Informatics in Education, 2004, Vol. 3, No. 1, 55–74

The Heuristic Evaluation of Web-Sites Concerning the Evaluators' Expertise and the Appropriate Criteria List

Athanasis KAROULIS, Andreas POMBORTSIS
Department o f Informatics, Aristotle University of Thessaloniki
PO Box 888-54124 Thessaloniki, Greece
e-mail: {karoulis, apombo}@csd.auth.gr

Received: January 2004

Abstract. Research for the evaluation of web-sites has already begun, however it is proceeding at a very slow rate. The main reasons for this are, in our opinion, the attempt to adapt existing methodologies to the particularities of the web, the individual structure of web-sites and the issue of finding the appropriate evaluators. This study copes exactly with these points and suggests a heuristic approach for the evaluation of web-sites.

In our study we tried primarily to train the evaluators in the particularities of the heuristic evaluation; in its classic form as well as in its web-adapted form. By doing this we try to answer the core question if we can augment the evaluators' expertise with a kind of training prior to the conduction of the evaluation itself. Next we used web-adapted heuristics, found in relative literature and tried to clarify them to the evaluators as well. Finally the evaluators were involved in a real evaluation of five web sites and they wrote down their comments on appropriately prepared questionnaires.

The results from this study confirm firstly two known conclusions, that the method is applicable to the Web and that the prior evaluators' expertise is of great importance. Yet, in addition to these, we concluded that it is possible to augment, under conditions, this expertise in a short way so they have an increased performance during the evaluation as well. Our main conclusion is, however, that the used heuristic list performed inadequately, but we noted the trend of the evaluators following a somewhat similar mode of thinking, thus providing us with the way to adapt these heuristics in a more holistic approach to the web.

Key words: user interface evaluation, usability, heuristic evaluation, web evaluation.

1. Introduction

Interface evaluation of a software system is a procedure intended to identify and propose solutions for usability problems caused by the specific software design. The term "evaluation" generally refers to the process of "gathering data about the usability of a design or product by a specified group of users for a particular activity within a specified environment or work context" (Preece *et al.*, 1994, p. 602). The main goal of an interface evaluation is, as already stated, to discover usability problems. A usability problem may be defined as "anything that inteferes with user's ability to efficiently and effectively

complete tasks" (Karat *et al.*, 1992). Evaluation of user interface design is of special importance in the overall software evaluation plan, for two major reasons: Firstly because it concerns exactly that part of the software product which enables users to communicate their instructions to the machine. Evaluation should verify that the interface design delivers a friendly, intuitive and transparent yet powerful environment to end-users for the accomplishment of their goals, which in our case is the acquisition of knowledge through the interaction with the instructional environment, which in its turn supports our claim that usability affects learnability. Secondly, because evaluation of the user interface should be carried out at the right time; early enough to offer designers the chance of getting valuable feedback about their design ideas and possibly proceed to interface redesign, while all important interface characteristics have been designed and are included for evaluation.

We distinguish two major evaluation categories: *formative* and *summative* evaluation (Scriven, 1976). The former is conducted during the design and construction phase, while the latter is conducted after the product has reached the end user. The results and conclusions of the former are used mainly for bug-fixing and improving the characteristics of the interface (detecting problems and shortcomings), while the results and conclusions of the latter are used to improve the interface as a whole and meet more user needs in a following upgrade.

What do we mean by the term "usability"? According to ISO-9241 (Ergonomic requirements for office work with visual display terminals) (ISO, 1998) standard, we have the following definition: *Usability of a system is its ability to function effectively and efficiently, while providing subjective satisfaction to its users.*

Usability of an interface is usually associated with five parameters (ISO, 1998; Nielsen, 1993), derived directly from this definition:

- *Easy to learn*: The user can get work done quickly with the system,
- *Efficient to use*: Once the user has learnt the system, a high level of productivity is possible,
- *Easy to remember*: The casual user is able to return to using the system after some period without having to learn everything all over again,
- *Few errors*: Users do not make many errors during the use of the system or if they do so they can easily recover them,
- *Pleasant to use*: Users are subjectively satisfied by using the system; they like it.

Two important conceptions regarding the usability of an interface are "transparency" and "intuitiveness" (Nielsen, 1993; Preece *et al.*, 1994; Shneiderman, 1998). Transparency refers to the ability of the interface to fade out in the background, allowing the user to concentrate during his work on *what* he wants to do and not on *how* to do it, in our case not interfering with the learning procedure (Roth and Chair, 1997), while intuitiveness refers to its ability to guide the user through it by the use of proper metaphors and successful mapping to the real world (Shank and Cleary, 1996), e.g., by providing him with the appropriate icons, correct labeling, exact phrasing, constructive feedback etc.

2. The Heuristic Evaluation

Maybe the most frequently encountered evaluation method, of any entity, is the provision of a list of criteria relative to this entity followed by questioning in order to express peoples' opinion. These people can be users or experts in the particular domain. So we distinguish, as already explained, between user based evaluations, known as "empirical evaluations" and expert based evaluations. However, at this point we have to make some clarifications about the notion of the user. Referring to the web we consider de facto that all involved persons are at the same time users, even if they deal with it as evaluators. "Real" user based evaluations assume that the users use the entity under consideration under conditions as realistic as possible, while, simultaneously, observations are collected about the evaluation procedure. However, as already mentioned, in the evaluation case under consideration there are some criteria set, which have to be followed during the evaluation. In the web evaluation approach, where every evaluator performs on his/her own, these are sometimes assessed without real use of the entity, but the user or the evaluator usually utilizes the *conceptual model*, as described by Norman (1988) for the entity and the way it performs, simulates its performance in his mind and concludes for every criterion. Alternatively he may use the entity, not to produce real work, but in order to assess the application of the criteria. So we can argue that in any case it is about an expert based evaluation approach, even if users are involved, as long as they are concerned about answering according to the set criteria.

What can we evaluate in this way? Makrakis (1999) says everything has to do with:

- the design;
- the organization;
- the function;
- the result of the entity under consideration.

To evaluate the above he assumes as nessesary:

- *Defining the evaluation axis.* They are the general questions set to be answered through the evaluation. They emerge
  1) from what we need to have evaluated and
  2) what the evaluation method allows us.

These axis are the basic principles of the *theoretical framework* of the evaluation.

- *Defining detailed criteria*. They are the concrete questions, usually measurable variables (so they are components of the *methodological framework*) to asssess the axis.

However, a number of problems arise from this approach.

- It provides all the disadvantages of the expert-based evaluations (Karat *et al.*, 1992; Nielsen, 1993b; Karoulis *et al.*, 2000).
- The axes and criteria list may become very long (Lewis and Rieman, 1994; Nielsen, 1993b). For example, the full interface usability criteria list suggested by Smith and Mosier (1986) includes 944 criteria.
- The evaluators' expertise plays a major role. (Lewis and Rieman, 1994; Nielsen, 1993). We discuss this issue in detail later.

To handle these problems Jacob Nielsen and Rolf Molich started their research in 1988 and in 1990 they presented the “heuristic evaluation” (Nielsen and Molich, 1990). The basic point was the reduction of the set criteria to just a few, at the same time being broadly applicable and generally agreed; simultaneously augmenting the evaluators’ expertise, and consequently their reliability. These “heuristic rules” or “heuristics” derived from studies, criteria lists, on field observations and prior experience of the domain.

The core point to evaluate in the initial approach is the usability of the interface. Based on the ISO principles about usability (ISO, 1998), Nielsen (1994) stated following heuristics, slightly modified and reorganized by us:

1. *Simple and natural dialog and aesthetic and minimalistic design.* Dialogs 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.
2. *Visibility of the system status – provide feedback:* The system should always keep users informed about what is going on , through appropriate feedback within reasonable time
3. *Speak the users’ language: match between system and real world*. The system should speak the user’s 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.
4. *Minimize the users’ cognitive load: recognition rather than recall.* Make 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.
5. *Consistency and standards:* Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.
6. *Flexibility and efficiency of use – provide shortcuts.* Accelerators – unseen by the novice user – may often speed up the interaction for the expert user to such an extent that the system can cater for both inexperienced and experienced users. Allow users to tailor frequent actions
7. *Support users’ 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.
8. *Prevent errors:* Even better than good error messages is a careful design which prevents a problem from occurring in the first place.
9. *Help users recognize, diagnose and recover from errors with constructive error messages*. Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution
10. *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.

In the heuristic evaluation we make two assumptions from the beginning (Lewis and Rieman, 1994), which have evolved from the observations of the application of the method:

- No distinct evaluator can find all the heuristically identifiable usability problems.
- Different evaluators find different problems.

The appropriate number of evaluators and their expertise are an issue of great importance. Researches up to now (Nielsen and Molich, 1990; Nielsen, 1992; Nielsen, 1993b) have shown that:

- *Simple or novice evaluators.* They do not perform very well. We need 15 evaluators to find out 75% of the heuristically identifiable problems. These are problems that heuristic evaluation can point out. As already mentioned and for different reasons, there are problems that are overlooked using this kind of evaluation. The research has shown that 5 of these simple evaluators can pinpoint only 50% of the total problems.
- *HCI experts (regular specialists)*. They perform significantly better: 3 to 5 of such evaluators can point out 75% of the heuristicaly identifiable problems and among them all major problems of the interface.
- *Double experts (double specialists)*. These are HCI experts with additional expertise on the subject matter, e.g., educators for educational interfaces. The reasearch has shown that 2–3 of them can point out the same percentage as the HCI experts.

The following figure by Nielsen (1992) summarizes these statements (Fig. 1).

It is obvious that there is no great difference between experts and double experts in seeking the involvement of the latter in the evaluation. However, there is a very distinct difference between experts and simple evaluators. As we can see in the figure, to point out 75% of the heuristically identifiable problems we need 15 simple evaluators, while 3 expert evaluators bring the same result.

The method refers mainly to traditional formative human-computer interface evaluation, yet a number of studies (Nielsen and Norman, 2000; Instone, 1997; Levi and Conrad, 1996) has proven its easy adaptability to the evaluation of web sites as well.

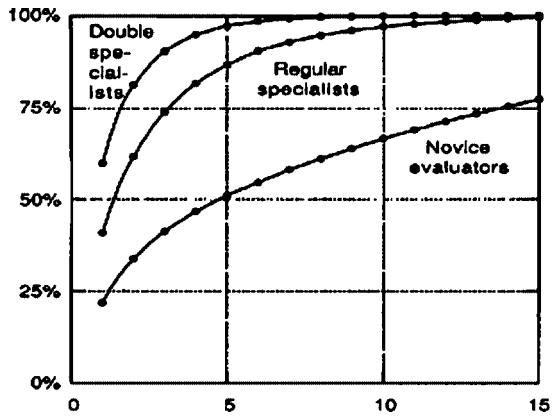

Fig. 1. Expertise of the evaluators.

3. Adaptation to the Web

Evaluation in the web differs from the traditional evaluation methodologies in many ways, due to the particularities of the web: every web site is an information space with non-linear structure, so two parameters, the download time and the ease of navigation, are of great importance. In addition to this, the evaluation procedure can be conducted by every evaluator on his/her own, redefining the notion of the "evaluation session" and introducing the notion of the "asynchronous evaluation", since the evaluators can perform their work from different places and at different time intervals. Finally, as already stated, in the web every evaluator is at the same time a user. Norman (2000) presents, for example, a cognitive walkthrough (Wharton *et al.*, 1992; Lewis *et al.*, 1990; Karoulis *et al.*, 2000) performed in the web, playing the role of the simple user and thus proving the efficiency of this combination. This particular occurance on its own adds the hue of the empirical evaluation to the expert based evaluations in the web as well, augmenting its reliability, since the combination of user based and expert based approaches seems probably to provide the best results (Karat *et al.*, 1992; Kantner and Rosenbaum, 1997; Karoulis and Pombortsis, 2000; Karoulis *et al.*, 2000). The adaptation of the heuristic evaluation in the web has been already studied by some researchers (e.g., Instone, 1997; Levi and Conrad, 1996) and the results are in agreement that, in general, it is effective. Other researchers however consider that this issue has not yet been researched enough (Trochim, 1996; Lowe, 1999), and we adopt that opinion too.

The five points of usability (easy to learn, efficient to use, easy to remember, few errors, pleasant to use) have to stand for all kinds of interfaces, but Lowe (1999) gives a more concrete picture, especially for the web, focusing on some more specific demands:

- effective underlying navigation structure;
- mechanisms for supporting ongoing maintenance;
- the extent to which the site achieves the objectives of the client for whom it was developed;
- compliance to standards and structural integrity;
- look, feel and "userfriendliness" of the site;
- consistence across different browsers and platforms.

As we shall discuss in the results of this study, these axes are proved to be more realistic and suitable for the web and based on them we shall propose a modified heuristic list.

4. Research Questions

Given the youth of the web technology and the speed at which the web is growing, it seems indispensable that researchers conduct steady studies on its parameters, which are more often than not very variable. Bearing this in mind, in this study we are concerned with the specialized adaptation issues of the heuristic evaluation to the web and our main questions are as follows:

1. Can we, mainly, apply the heuristic methodology in the web?
2. Can the power users' expertise be augmented through some kind of "training", so that they can perform as well as expert evaluators?
3. Is the same list of heuristics valid for the web as for the evaluation of traditional interfaces?

The first question, as already mentioned, is almost answered in the affirmative, so one more piece of evidence will strengthen this view.

Creating Experts

Let us now consider the second question. It is known that it is possible for computer scientists to easily learn the evaluation methodologies and apply them sucessfully (Nielsen, 1992a; Wright and Monk, 1991). But computer scientists (the "experts") are not yet available in great numbers, so one can't argue that he/she will find an adequate number eager to conduct the evaluation. So the following question arises; can some power users be trained in the heuristic methodology and be allowed to play the role of the expert? These "power users" could be, for example, computer science students. Let us note at this point that simple users tend to be outside of the scope of this study, because of their already reported inadequate performance during the evaluation.

The transistion of the novice to expert with the passing of time has occupied many researchers. Is starts mainly from the question of "how do we define the novice and the expert user". Demetriades (2000) argues that it is not about a quantitative differentiated accumulation of knowledge between two different human categories. What differentiates novice from experts is basically the different representations they possess for the entity, and, consequently, for the problems they are supposed to solve. Indeed, a series of studies (Larkin, 1983; Chi *et al.*, 1981) shows that the mental representations of the novice are strictly restricted to the surface characteristics of the problem, which is expected, since they are known and familiar. Contrary to the above, experts possess the ability to correlate these surface characteristics to deeper principles, in representations and subtractions of a higher level and proceed to efficient solutions.

Anderson (1995) gives an analytical description of the procedure of the development of the novice to knowledgeable expert for different cognitive domains. But it is Vygotsky (1978) who defined the "zone of proximal development", providing thus a complete explanation of how, according to him, the augmentation of the knowledge happens. According to Vygotsky, every human finds him/herself at a particular level of development, where he/she can solve a set of problems using his/her own abilities. However, there is a superset of problems the person could solve as well with the aid of more competent agents, such as his/her books or his/her teachers. This almost mathematical difference between the "level of real development" to the "level of possible development" builds, according to Vygotsky, the "zone of proximal development". That is to say that this zone consists of functions and skills not yet developed, but ready to emerge or have just entered the state of maturity. These abilities can be developed, in a slower manner on their own as well, while the experience of the person is augmented by dealing with the problems

under consideration. In our particular case, while the user occupies him/herself with the particular system, he/she gains experience, which in its turn is the crossing of the "zone of proximal development" by Vygotsky. Obviously, this zone doesn't cease existing, but it is translocated further.

So our question refers in particular to how far a short training period can help the evaluators cross this zone in a feasible time towards the application of the heuristic evaluation methodology.

The Heuristic List

The initial heuristics have been adapted and commented by Instone (2000) for their application in web-based heuristic evaluations, as follows:

1. *Visibility of system status*. The system should always keep users informed about what is going on, through appropriate feedback within reasonable time. Probably the two most important things that users need to know at your site are "Where am I?" and "Where can I go next?"
2. *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.
3. *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.
4. *Consistency and standards*. Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.
5. *Error prevention.* Even better than good error messages is a careful design which prevents a problem from occurring in the first place.
6. *Recognition rather than recall*. Make 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.
7. *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.
8. *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.
9. *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.
10. *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.

So the third question set in this study is if these heuristics are appropriate for the web and really efficient in the way Instone (2000) declares them to be.

5. Adaptation, Organization and Conduct of the Evaluation

From what has been stated up to now it is obvious that our evaluation provides the following schematical form (Fig. 2).

The application of this approach has been materialized as follows: The evaluators conducted the evaluation in their own environment and using their own mode of internet connection. This is obviously the first major difference between the web-based and the traditional evaluation approaches, the transition therefore from the "evaluation session" to the "asynchronous evaluation session". The first immediate consequence is the need to train the evaluators in a written manner in all necessary detail to fully clarify the procedure, but not to an excessive degree, so that phenomena of discouragement occure.

Therefore we prepared a booklet, which we titled "Notes to the Evaluators", consisting of 7 pages, and describing the methodology of the heuristic evaluation by Nielsen (1992; 1994a and 1994b), as well as its adaptation in the web, and finally the description of the procedure the evaluators had to follow to complete their work. In addition to this, they have been equiped with another booklet, consisting of 5 pages, containing the web-adapted heuristics of Instone (2000) and their comments. During the procedure we were asked to translate them in Greek, and we did that. The heuristics in this booklet were more detailed than provided here, in an attempt to clarify them fully and augment the level of the provided training. From that point the evaluators were asked to follow these web-adapted heuristics instead of the "originals" by Nielsen, to conduct their evaluation.

The training material included an "Evaluator's Notebook" as well, where the participants could note down their assessments and opinions.

The session started at the beginning of December 2000 and was completed at the end of February 2001, a duration of three months. During this period no web site changed its structure, which is considered to be good luck, since another cause for concern is the frequent change of the form and the content, due to steady maintenance, of the web sites; a fact that is completely contrary to the prolongation of the duration of the evaluation session.

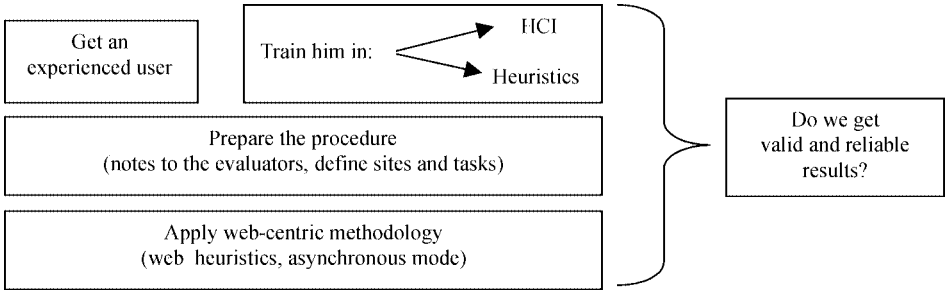

Fig. 2. The preparation of the evaluation.

We used in total 40 evaluators working in 5 sites, 2 major international sites, 1 small international site and 2 small Greek sites. So we had a total of $40 \times 5 = 200$ evaluations performed. In addition, the fact that we used 3 small sites for our evaluations provided us with an adequate number of heuristically identifiable problems, since smaller sites usually lack the possibility for extended usability testing, as major sites do.

6. Results

Before starting the presentation of the results of this study, we would like to emphasize that the object of this study was not to evaluate the web sites under consideration, but to answer our research questions on the efficiency of the method and the chosen evaluators on the web. With this view in mind, we omit the results of the evaluations concerning the usability of each particular site. A direct consequence is that the aggregation of the evaluators' opinions is no longer necessary, as suggested by Nielsen (1994b) in order to obtain the evaluation results about the sites. It was more necessary to categorize the opinions of the evaluators to make the answers visible to our research questions. In order to achieve this, we set some secondary questions during the elaboration phase which in their turn would lead us to clarify the main questions. These secondary questions are:

1. How many of the given heuristics were understood by the evaluators in a manner that they were able to apply them in the evaluation procedure? (3rd question – reliability of the heuristics)
2. How many usability problems did they recognize in the sites visited? (2nd question – successful training of the evaluators)
3. Did they successfully rate the severity of every encountered problem? (2nd question – successful training of the evaluators)
4. Did they successfully correspond to every problem with the appropriate heuristic or heuristics? (2nd question – successful training of the evaluators and 3rd question – reliability of the heuristics).

To proceed, we needed to group and categorize all the opinions of the evaluators in a separate supplementary document. Below we provide a sample of this document.

Evaluator: Mr. X
Site: www.a-site.com

Problem	Evaluat. Severity	Real Severity	Evaluator's Heuristic	Real Heuristic
Navigation (where am I / where can I go?)	minor	Major	1	1,2,3,6,10
Badly designed links (not standard)	Major	Major	4	4
Missing help	Major	Major	10	10
Insufficient functionality (not user-centered)	Major	Major	2,5	2,3,7
Inappropriate function	Major	minor	6,10,8	8
Semantic (inappropriate / misleading etc)	Major	Major	2	2,5

In this document we provide the evaluators' opinions, as well as our assessments about the different heuristically identifiable problems. This approach can be found in the relevant literature (Nielsen, 1992; Lewis and Rieman, 1994) and relies on the observation

that the conductors, in this case us, who are obviously HCI experts, can point out the majority of the usability problems during the preparation phase of the evaluation, that the evaluators will discover later on. This fact is a consequence of the discovery that a few double experts can pinpoint most of the problems. However, in our study this issue has proved to be insignificant, because the problems that were not discovered by us, but were found by the evaluators during the session, could be rated on their severity afterwards and matched with the corresponding heuristics. We noted our opinions in columns, next to those of the evaluators'.

It is obvious from the table above that we consider that every problem can oppose more than one heuristic. Let's consider an example: Evaluator X has noted: "it is not clear to the user if the particular graphic element means 'back to previous page' or is just decorative". This opinion may have more than one interpretation, as "badly designed graphics" and "difficult navigation" and "possibility for a user error", it therefore opposes heuristic 8 (aesthetic and minimalist design), heuristic 1 (visibility of system status) and heuristic 5 (prevent errors), but also heuristic 4 (consistency and standards), 6 (recognition than recall) and 7 (flexibility and efficiency of use), as well as setting its severity on 'Major'. Consequently, if the evaluator noted opposition to heuristic 8 (aesthetic and minimalist design) and 6 (recognition than recall), is considered to be successful, yet if he noted opposition to heuristic 10 (help and documentation), is considered to be a failure. We have to point out that we avoided assigning many heuristics to every problem, but limited it to the maximum of 5 heuristics per problem. As well as there are obviously many problems that correspond to exactly one heuristic, for example heuristic 10 (help and documentation – there is either help, or not).

For the severity rating of the usability problems we did not use the categorization Nielsen (2000c) suggests, that is to say a five graded scale from 0 to 4 (0 no problem – 4 catastrophic problem). The reason for this decision is that in the same study Nielsen reports that characterization of the severity of the problem from only one evaluator as being very unreliable and suggests the aggregation of the characterizations of at least three evaluators. While we have also pointed out the same difficulty in some of our former studies (Karoulis and Pombortsis, 2000; Karoulis *et al.*, 2000b; Karoulis *et al.*, 2001), and, in addition to this, difficulty with the statistical elaboration of the results. So we decided to ask the evaluators to categorize the problems into two severity categories, as "Major" and "minor", as Nielsen suggested in a former study (Nielsen, 1992). We consider "Major" problems to be the ones that have serious potential for confusing users or causing them to use the system erroneously, while "minor" problems may slow down the interaction or inconvenience users unnecessarily.

Finally we applied a non-structured interview approach with the evaluators. In an informal conversation with every one of them we asked some general questions, such as:

- Their general impression from the evaluation (Question 1 – can a heuristic approach be applied in the web?)
- If the booklet has helped (Question 2 – can the evaluators' expertise be augmented?)

- Do they think that they now know some things about heuristic evaluation and the evalation methodologies in general (Question 2 – can the evaluators' expertise be augmented?)
- Their opinion about the heuristics (Question 3 – appropriateness of the heuristics list)
- If the whole procedure was interesting to them (subjective satisfaction)
- If they finally think that such an evaluation can improve the quality of web sites (Question 1 – can a heuristic approach be applied in the web?)

7. Conclusions

The answers to our questions can be given briefly as follows:

- *Can we, mainly, apply the heuristic methodology in the web?*
  The answer to this question is affirmative, which is in agreement with most of the studies up to now. However, in order to apply the method effectively, the results of the following points must be taken into consideration as well.
- *Can the power users' expertise be augmented through some kind of "training", so that they can perform as well as expert evaluators?*
  Yes and no. This also confirms the results from previous studies, that report the experts performing very differently from the simple users. However, this question is more complicated and will be discussed in detail later.
- *Is the same list of heuristics valid for the web as for the evaluation of traditional interfaces?*
  The answer, according to our study, is negative. The heuristics we used seemed not to facilitate the evaluators in their work. They stated that they "interpreted" them to be applicable in different instances, and they provided us with some hints as well.

In more detail, heuristic evaluation performs well even in the web, yet the main issues of the evaluators' expertise and the validation of the web-heuristics remain.

The starting point for this study was the question if we could involve only power users, eg computer science students, instead of the difficultly in finding HCI experts. Of the 40 evaluators, 28 were skilled computer scientists or information technology teachers, while the rest were computer science students with "normal" expertise, yet nobody had experience in the HCI domain. The results have shown that five of the "experienced" evaluators had no difficulty in interpreting and applying the method, while eight "experienced" and the twelve "normal" did not fully understand the method and did not perform well. So we can say that, in general, it is possible to create experts, who are able to participate in web-heuristic evaluations, since twenty of our subjects succeded; however we can not ignore the fact that the other twenty could not perform well. The separating line between these groups is not clear and our study can finally only approve the results of former studies (Nielsen and Molich, 1990; Nielsen, 1992) that suggest careful selection of the evaluators.

According to the mode of the training, 16 of the "successful" evaluators considered the booklet as "very lucid and enlightening", 12 considered a face-to-face seminar as a better

solution without an optional booklet, while the rest had no opinion about this issue. There is additional evidence on this issue by Nielsen and Mack (1994), that heuristic evaluation can be taught in a half-day seminar, so this proposed approach seems to be a better one.

Regarding the third question on the appropriateness of the heuristics, it was clear that the used heuristics did not even facilitate the "successful" evaluators. On the contrary, some suggestions were made to us, as well as our collecting the evaluators' comments which resulted in a more lucid web-adapted heuristic list that seems to be more familiar and appears to facilitate the procedure.

8. Discussion

Collecting the evaluators' answers we distinguished the following categories of discovered problems:

- Navigation (where am I / Where can I go?)
- Semantic (inappropriate / misleading etc)
- Switching between Greek and English
- Restriction (navigation and user tasks)
- Lack of feedback
- Misfunction (software/design errors)
- Insufficient functionality (not user-centered)
- Lack of search function
- Inconsitency between menus and functions
- Bad spatial management – bad graphics
- Dangling and dead links
- Insufficient / too much information
- It is an on-line booklet
- Different icons lead to the same point
- Inconsistency of title and link (semantic)
- Scarring (errors, changing site, pop-up windows)
- Removal of user control
- User's cognitive overload
- Impossible to find a communication mode
- Unsuitable (improper) function
- Bad design (for functionality)
- Lack of help
- Badly designed links (not standard)
- Irrelevant information
- Obscure information

Most of these categories adhere to one or many of the above mentioned heuristics, as already commented. However, in this list there are categories that refer to the content of the web-site, its design as regards functionality or for not supporting the task the user wants to perform. These issues are obviously the concern of the user-centered design, which in the evaluators' opinion has not been applied, but is unavoidable, especially if one is designing for the web (Lewis and Rieman, 1994).

At the same time, the heuristics for the help era have been aggregated to just one: "lack of help". The evaluators, being at the same time users, as previously mentioned, do not seem eager to distinguish *why* an error happened, how the user could *avoid it*, and how he is going *to recover* from it. Instead, they demand good error handling from the designers, which includes all these subcategories.

We are able after all to propose the list of heuristics that emerged from this study to be more web-centric and more user-centric. Let us mention at this stage that in the assembling of this list we mainly took into consideration the results of our study and the evaluators' comments. However, all the aforementioned issues, such as the heuristics of Nielsen (2000a) and Instone (2000), as well as the prososals of Lowe (1999), which have been mentioned earlier, still remain as the underlying structure. Finally we also took into consideration the work of Togniazzini (2000), who proposes criteria, not for

the evaluation process itself, but for the design of web-sites, which are very close to the results of this study.

Below we present the list. Its structure is slightly different than usual: it consists of axis, which contain criteria as follows:

Axis 1: **Visible system status and in correspondence to what the user expects**

- *Criterion* 1.1: Navigation. Is it obvious where I am and where I can go next?
- *Criterion* 1.2: Are all the icons and/or navigation possibilities visible and is it clear where they lead?
- *Criterion* 1.3: Are all semantics clear and all functional graphics (this doesn't include decorative graphics) clear as to what they do? Are the used "metaphors" of the icons and the graphics clear as to what they mean?
- *Criterion* 1.4: Is consistent language used, are international standards respected, is there a unique and consistent way of presenting the information?

Axis 2: **Flexibility of use and structural integrity**

- *Criterion* 2.1: Are there the necessary "accelerators" available? For example can all pages be bookmarked with correct titles?
- *Criterion* 2.2: Has the site been debugged? Are there any empty areas or dangling and dead links? Is the encoding correct?
- *Criterion* 2.3: Does the site follow the conventions of the web, e.g., colour of hyperlinks, presentational structure etc.?
- *Criterion* 2.4: Does the site support its exploration? Is there a site map, search function etc.?
- *Criterion* 2.5: Can the user easily remember the stucture, the functional and navigational mode at his/her next visit to the site?

Axis 3: **Efficiency of use**

- *Criterion* 3.1: Are the technologies wisely used? That is to say, does the site use the exact technologies the user awaits to see? Are these technologies acceptable for all user configurations that will visit the site?
- *Criterion* 3.2: Are the response times of the site in line with what the user expects? This means, have the designers taken into consideration the different users' speeds?
- *Criterion* 3.3: Does the site adhere to the independent philosophy of the web? Therefore, can every user with any equipment and any browser perform the tasks he aims to perform in the site?
- *Criterion* 3.4: Does the site provide direct access to the most common tasks one can perform in the site, or does the user have to cope with dialogs and choices?

Axis 4: **User control, user-centered design and interaction**

- *Criterion* 4.1: Can the user completely control all the interactive elements? Is this control taken away from the user at any time?

Criterion 4.2: Are there the appropriate interaction elements corresponding to the tasks that the user aims to perform in the site?

Criterion 4.3: Is the feedback of these interaction elements of the kind the user expects, or do they surprise him/her?

Criterion 4.4: Does the site support all the tasks the common user of the particular site aims to perform?

Criterion 4.5: Can the user perform the tasks of his/her interest with minimal cognitive load? This means, does the system facilitate him/her, or does he/she have to find out the way by him/herself?

Axis 5: **Content and presentation**

Criterion 5.1: Is there the right amount of information in the site (not insufficient or excessive)? If vast amounts of information must be included (e.g., older files and/or records) is it structured in levels?

Criterion 5.2: Is there the right quality of information in the site (valid, clear, apropos) that the user aims to find?

Criterion 5.3: Does the site give the impression of having been constructed and then left on its own, or is it regulary maintained? Is there any outdated information?

Criterion 5.4: Is the information presented in a web-centric way, or is it just an adaptation of printed material?

Criterion 5.5: Is the information presented graphically acceptable, in accordance with the web-publishing principles, e.g., colours, white space management, navigational elements etc.? Easy to read?

Axis 6: **Subjective satisfaction, communication and help**

Criterion 6.1: Does the user feel he/she is isolated or left on his/her own? Does he/she have any communication facility, or is the site very impersonal?

Criterion 6.2: Is the site, in general, pleasant to use? Encourages exploration? Has the site alterations? Does the user feel the site is under his/her control?

Criterion 6.3: Is there help, search function, external help, glossary or something else to facilitate the user in performing his/her tasks?

The approach of building the list in axis containing criteria supports its application in two forms, as it may be obvious. One is the compact form – only the axis – if there is a shortage of resources (time, money etc.) or if we have very experienced evaluators available (double experts). The other one is the analytic form – all the criteria – for a more detailed evaluation of the site.

The Severity Rating of the Problems

Completing the discussion of the results we state that the issue of the categorization of the problems as “Major” or “minor” has shown the following. The evaluators that understood the method and performed well had a great percentage of agreement with our categorization of the problems as “Major” or “minor”. Twelve of them had double

the amount of agreements than disagreements, while all agreed to more severity ratings with us, even marginally. On the other hand, the evaluators that had difficulties, provided non-identifiable behaviour. Eight of them provided more disagreements than agreements with us, eight provided a very large percentage of agreement with us, more than the double the other team provided, and four evaluators almost unanimity (12 agreements – 2 disagreements). So we can argue that, under the condition that our evaluators perform well, the severity ratings are in accordance with all previous studies of the method up to now, that is that one evaluator's ratings are not reliable, in contrary we have to aggregate the ratings of more than one evaluator for every heuristic. However, we do not recommend analyzing the severity in more levels, especially for power users who are being trained in the heuristic evaluation methodology, but remaining in the bipolar characterization as "Major" and "minor" is recommended.

Proposals for Further Research

Further research has to take into consideration some core points. The general impression is that the method is applicable and provides some great advantages as well, which are advantages of the heuristic evaluation. In general it is cheap, fast and easy to apply; the experts are more easily located than the users, despite the difficulty in locating them, and it is very efficient, according to the problems it discovers in relation to the effort and the resources needed (Nielsen, 1990; Nielsen, 1992; Levi and Conrad, 1996). Consequently, the efficiency of the form used in this study can be considered as a good starting point, as we estimate that any further improvement of the method from now on will provide significantly improved results. Under this point of view we propose some modifications, which however have first to be theoretically stated and experimentally validated as well.

In more detail, we propose the use of the former cited criteria. Levi and Conrad (1996) argue that, finally, every system has to be assessed according to its usability. This is of greater importance in the web, where the main concern is in how far the system facilitates the users to perform their intended tasks and how far the users are satisfied with their experience. The proposed list has its locus on the notions of *web-centricity* and *user-centricity*, which are considered to be of major importance to the web, yet they are constantly neglected.

This list can be combined with a Likert scale as well, in order to make a quantitative assessment of the evaluators' opinions, an approach that could lead to a kind of "gradation" of the site according to these heuristics. However, the Likert scale gradation is criticized to be monosemantic. Mathematicians will love it, however sociologists are concerned about the validity of the results it provides. These are known side-effects and are not within the scope of this work. We propose its use because of the formative nature of the heuristic approach: it can be iteratively applied during the design cycle and designers need "quick and dirty" results to improve the system in feasible time, so the evaluation approach has to be cost effective. On the other hand, if, for example, one wants to investigate issues specific to students from different cultures or other socially relevant issues,

then a more open-ended questioning format should be more appropriate. However, one should be aware that the elaboration of such kind of data is very time consuming, and during a formative evaluation procedure this could sometimes even be unacceptable. This is obviously another concern for further investigation.

Another issue for further study is the transition to the "asynchronous evaluation". This fact brings with it some positive and negative implications, such as the different access speed of the evaluators to the site under consideration, or the absence of the instant access a conducter could provide, since he/she is spatialy absent from the session, or the great prolongation of the session time. These issues have not yet been investigated enough in the studies conducted so far, but they were also not in the scope of this study. However it was obvious from the evaluators' reactions that the problem existed. We have been asked repeatedly for clarifications concerning the methodology or the sites themselves. So if a question arose, should we answer it and to whom? Only to those who asked (training our evaluators in an unequal way), to all (modifying thus our approach), or not at all (making their life difficult)? Finally we opted for the last choice that is not changing our approach, however the issue of the asynchronous evaluation remains an open issue.

Summarizing the above, we argue that firstly special care must be taken in carefully selecting the evaluators, so that they have the nessesary expertise in computer science. Secondly, one has to follow a training approach with a seminar in addition to the booklet and finally use our proposed list of criteria that seems more familiar to the particular evaluator category. As a final conclusion to all the above, we believe that the method will finally have enough potential to provide an alternative solution in a situation where there are not HCI experts available to perform the evaluation.

Acknowledgements

The authors want at this point to thank our evaluators for their willingness in participating in this evaluation, their consistency and for having volunteered their time. It is clear that without their help this study may never have been finished.

References

Anderson, J.R. (1995). *Cognitive Psychology and its Implications*, 4th edt. W.H. Freeman and Co., New York.

Chi, M.T.H., P.J. Feltovich and R. Glaser (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, **5**, 121–152.

Demetriades, St. (2000). *Mutlimedia Technology and its Application in the Educational Procedure.* Doctoral Dissertation, A.U.Th, Dept. of Informatics (in Greek).

Instone, K. (1997). *Site Usability Evaluation*. Retrieved 9 Dec. 2003:
`http://user-experience.org/uefiles/writings/siteeval.html`

Instone, K. (2000). Usability Heuristics For The Web. Retrieved 9 Dec. 2003:
`http://user-experience.org/uefiles/writings/heuristics.html`

ISO 9241 – International Standarization Organization (1998). *Ergonomic Requirements for Office Work with Visual Display Terminals (VDT's)*.

Karat, C., R. Campbell and T. Fiegel (1992). Comparison of emperical testing and walkthrough methods in user interface evaluation. In *Proceedings of ACM CHI '92*. Monterey, CA, pp. 397–404.

Karoulis, A., S. Demetriades and A. Pombortsis (2000). The cognitive graphical jogthrough – an evaluation method with assessment capabilities. In *Applied Informatics 2000 Conference Proceedings*. Innsbruck, Austria.

Karoulis, A., S. Demetriades and A. Pombortsis (2000b). Evaluation of multimedia educational interfaces for the junior highschool classes using two different methodologies: the "Perivallon" experience (under submission).

Karoulis, A., and A. Pombortsis (2000). Evaluating the usability of multimedia educational software for use in the classroom using a "Combinatory Evaluation" approach. In *Proc. of Eden 4th Open Classroom Conference*. Barcelona, Spain.

Karoulis, A., St. Demetriades and A. Pombortsis (2001). Evaluation of multimedia educational interfaces using a "combinatory evaluation": the "Perivallon" exerience. In *Proceedings of Binding Together Secondary and Tertiary Education Conference*. Thessaloniki (in Greek).

Larkin, J. (1983). The role of problem representation in physics. In D. Gentner and A. Stevens (Eds.), *Mental Models*. Lawrence Erlbaum, Hillsdale, New Jersey, pp. 75–98.

Levi, M.D., and F.G. Conrad (1996). A heuristic evaluation of a world wide web prototype. *Interactions Magazine*, **III**(4), 50–61.

Lewis, C., P. Polson, C. Wharton and J. Rieman (1990). Testing a Walkthrough methodology for theory-based design of walk-up-and-use interfaces. In *Proceedings of ACM CHI '90*. Seattle, Washington, pp. 235–242.

Lewis, C., and J. Rieman (1994). *Task-Centered User Interface Design – A practical Introduction*. `ftp.cs.colorado.edu/pub/cs/distribs/HCI-Design-Book`

Lowe, D. (1999). Web site evaluation. *WebNet Journal*, **1**(4), Oct-Dec.

Makrakis, B. (1999). Evaluation of open and distance learning systems. In *Open and Distance Learning*, Vol A', Greek Open University, Patras, Greece, pp. 245–301.

Nielsen, J. (1990). Big paybacks from "discount" usability engineering. *IEEE Software*, **7**(3), 107–108.

Nielsen, J. (1992). Finding usability problems through heuristic evaluation. In *Proceedings of ACM CHI '92*. Monterey, CA.

Nielsen, J. (1992a). Evaluating the thinking aloud technique for use by computer scientists. In H.R. Hartson and D. Hix (Eds.), *Advances in Human-Computer Interaction*, vol. 3. Ablex, Norwood, New Jersey, pp. 69–82.

Nielsen, J. (1993a). Usability evaluation and inspection methods. In S. Ashlund, K. Mullet, A. Henderson, E. Hollnagel, T. White (Eds.), *Bridges between Worlds, INTERCHI '93*. Tutorial notes 22. Reading, MA: Addison–Wesley.

Nielsen, J. (1993b). *Usability Engineering*. Academic Press, San Diego.

Nielsen, J. (1994a). Enhancing the explanatory power of usability heuristics. In *Proc. ACM CHI'94 Conf.* Boston, MA, pp. 152–158.

Nielsen, J. (1994b). Heuristic evaluation. In J. Nielsen and R.L. Mack (Eds.), *Usability Inspection Methods*. John Wiley & Sons, New York, NY.

Nielsen, J. (2000c). *Severity Ratings for Usability Problems. Severity Ratings in Heuristic Evaluation*. Retrieved 14 Feb. 2000: `http://www.useit.com`

Nielsen, J., and R.L. Mack (Eds.) (1994). *Usability Inspection Methods*. John Wiley & Sons.

Nielsen, J., and R. Molich (1990). Heuristic evaluation of user interfaces. In *Proc. of Computer-Human Interaction Conference* (CHI). Seattle, WA, pp. 249–256.

Nielsen, J., and D. Norman (2000). *Web-Site Usability: Get the Right Answers From Testing*. Retrieved 14 Feb. 2000: `http://www.useit.com`

Norman, D.A. (1988). *The Psychology of Everyday Things*. Basic Books, New York.

Norman, D. (2000). *Web-Site Usability: Walk-Through: a Usability Experiment*. Retrieved 9 Dec. 2003: `http://www.informationweek.com/773/we2.htm`

Smith, S., and J. Mosier (1986). *Design Guidelines for Designing User Interface Software*. The MITRE Corp. Retrieved 9 Dec. 2003: `http://www.hcibib.org/sam/` and `ftp://ftp.cis.ohio-state.edu/pub/hci/Guidelines`

Togniazzini, B. (2000). *Ask Tog: First Principles*. Retrieved 9 Dec. 2003: `http://www.asktog.com/basics/firstPrinciples.html`

Trochim, W. (1996). *Evaluating Websites*. Cornell University, Ithaca, NY. Retrieved 9 Dec. 2003: `http://trochim.human.cornell.edu/webeval/intro.htm`

Vygotsky, L. (1930/1978). *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge.

Wharton, C., J. Bradford, R. Jeffries and M. Franzke (1992). Applying cognitive walkthroughs to more complex user interfaces: experiences, issues and recommendations. In *Proceedings of ACM CHI '92*. Monterey, CA, pp. 381–388.

Wright, P.C., and A.F. Monk (1991). A cost-effective evaluation method for use by designers. *International Journal of Man-Machine Studies*, **35**(6), 891–912.

A. Karoulis has a BSc in mathematics from the Aristotle University of Thessaloniki, a degree in educational technologies from the University of Macedonia (Greece), a degree in open and distance learning from the Greek Open University, a MSc in information systems from the University of Macedonia (Greece), and a PhD in informatics, in the domain of human-computer interaction from the Aristotle University of Thessaloniki, Greece. He is currently active as an instructor in educational techologies for secondary education, as a multimedia and web project manager and as a researcher in the domains of HCI and of distance learning, regarding the application of new technologies at the Department of Informatics at Aristotle University. He is author of two books and co-author of another five which are published in Greece and he has managed more than six multimedia projects. His scientific interests concern human-computer interaction, multimedia and web design, educational technologies and distance learning.

A. Pombortsis received a BSc degree in physics and an MSc degree in electronics and communications (both from the University of Thessaloniki), and a diploma degree in electrical engineering from the Technical University of Thessaloniki. In 1987 he received a PhD degree in computer science from the University of Thessaloniki. Currently, he is professor in the Department of Informatics, Aristotle University Of Thessaloniki, Greece. His research interests include computer architecture, parallel and distributed computer systems, and multimedia systems.

Euristinis žiniatinklių vertinimas atsižvelgiant į vertintojo kompetenciją ir tam tikrą kriterijų sąrašą

Athanasis KAROULIS, Andreas POMBORTSIS

Nors žiniatinklių vertinimo tyrimai jau atliekami, tačiau tai vyksta dar pakankamai vangiai. Tai galima pateisinti keliomis pagrindinėmis priežastimis: išrastų metodologijų taikymu žiniatinklio specifikai, individualia žiniatinklių struktūra bei tinkamu vertinimų kriterijų nustatymo keblumu. Straipsnyje nagrinėjami, būtent, šie atvejai, taip pat pasiūlomas euristinis žiniatinklių vertinimo metodas. Atliekant tyrimą mėginta pirmiausia parengti vertintojus taip, kad jie įsisavintų euristinio vertinimo specifiką, tiek klasikiniame, tiek ir interneto kontekste. Tai buvo daroma siekiant atsakyti į esminį klausimą: ar prieš vertinimo procesą atlikti mokymai gali pagerinti įvertinimo kokybę. Vėliau buvo pasitelktas žiniatinkliui pritaikytas euristinis metodas, aprašytas atitinkamoje literatūroje, vėliau su juo buvo supažindinti vertintojai. Galiausiai vertintojams buvo pateikta užduotis įvertinti penkis žiniatinkliui bei aptarti savo pastebėjimus specialiai tam paruoštose anketose. Gauti rezultatai patvirtino du jau žinomus dalykus: kad taikomas metodas iš tiesų tinka žniatinklio specifikai ir kad vertintojų kompetencija yra ypatingai svarbi. Be to, prieita išvados, jog esant tinkamoms sąlygoms vertintojų kompetenciją galima pagilinti per pakankamai trumpą laiką. Pagrindine analizės išvada būtų ta, jog nors ir naudotas euristinis sąrašas nebuvo pakankamas, tačiau pastebėta tendencija, kad vertintojai linkę vadovautis panašia mastymo paradigma, o tai leidžia daryti prielaidą, jog toks euristinio metodo taikymas gali būti prasmingas nagrinėjamu atveju.