messages area in the "shopping cart" page is in unexpected location.

	24. Too much scrolling. 25. No link to jump back to the top of the page after long scrolling. 26. Users spends fairly long time to understand how to use filtration features because no consistency in the graphical design and functionality. 27. The location of shopping cart is confusing and users' often can't find it easily. 28. The price of some items only appears when a user checks out. 29. No Spell out for abbreviations (e.g. QTY) 30. In the home page there are two areas for "all products" each one has slightly different organization. 31. Ads of outside companies are not on the periphery of the page. 32. Some items are without "Add to basket" button and no justification is provided. 33. External link is mixed with the websites links. 34. The use of "Important message!" instead of "Error" makes users unable to recognize it. 35. Filters in some cases provide irrelevant results 36. "Shop by brand" is a filter exists twice in many pages and each one gives different results.

Table 3: Comparison between some problems discovered by HE and UT

The Analysis of usability testing

This section discusses and analyses the number and types of detected usability problems. Then the three quality attribute (effectiveness, efficiency and user satisfaction) are measured and analysed. Graphs will also be introduced for further clarification.

The number of usability problem discovered

After collecting all the errors encountered during the usability testing sessions, the researchers listed all of them in two separate tables; one for each website. Then, all the redundant problems were removed. Actually, the redundancy helped in assessing the frequency of each problem. Subsequently, all the video recordings were reviewed carefully in order to assess the impact and persistence of each error. Analyzing these three factors was vitally important for assigning impartial severity ratings. Figure 11 shows that the number of usability problems encountered in Buy.com was 48% higher than in Qvc.com (74% vs. 26%). Again, this has to be further checked by investigating the types of these problems, and by identifying the aspects of the catalogue in which these problems exist (Section 5.2.2). For example, problems that might affect the process of finding products (i.e. search engine

features and filtration tools) are likely to impact the users' experience more than problems associated with interactivity and participation features (i.e. writing reviews and rating items).

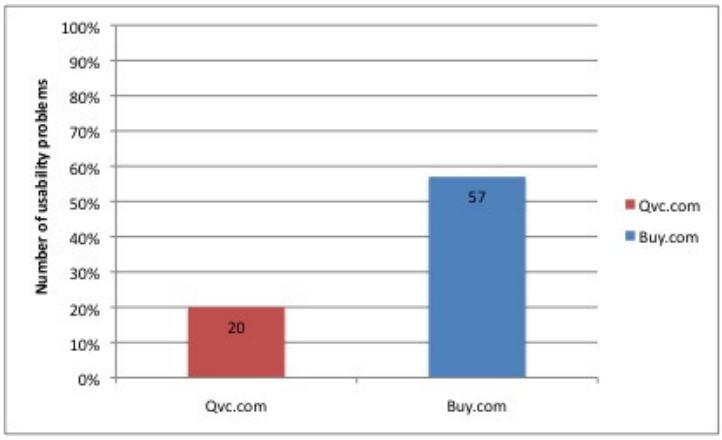

Figure 11: Total usability problems found by usability testing.

The types of usability problem discovered

Figure 12 illustrates the number of problems classified by the type. It can be seen that Buy.com achieved worse results for all types of problems.

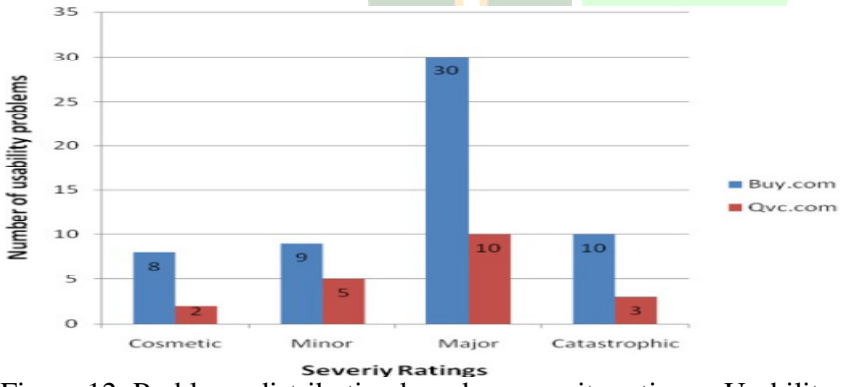

Figure 12: Problems distribution based on severity ratings - Usability Testing.

Referring again to Figure 12, it can be seen that the users encountered 30 major usability problems in Buy.com, while in Qvc.com there were only 10 of them. Major problems that are the problems that could make users stumble or cause difficulties in using the e-catalogue. It should be pointed out that the 10major problems and 3 catastrophic ones in Qvc.com could also cause deterioration in user acceptance toward the website, especially as these figures are associated with just one part of the website, which is the e-catalogue. Therefore, the next section will evaluate users' performance to assess the real influence of these numbers.

Users' performance measurement

This section assesses the usability of the both websites by analyzing users 'performance. The three quality attributes that will be used to achieve this goal are: effectiveness, efficiency and satisfaction. Effectiveness will be assessed by measuring

success rate and number of errors. On the other hand, efficiency will be assessed by measuring the time spent on tasks and finally satisfactory questionnaires and observation are the methods to measure users' satisfaction.

Task Completion Rate

Users were given, in total, six distinct and criteria-based tasks. The first one was just to familiarize the users with the websites, testing equipment and materials. Therefore, its results were not included in the analysis. Consequently, each participant performed 10 tasks in total. In other words, there were 45 tasks for each website. Figures 13 and 14 compare the percentage of users who completed tasks successfully, partly or failed to complete in both websites.

It is noticeable that only 33.33% of the users were able to perform task 4 successfully on Qvc.com. This reveals the relatively poor quality of the website in terms of searching and filtering features. Another important observation in Buy.com is that the successful tasks followed a steadily increasing trend. This can be associated with two factors, which are familiarity and learnability. Familiarity can be defined as, "degree to which a user recognizes user interface components and views their interaction as natural; the similarity of the interface to concrete objects the user has interacted with in the past" [Usabilityfirst, 2011a]. In fact, this factor was excluded due to the fact that the users were unable to understand many of the website features. Learnability, on the other hand, cannot be measured in this study due to resources limitation (i.e. time frame and participants).

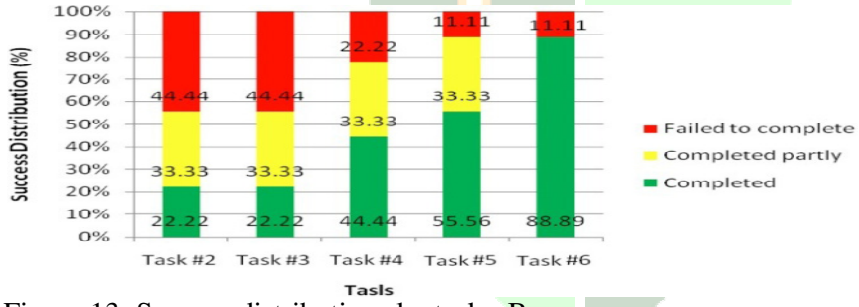

Figure 13: Success distributions by task - Buy.com.

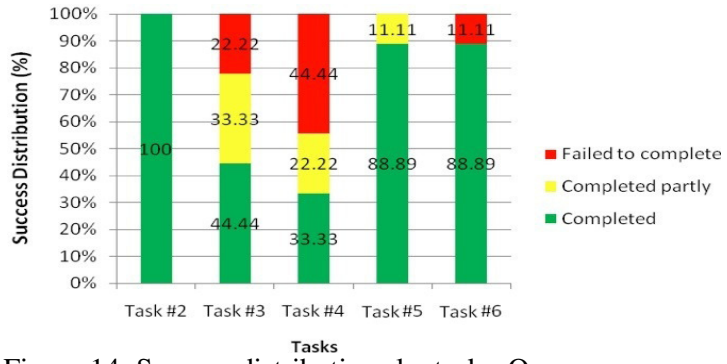

Figure 14: Success distributions by task - Qvc.com.

Usabilityfirst (2011a) defined task completion rate (i.e. success rate) as "the percentage of tasks that users complete correctly". He also used the following formula for measuring success rate:

- $Success\ rate = \frac{(successful\ tasks + (partially\ successful\ tasks) * 0.5)}{Total\ number\ of\ tasks}$

Table 4 shows the completed, partly completed and failed to be completed tasks in both websites. These figures were used to calculate the success rate. As can be seen, Qvc.com scored the higher rate (77.8%); Buy.com only scored 60%.

	Buy.com	**Qvc.com**
Successful tasks	21	32
Partially successful	12	6
Failed	12	7
Total number of tasks	45	45
Success rate	**60%**	**77.8%**

Table 4: Task success rate

Usabilityfirst (2011a) pointed out that the success rate for the majority of websites is below 50%. Maybe because this study is testing only particular aspects of the websites, both of them achieved a success rate of more than 50%. In Figure 13 and 14, it can be seen that only one user failed to complete the last task. This might be because the task was easy to accomplish, and the users had become more familiar with the websites.

Number of Errors

After finishing all the test sessions, all the errors encountered were aggregated by task, separately for each website. It can be seen from Figure 15 that 42.11% of the errors in Buy.com were discovered in Task 2. It seems that this considerable portion was due to the nature of the task, in which users were asked to use the catalogue without the search engine. The researchers designed this task to investigate how users develop different strategies to accomplish it. This resulted in detecting many usability problems. It has been noticed that the errors discovered were repeated as the test progressed. In Qvc.com, it seems that the main problem is associated with information, rather than functionality. For example, some information is missing, not clear (due to terminology), or placed in an unexpected location within the page.

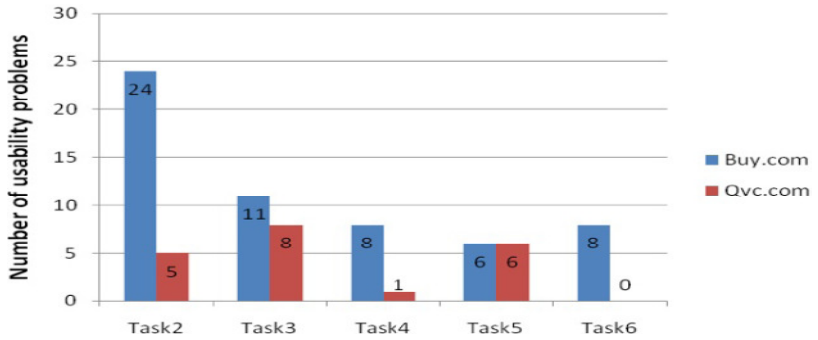

Figure 15: Number of usability problems detected by the usability testing per task

Time spent on tasks

Figure 16 compares the total time spent by all users per task. It can be seen that the users spent more time in all tasks on Buy.com. Table 5 shows the total time spent by all users

for all tasks, and also the average time that users spent in performing one task. The results of task 2 on Buy.com were clearly not satisfactory (Figure 16). This could be because the users were making the effort to understand the complex structure of the website. Also, the five minutes that was given in task 1 seemed to be not enough for familiarizing the users with the website, especially when taking into account the overuse of technological features. However, the researchers believe that if Buy.com applied some enhancements only to the website structure and to the manner in which the information is organized, the results would improve significantly.

	Buy.com	**Qvc.com**
Total time spent by all users (minutes)	17.6	11.6
Average time spent per user per task (minutes)	3.4	2.2

Table 5: Time spent on all users, all tasks

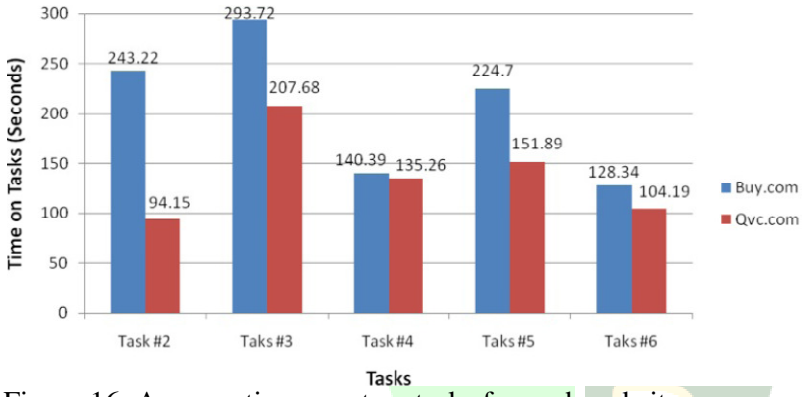

Figure 16: Average time spent on tasks for each website.

Satisfaction questionnaires analysis

User satisfaction is the third quality measure in the usability testing of this study. The main aim of this quality component is to gain better understanding of how the users perceived both websites. This was achieved by providing them with a questionnaire consisting of three types of questions [Rubin and Chisnell, 2008]: 1) Likert scale questions: users can register their degree of agreement or disagreement for each question on a five-point scale; 2) Check-box questions: users can select multiple statements as they apply to them; 3) Dichotomous question: users can make a final judgment on both websites by answering whether they would visit it again or not.

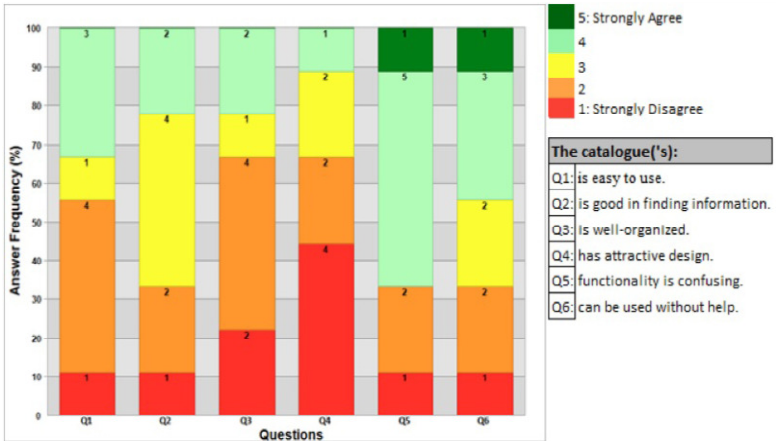

Figure 17: Users responses to the Likert scale questions - Buy.com.

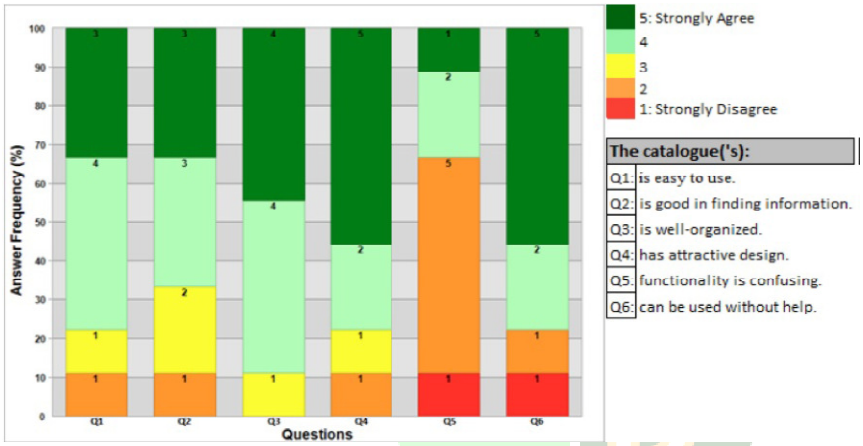

Figure 18: Users responses to the Likert scale questions - Qvc.com.

Figures 17 and 18 illustrate the percentages of the answers to each question. In these figures and for more clarification, the number of users who responded to each point scale was added at the top of each part within the columns. Generally, the light green and green colours show the users' positive responses. This is except question 5 only, as the question was originally formed as a negative. Overall, it can be seen that the participants' experience in Qvc.com was positive in comparison with Buy.com. The main differences between the websites were on catalogue organization, aesthetic design and user-friendliness, on which Qvc.com significantly scored good results. Although Buy.com employs a great many searching and filtration features, it failed to gain the anticipated acceptance. This might be related to two factors; one is the functionality of these tools because 66.67% of the users said the catalogue functionality was confusing. The other factor is that the users may have been overwhelmed by the quantity of these tools, as the catalogue offers more than 6 types of filtration tools and options. Finally, it was observed that some filters provided irrelevant results. On the other hand, Qvc.com offers only one type of filter (with three options: filter by category, brand and price). This basic approach (i.e. minimalist design), with such few options, however, did not affect the users' experience negatively as the number of users who responded positively to Question 2, which is about the ease of finding information, accounted for 66.67%. Table 6 show the users' responses to the check-box questions, which were aimed at investigating how the users found the websites' structure and navigation mechanisms.

Question	Buy.com	Qvc.com
I didn't know where to go or to look at first	55.56%	11.11%
I found that the navigation between the pages is difficult	55.56%	22.22%
I felt that I needed more time to understand the website	77.78%	11.11%

Table 6: Websites' structure and navigation mechanisms

Finally, the users were asked a fundamental question, which could summarise their overall experience. The question was about whether the user would ever choose to use the website again; a further option, which is 'not sure', was added for those who could not form a definite decision. The pie charts in Figure 19 show that 67% of the users reported that they would not use Buy.com again, while none of the users in Qvc.com registered a 'no' answer.

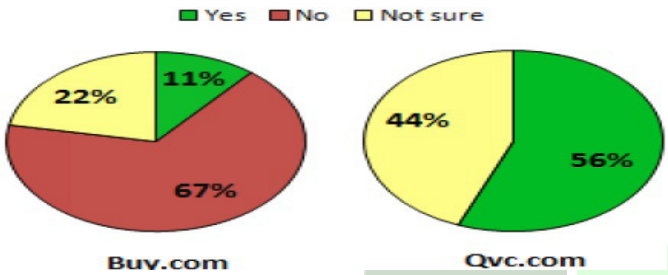

Figure 19: Users who would use the website again.

Comparison between Two Methods

After discussing the quality of the e-catalogues in the selected websites, this section will introduce the results of evaluating the performance of the modified heuristics and usability testing in terms of the efficiency, validity thoroughness and effectiveness. This will help in discovering how each method performed on two different website designs and more importantly helping the researchers in supporting this conclusion by analysing the experiments' validity.

Efficiency

Efficiency of UEMs is the "ratio between the number of usability problems detected and the total time spent on the inspection process" [Fernandez et al., 2010]. In other words, it is the relation between the quality of a UEM in terms of finding as much as possible of problems and expended resources (i.e. time). The formulae can be used to measure the efficiency of any UEM is:

$$Efficiency = \frac{Total\ number\ of\ usability\ problems\ found\ by\ a\ UEM}{Average\ time\ spent}$$

Figure 20 shows that on Buy.com, the usability testing achieved 3.31, while the heuristic evaluation scored 1.6. Consequently, the former was more efficient as less time was needed to find more usability problems. On the contrary, the same UEM was not as efficient as expected on Qvc.com in that it was less efficient by 0.30 in comparison to the heuristic evaluation.

Validity

Nielsen (1994) defined the validity, as "a question of whether the usability test in fact measures something of relevance to usability of real products in real use outside the laboratory". The UEM that is able to find a great deal of usability problems but with significant portion of unreal ones has, in fact, less validity. Sears (1997) pointed out that using the following formula will help in identifying whether a UEM is valid or not in a particular experiment:

$$Validity = \frac{Total\ number\ of\ \mathbf{real}\ usability\ problems\ found\ by\ a\ UEM}{Total\ number\ of\ problems\ identified}$$

In fact that, the validity of the heuristic evaluation on both websites was not satisfactory as it scored worst results in comparison to the usability testing in this study and also other studies (See Figure 20). On the other hand, the usability testing on both websites has achieved good and very close results. In other words, the usability testing was better in decreasing the "the false alarms" [Hartson et al., 2001]. For example, the degree of accuracy of the usability testing on Buy.com was 0.7 (i.e. 60%) higher than the heuristic evaluation. Low validity in this context refers to some problems that might have affected the experimental design. Jacko and Sears (2003) explained that there are two types of UEMs' validity. These are: internal and external validity. The former is about the extent of which the testing model is implemented correctly in that any observation can be accurately associated to particular factor(s). The latter is about the model if it can be generalized and applied to another cases out of the current study. This study is only considering external validity only.

Thoroughness

Sears (1997) defined thoroughness as the capability of a UEM to evaluate all an interface's components and characteristics in depth. The same author also explained that thoroughness can be measured by calculating the ratio of real problems discovered by a UEM to the real problems existing in the interface.

$$Thoroughness = \frac{Total\ number\ of\ \mathbf{real}\ usability\ problems\ found\ by\ a\ UEM}{Total\ number\ of\ \mathbf{real}\ problems\ exist}$$

The results show that the usability testing (0.83) was significantly more thorough than the heuristic evaluation (0.17) in Buy.com, while both UEMs achieved equal results in Qvc.com.

Effectiveness

Effectiveness of UEMs can be defined as the accuracy (validity) and completeness (thoroughness) of the results of performing specified goals [Jacko, 2007]. This is in line with the formula created by [Sears, 1997]:

$$Effectiveness = Thoroughness \times Validity$$

Achieving better results in the Thoroughness and Validity attributes for the usability testing, reflects on the overall effectiveness. The result shows that this UEM is more effective than the heuristic evaluation by (0.58) in Buy.com and by (0.21) in Qvc.com.

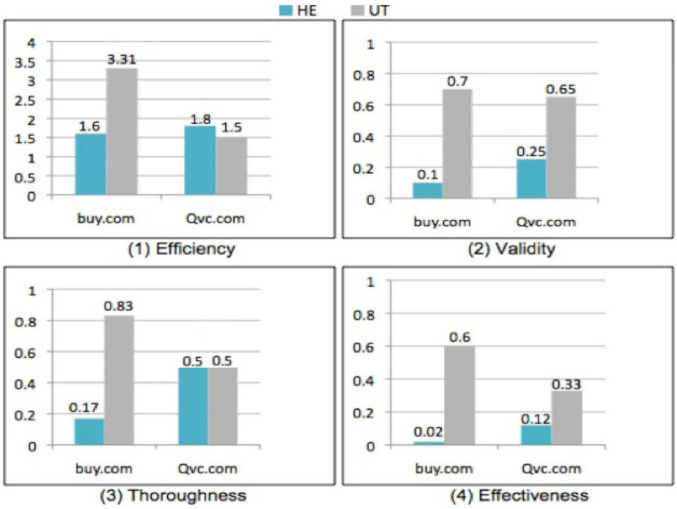

Figure 20: UEMs' efficiency, validity, thoroughness and effectiveness.

Conclusion

This study has investigated the strengths and weaknesses of modified heuristics and user testing methods. Obviously, employing different usability evaluation methods on carefully selected websites has provided some interesting results. These results can provide useful insights to identify important aspects for designing catalogue systems on shopping websites. The results show that each design of the e-catalogues has some good and bad aspects. However, Buy.com has provided better insights into common usability problems prevalent in e-catalogue systems, some of which are likely lead to confusion, such as overuse of different assistive technologies and poor consistency. In respect of UEMs, the results suggest that each method has advantages and disadvantages in terms of overall performance. For example, the modified heuristics evaluation method revealed more usability problems, while usability testing was better in detecting serious ones. Comparing the results derived from both experiments on the Buy.com e-catalogue shows that user experience is severely affected when Web 2.0 sites have poor compliance with Nielsen's traditional usability heuristics . Consequently, the findings of these studies (Hart et al., 2008; Thompson and Kemp, 2009) cannot be applied on the e-catalogue systems of shopping websites. Those studies investigated some social websites such as youtube.com and facebook.com and concluded that traditional HE ignores what is called "felt experience" such as "Significantly, pleasure, curiosity and fun, identification and self- expression, surprise and serendipity, and privacy". The results of the experiments also suggest that the usability testing can provide better insights into usability problems. In fact, this is not in line with the findings of [Nielsen, 2005c], which argued that usability testing could achieve better results only in "highly domain-dependent" systems (i.e. systems that rely on a specific knowledge-base such as "internal telephone company systems"), not in normal websites such as Buy.com and Qvc.com. Heuristics evaluation method, on the other hand, was more effective in finding more usability problems at the lowest cost and with the fewest resources. For example, modified heuristics evaluation detected 77 and 53 problems in Buy.com and Qvc.com respectively. Usability testing, in contrast, detected only 57 and 20 problems in Buy.com and Qvc.com respectively. In spite of that, heuristics evaluation was not better in finding more serious ones, and therefore, this is in line with [Jeffries et al., 1991] only in respect of finding more problems. In terms of the good and bad practices, the researchers believe that the product classification has proved to be the backbone of all online catalogues. An inadequate classification scheme, coupled with a lack of appropriate user language for describing

products, and also for using the correct keywords, will have a significant negative impact on a product's findability. It has also been observed that overuse of searching and filtration features could lead to more confusion especially when the results are irrelevant, inaccurate or unexpected. Moreover, the complex functionality of these tools is highly likely to impact user experience. On the other hand, applying a basic approach to finding information on an e-catalogue system (e.g. qvc.com), often improves end-user's interaction. However, complete, clear and organised information are further success factors of any e-catalogue system. In other words, functionality on its own does not always imply usable design. This explains why Qvc.com did not achieve expected results, as some information related to products is missing and sometimes unclear. Regardless of the variety of assistive tools in an e-catalogue system, aesthetic and minimalist design, organization and user-friendliness are the main determinants of usable e-catalogue model as extracted from the satisfaction questionnaires analysis. The security means has to be addressed carefully as it is considered one of the vital aspects of any system. However, it has been observed that presenting links to external sources within an e-catalogue can severely affect security due to the fact that the user will be moved to an entirely new website while the profile remains open.

REFERENCES

Albers, M. and Still, B. (2010). *Usability of Complex Information Systems: Evaluation of User Interaction*. Taylor & Francis.

Alshamari, M. (2010). *Task Formulation in Usability Testing*. PhD thesis, University of East Anglia, Norwich.

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

Dumas, J.S. and Redish, J.C., (1999). *A practical guide to usability testing*. Lives of Great Explorers Series. Intellect Books, Portland.

Fernandez, A., Abrahão, S., and Insfran, E. (2010). Towards to the validation of a usability evaluation method for model-driven web development. In *Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, ESEM'10, pages 54:1–54:4, New York. ACM.

Hart, J., Ridly, C., Taher, F, Sas, C., & Dix, A. (2008). Exploring the Facebook experience: A new approach to usability. *In Proceedings of NordiCHI Conference*, pages 471–474, Lund, Sweden.

Holzinger. Usability engineering methods for software developers. *Communications of the ACM*, 48(1):71-74, 2005.

Idea Group Pub., London.

ISO (1998). Ergonomic requirements for office work with visual display terminals (vdts) - part 11: Guidance on usability. Technical report, International Organisation Standards.

Jacko, J. (2007). *Human–Computer Interaction: Interaction Design and Usability*. Lecture notes in computer science. Springer, Beijing.

Jacko, J. and Sears, A. (2003). *The Human–computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications*. Routledge, Mahwah.

Jeffries, R., Miller, J., Wharton, C., and Uyeda, K. (1991). User interface evaluation in the real world: a comparison of four techniques. In *Proceedings of the SIGCHI conference on Human factors in computing systems: Reaching through technology*, CHI '91, pages 29–36, New York. ACM.

Kamper, R. (2002). Extending the usability of heuristics for design and evaluation: Lead, follow, and get out of the way. International Journal of Human–Computer Interaction, 14(3-4):447–462.

Lee, S. and Koubek, R. (2010). The effects of usability and web design attributes on user preference for e–commerce websites. *Computers in Industry*, 61(4):329–341.

Ling, C. and Salvendy, G., (2005) Extension of heuristic evaluation method: a review and reappraisal.

Magoulas, G.D., Chen, S.Y. and Papanikolaou, K.A., (1990.), Integrating layered and heuristic evaluation for adaptive learning environments. *UM2001*, page 5.

Manning, H., McCarthy, J., and Souza, R. (1998). Why most web sites fail? *Interactive Technology Series*, 3(7).

Nielsen, J. (1994). *Usability Engineering*. Morgan Kaufmann, San Diego.

Nielsen, J. (2000a). *Designing Web Usability*. New Riders, Indianapolis.

Nielsen, J. (2000b). Is navigation useful?, available at: [http://www.useit.com/alertbox/20000109.html], accessed on 5/7/2012.

Nielsen, J. (2001). Success rate: The simplest usability metric, available at: [http://www.useit.com/alertbox/20010218.html], accessed on 5/7/2012.

Nielsen, J. (2001c). Usability metrics. http://www.useit.com/alertbox/20010121.html, [Website accessed 19th Jul, 2011].

Nielsen, J. and Landauer, T. (1993). A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT '93 and CHI '93conference on Human factors in computing systems*, CHI '93, pages 206–213, New York. ACM.

Nielsen, J. and Loranger, H., *Prioritizing web usability*. New Riders Publishing Thousand Oaks, CA, USA, 2006.

Nielsen, J. and Molich, R, (1990a), Improving a human-computer dialogue. In *Communications of the ACM* 33, 3 (March), pages 338-348. ACM

Nielsen, J. and Molich, R, (1990b) Heuristic evaluation of user interfaces. *Proceedings of the SIGCHI conference on Human factors in computing systems: Empowering people,* 249-256.ACM

Nielsen, J., (1994), Heuristic evaluation. *Usability inspection methods*, pages 25-62.

Nielsen, J., (2007). *Human–Computer Interaction: Interaction Design and Usability*. Lecture notes in computer science. Springer, Beijing.

Porrero, P. (1998). *Improving the Quality of Life for the European Citizen: Technology for Inclusive Design and Equality*. IOS Press, Amsterdam.

Qin, Z. (2009). *Introduction to E–commerce*. Springer, Beijing.

Rubin, J. and Chisnell, D. (2008). *Handbook of Usability Testing: How to Plan, Design and Conduct Effective Tests*. Wiley India Pvt. Ltd.

Sauro, J. and Kindlund, E. (2005). A method to standardize usability metrics into a single score. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '05, pages 401–409, New York. ACM.

Sears, A. (1997). Heuristic walkthroughs: finding the problems without the noise. International Journal of Human–Computer Interaction, 9(3):213–234.

Shackel, B. and Richardson, S. J. (1991). *Human Factors for Informatics Usability*. Cambridge University Press, Melbourne.

Suh, W. (2005). *Web Engineering: Principles and Techniques*. ITPro collection.

Tana, W. S., Liub, D., and Bishua, R. (2009), Web evaluation: Heuristic evaluation vs. user testing. *International Journal of Industrial Ergonomics*, 39(4):621–627.

Techsmith (2011). Morae: usability testing and market research software. http://www.techsmith.com/morae/record.asp [Website accessed 18th Jul, 2011].

Thatcher, J., Burks, M., Heilmann, C., Henry, S., Kirkpatrick, A., Lauke, P.,Lawson, B., Regan, B., Rutter, R., Urban, M., and Waddell, C. (2006). Web accessibility: Web standards and regulatory compliance. In *Understanding Web Accessibility*, pages 1–51. Apress, New York.

Thompson , A. and Kemp, E. (2009). Web 2.0: extending the framework for heuristic evaluation. *In Proceedings of the 10th International Conference NZ Chapter of the ACM's Special Interest Group on Human-Computer Interaction*, CHINZ '09, pages 29–36, Auckland. ACM.

Tullis, T. and Albert, W. (2008). *Measuring the User Experience: Collecting, Analyzing, and Presenting*. Morgan Kaufmann, Burlington.

Usabilityfirst (2011a), Familiarity, available at: [http://www.usabilityfirst.com/glossary/familiarity/], accessed on 5/7/2012.

Usabilityfirst (2011b), Learnability, available at: [http://www.usabilityfirst.com/glossary/learnability/], accessed on 5/7/2012.

Webcredible (2009), Energy and water supplier website usability, available at: [http://www.webcredible.co.uk/user-friendly-resources/white-papers/utility-2009.shtml], accessed on 5/7/2012.

Webcredible (2010a), 2010 ecommerce usability for high street retailers, available at: [http://www.webcredible.co.uk/user-friendly-resources/whitepapers/ecommerce-usability-2010.shtml], accessed on 5/7/2012.

Webcredible (2010b), Local council websites: The devil is in the detail, available at: [http://www.webcredible.co.uk/user-friendly-resources/whitepapers/council-2010.shtml], accessed on 5/7/2012.

Yen, B. and Kong, R. (2002), Personalization of information access for electronic catalogs on the web, Electronic Commerce Research and Applications,1(1):20–40.

Zhou, L., Dai, L., and Zhang, D. (2007). Online shopping acceptance model – a critical survey of consumer factors in online shopping. *Journal of Electronic Commerce Research*, 8(1):41–62.