Introduction

Usability – the concept and methods – constitutes perhaps the single most important contribution that research in Human-Computer Interaction (HCI) have made to practice in the field. The first HCI study of usability was reported by Roberts and Moran in 1982. Since then, results of usability-related HCI research have been widely spread both within the research field and in practice, and today the concept and methods of usability are employed as an important instrument for quality assurance in the design of IT artifacts worldwide.

Traditionally, usability has generally been about understanding system functionality, i.e. the users' understanding of how to use a system. The concept of usability was operationalized through identifying and measuring specific aspects of human-computer interaction, such as learnability, memorability, efficiency and user satisfaction. The underlying idea can be formulated as follows: if designers cannot build systems that are sufficiently comprehensible to the intended users – then something is wrong.

Currently, usability research and practice are facing a serious challenge. The focus of design concerns is expanding from predominantly functional aspects of IT systems to overall user experience. This is partly due to a wider scope of current applications of IT, which now include not only systems with function as their most critical aspect, but also systems designed to target the overall user experience. The shift in design focus also partly originates from more demanding users, who are becoming increasingly mature in their use of IT systems, and therefore also expect the systems to provide them with experiences and not just functionality.

The idea of experience is playing an increasingly central role in society as a whole. For instance, based on goals for sustainable economic growth for Europe formulated by the European Union (EU), the Swedish Agency for Innovation Systems (VINNOVA) indicated eighteen key issues for growth in Sweden. In particular, development of the experience industry was identified as very important:

> *"For a couple of years now the experience industry has been a rapidly developing area, nationally as well as internationally. There is a great unexploited market potential; at the same time "experience" is the fastest growing industry in the Swedish labor market. Technological development in the area of IT is further strengthening its potential through the creation of new possibilities for developing attractive services and products".*
> http://www.vinnova.se/(2003-10-10)

The Swedish organization KK-stiftelsen (The Knowledge Foundation) has presented statistics plotting the progress of the Swedish experience industry between the years 1995-2001. The report based on these statistics points out that:

> *"The experience industry is advancing from strength to strength. Between the years 1995 and 2001, this industry grew in Sweden by as much as 45 %..[and] is now larger than many of the traditional Swedish industries, such as forestry and timber. The growth of the experience industry was 9 % higher [than the growth of other industries] and contributed 4,8% of the GNP in 2001 [in Sweden]."*
> http://www.kks.se (2003-10-14)

Therefore, the category of experience is no longer limited to domains of academic discourse or product promotion. As the above citations clearly indicate, understanding, support, and evaluation of experience is also becoming an economic necessity.

The trend towards experience has direct implications for usability evaluation. Since experience is considered an important aspect of the quality of various products, it should be evaluated. One possible, if not universally accepted, approach is to consider evaluation of experience as a case of usability evaluation. However, the existing methods cannot be employed. When the focus is on experiences rather than on more functional aspects of systems, a revision of usability methods is required. The traditional usability evaluation methods used in research and in large corporations, deal adequately with functional aspects, but to date are not particularly suitable for dealing with the experiential aspects of systems. Evaluating users' experiences and not just users' understanding of the system requires new procedures, indicators and analysis. In other words, the basic concepts and methods of usability need to be further developed before they can be applied to this new field.

This thesis is an attempt to deal with the problem of how systems which are intended to induce or facilitate certain experiences, can be evaluated in the context of usability. This is a very broad area including as it does a wide range of theoretical as well as methodological issues. There are a number of ways the problem can be approached. One possibility is to consider traditional usability as obsolete and discard it completely. If this is done, a new system of concepts, methods, and procedures would need to be developed "from scratch". Another possibility is to take existing usability methodology as the point of departure and extend it to cover user experience. The latter approach has been adopted in this thesis. In other words, the underlying intention was to investigate how far one could advance in evaluating "experiential" aspects of usability following the lines of traditional usability.

It should be noted that this thesis does not intend to solve the problem of evaluating user experience once and for all. It has a more modest ambition. The thesis represents only one possible approach, which can and should be combined with further research based on different presuppositions. In addition, the scope of the empirical study conducted within this thesis is limited to a subset of the general scope of usability. The empirical study reported below specifically focuses on *fun and entertainment*, deals with *web usability*, and, furthermore, with usability of a special type of website, namely, the so-called *entertainment websites*. Therefore, the thesis should be considered as taking a step towards "experiential" usability rather than creating a new methodology.

To accomplish its objective the research reported in this thesis progressed through a series of steps. First, existing studies in the area of usability evaluation were examined to determine the extent to which these studies can shed light on evaluation of fun and usability. It was concluded that evaluation of fun and usability remains an open issue. Then, the theories that appeared most relevant for fun and usability were analyzed to establish if they could help in *operationalizing* fun and entertainment as aspects of web usability. Since the input from theories was judged to be not sufficiently specific to guide a revision of usability evaluation methodology, the study defined fun and entertainment, in terms of relevance for usability evaluation, as properties intentionally implemented by designers. At a concrete methodological level the problem of thus identifying fun and entertainment was addressed through selecting entertainment web sites as the primary object of empirical investigation. The empirical investigation was designed as a three-phase study. In the first phase an initial set of traditional usability methods (both empirical ones and inspection ones) were applied in evaluation of a variety of websites. In the second phase the findings of the first phase were used to

refine and redesign the initial set of methods. These revised methods were applied in a new round of usability evaluations of actual websites. This research strategy allowed a number of specific conclusions to be drawn about usability evaluation of fun and entertainment.

Aim, research questions, and intended outcomes of the study

The aim of the study reported in this thesis was to explore the potential of traditional usability evaluation approaches to deal with issues related to user experience, such as fun and entertainment. To achieve this aim the study focused on the following research questions:

- Can traditional usability methodology be applied in principle to evaluation of fun and entertainment?
- What are the limitations of traditional usability methods in evaluating fun and entertainment?
- In what ways can the above limitations be overcome?
- How can existing theories of fun and entertainment help develop an experience-oriented usability evaluation methods?
- How can fun and entertainment be defined so that usability evaluations can be applied to them?
- What research strategy can be used to evaluate and refine usability evaluation methods?
- What are the factors that contribute to the effectiveness of employing usability evaluation methods in evaluating fun and entertainment?

The intended outcome of the study was twofold. On the one and, the study intended to produce theoretical analyses and empirical findings that attempt to provide answers to the above questions. These were the intended research outcome of the study. On the other hand, the study was expected to generate more practical contributions, namely, a refined methodology for empirical usability evaluation suitable for evaluating entertainment web sites, a re-designed list of heuristics for Heuristic Evaluation of entertainment web sites, and a new methodology for expert evaluation of entertainment web site after the collaboration.

Cooperation with software industry

This thesis partly describes a joint collaboration project between Umeå University and Paregos Media Design AB, a leading Swedish web design company. Paregos Mediadesign AB provided access to the web sites in their design projects and all related materials in these projects. Throughout the process the company also provided the tools and input needed to perform evaluations as part of the continuous quality assurance in its design process. The figures below originate from Paregos Mediadesign AB and describe the changes they made in the design process as a result of our collaboration. The first figure shows the production process before the collaboration and the second visualizes the changed view of the same process.

Figure I.1: The production process at Paregos Mediadesign before the collaboration (Paregos, 2001)

The figures show that before the collaboration usability issues were practically not considered important in the design process. System evaluation took place during the very last phase of the production process and dealt almost exclusively with code bugs and other technological aspects. As a result of collaboration with the research team, the company changed the process, so that usability became a concern of designers at every phase of the design process. As stated by the management of Paregos:

"Usability aspects have become a central 'measure of success' as a result of the collaboration with the research team", (Niklas Forslund, founder of Paregos, 2001-10-20).

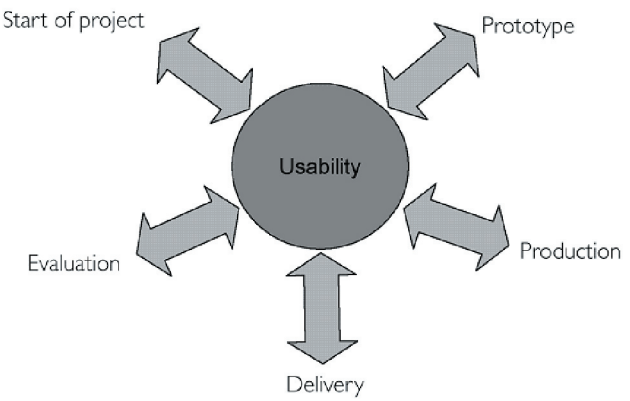

Figure I.2: The production process after the collaboration (Paregos, 2001)

Paregos came to understand the importance of user-centered design, where usability is a key issue, and during the collaboration process strategies were developed to involve usability at all stages in the design process. Usability was also used as an argument in promoting Paregos design.

Publications

The study reported in this thesis was presented at a number of international conferences and published in conference proceedings. Some of these publications were incorporated into the thesis, mainly in a revised form.These publications are:

Olsson, C. (2000). The usability concept re-considered: A need for new ways of measuring real web use. In proceedings of IRIS23 Informations systems Research seminar In Scandinavia, Doing IT Together. Eds. L. Svensson, U. Snis, C. Soerensen, H. Fägerlind, T. Lindroth, M. Magnusson, . Östlund. Laboratorium for Interaction Technology, University of Trollhättan Uddevalla.

Olsson, C. (2000). To Measure or not to Measure: Why web usability is different from traditional usability. In proceedings of WebNet 2000, Association for the Advancement of Computing in Education, Charlottesville, VA.

Jegers, K. & Wiberg, C. (2001). Evaluating Experience: Implications for usability tests conducted on entertainment web sites. In proceedings of IRIS24 Information systems Research seminar In Scandinavia

Wiberg, C. (2001). Join the Joyride: An Identification of Three Important Factors for Evaluation of On-line Entertainment. In proceedings of WebNet 2001, Association for the Advancement of Computing in Education, Charlottesville, VA.

Wiberg, C. (2001). From ease of use to fun of use: Usability evaluation guidelines for testing entertainment web sites. In Proceedings of Conference on Affective Human Factors Design, CAHD, Singapore

Wiberg, C. (2001). Bridging the Gap Between Designer and Ethnographer by Using a Facilitator. In Glimell, H. & Juhlin, O. The Social Production of Technology: On the everyday life with things.(eds.) (pp.222-241)

Danielsson, K. & Wiberg, C. (2002). IT Basketball – A Sporty Virtual Environment: An Evaluation of Usability, Presence and Interest in Service. In Proceedings of IRIS25 Informations Systems Research Seminar in Scandinavia.

Jegers, K. & Wiberg, C. (2003) Satisfaction and Learnability in Edutainment: A usability study of the knowledge game 'Laser Challenge' at the Nobel e-museum. In Proceedings of the International Conference of Human Computer Interaction, Crete, Greece, June, 22-25, 2003

Jegers, K. & Wiberg, C. (2003) FunTain: Design Implications for Edutainment Games. In Proceedings of ED-MEDIA 2003, AACE, Honolulu, Hawaii, June 22-25, 2003

Structure of the thesis

This thesis follows a structure typical of many academic dissertations. That is, it can be described as a double funnel or trumpet, with the openings pointing in opposite directions. The width of the funnel represents the general vs. the particular, i.e. the thesis starts at a general level, moves towards the particular and finally opens up to the general again (c.f. Holtom & Fisher, 1999). A visualization of the thesis, based on this type of thinking, with chapters in the center and the parts underneath, is shown in Figure I.3.A more detailed structure of the contents of each part of the thesis is as follows:

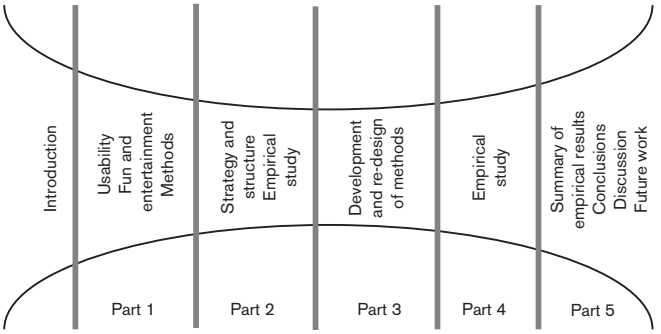

Figure I.3: The thesis visualized as a 'double funnel'. (Based on a similar model presented by Holtom & Fisher, 1999)

Part 1

This part provides an overview of main concepts and relevant research. The concept of usability and traditional usability methods and techniques are described and discussed in this section. Usability studies related to fun and entertainment, as well as theoretical frameworks and definitions of these phenomena are then discussed. Finally, methods as objects of study are further analyzed. The chapters included in Part 1 are:

- Chapter 1 – Usability
- Chapter 2 – Entertainment and fun as aspects of web usability
- Chapter 3 – Usability evaluation methods as objects of study

Part 2

In the second part, the overall research strategy and structure are described. The first phase of the empirical study, i.e. the application of traditional usability evaluation methods, is presented. The chapters included in Part 2 are:

- Chapter 4 – Strategy and structure
- Chapter 5 – Using traditional empirical usability evaluation methods
- Chapter 6 – Using traditional inspection methods –Experts
- Chapter 7 – Using traditional inspection methods – Novices

Part 3

The third part describes the process of revision and re-design of the methods under investigation. The chapters comprising Part 3 are:

- Chapter 8 – Revision and re-design of empirical usability evaluation methods
- Chapter 9 – Revision and re-design of inspection methods

Part 4

Part 4 includes the studies conducted using the revised methodological approaches. The chapters included are:

- Chapter 10 – Evaluations using revised inspection methods
- Chapter 11 – Testing of a new methodology for empirical usability evaluation methods
- Chapter 12 – Testing of new methodology for inspection methods

Part 5

The final part contains a discussion of the study specifically and evaluation of entertainment in general. The chapters included are:

- Chapter 13 – Summary of empirical findings
- Chapter 14 – Discussion

The empirical study presented in the thesis employed an extensive range of materials, such as web sites, questionnaires, evaluation forms, lists of heuristics etc. These materials are presented in Appendices I - III.

How to read the thesis

This thesis is intended for a wide audience. Thus different groups of readers will come to it with different intentions. A general academic reader might show interest in the thesis not because of interest in the specific topic of usability in relation to fun, but rather out of curiosity about how the topic is related to the 'discipline' of informatics. For such a reader, the introduction, strategy and structure, and finally the conclusions might be of most interest.

For academic readers with more specific knowledge and interest in informatics in general and HCI in particular, the introduction, strategy and structure, empirical study and conclusions might be of greatest interest and value, while they would probably already be familiar with the ideas described in the theoretical part.

The third and final group of potential readers that this thesis might reach is practitioners in the software industry producing entertainment web sites. From my experience of collaborating with this group of people, the strongest lasting impression is the time pressure they work under. Ironically, this audience is the only one recommended to read the whole thesis. As I am quite sure this will never happen, a more reasonable recommendation is to read the introduction, discussion of usability methods, discussion of theories of fun, and summary of results.

In a perfect world, of course, everyone would read the entire thesis and leave it with a richer understanding of the whole picture of the evaluation of entertainment web sites.

Part I

Methods and theories in usability and fun

Part 1 comprises the central theoretical and methodological concepts that form the basis of for the study conducted in this thesis. In general, three main areas are covered:

- Chapter 1 – Usability
- Chapter 2 – Entertainment and fun as aspects in web usability
- Chapter 3 – Usability evaluation methods as objects of study

Chapter 1 discusses the attributes of usability and presents general descriptions and definitions of the concept of usability. Further, an overview of various usability evaluation methods is presented together with a number of examples of related research work. Finally, the chapter provides an overview of research in areas of Human Factors and Human-Computer Interaction, related to fun and entertainment. Chapter 2 considers a variety of theories in relation to fun, entertainment, pleasure and experience. These frameworks are explored to see if they could provide support in operationalizing entertainment and fun in the study.

The concept of *entertainment web site* is defined and further explored in relation to evaluation methods. Chapter 3 develops the concept of methods. They are defined and discussed from a number of points of view, for instance they are divided into product- and process-oriented methods. Finally, possible frameworks within which judgments can be made are presented to provide an understanding of the ways that methods can be assessed with the object of redesigning and refining them to better suit our purposes.

Chapter 1
Usability

Usability is a key concept in HCI. It concerns, for instance, making systems safe, easy to learn and easy to use (Preece, 1994). It originates from 'software psychology' in the 1970s, which was a related discipline to experimental psychology (Shneiderman, 1980 in Ehn & Löwgren, 1997)

The first usability study reported in HCI research, was presented by Roberts and Moran (1982), and was an evaluation of text editors. Here, the first attempt to divide the concept of usability into various dimensions was made. Thus, usability was divided into:

- *Time* to perform edit tasks by experts
- *Errors* made by experts
- *Learning* of basic edit tasks by novices
- *Functionality* over all possible edit tasks

As will be shown below, this initial division has had a great influence on many later divisions of the concept, and has itself also been developed further.

The HCI community is mainly an inter-disciplinary research community, including people from a range of research disciplines such as Psychology, Ergonomics, Sociology, Anthropology, Computer Science, etc. (c.f. Monk and Gilbert, 1995). Researchers in Psychology, and more specifically Experimental Psychology, can draw on a long line of various kinds of experiments in human behavior. These experiments have also had a major impact on the evaluation of usability in the HCI community in that many of the methods used in HCI originate from Experimental Psychology. Then methods in the latter field have

focused mainly on first generating a hypothesis and then gathering quantitative data in order to test the hypothesis. The focus on data is vital, arising as it did from a collective reaction to the theories of "armchair" psychologists at the turn of the century (Monk and Gilbert, 1995).

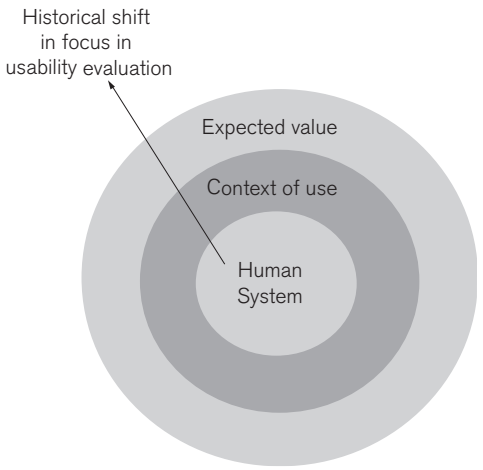

Figure 1.1 The shift in focus in usability research and practice over time. (Ottersten & Berndtsson, 2002)

Historically, the focus in usability research has moved from concern only with the human system, to concern with the context, and has finally become focused on expected value (Löwgren, 1993; Ottersten & Berndtsson, 2002). The model below shows the shift in focus:

This shift might also imply the re-designing of evaluation methods, which has occurred to some extent, but the results in this thesis show that more work needs to be done on this. Another shift in focus in evaluating software in HCI has been away from hypothesis testing and the gathering of extensive quantitative data material to a view of evaluation as an information search used to inform iterative design (Karat 1997). There are many sub-disciplines in HCI working with usability and usability testing of which some are User-Centered Design (Karat, 1997), Usability Engineering (Nielsen, 1993), Task Analysis (Diaper, 1989) and more.

The usability concept

The term usability employed in daily talk may suggest both something it is, and even something it is not. Karat (1997)defines usability as follows:

"The usability of a product is not an attribute of the product alone it is an attribute of interaction with a product in a context of use."

Figure 1.2 shows usability and its context, as described by Jakob Nielsen (1993). Note however that this is only one of many categorizations of usability.

In short, descriptions of some of the general concepts above are[1]:

- *System acceptability*. The ability of the system to meet all needs and requirements of all stakeholders, from direct users to customers etc.
- *Social acceptability*. The correspondence of the system to the social rules and norms that apply in a given context.
- *Practical acceptability*. The acceptability of the systems as regards cost, reliability, compatibility, etc.
- *Usefulness*. The ability of the system to achieve a desired goal. This can be

broken down into utility and usability.

- *Utility*. The ability of the system to do what is needed.
- *Usability*. The practical usability needed by the user of the system's functionality.

Further, Nielsen (1993) defines usability as containing at least the following aspects:

1 *Learnability*: The system should be easy to learn so that the user can rapidly start to get some work done using the system.

2 *Efficiency*: The system should be efficient, so that once the user has mastered it, a high level of productivity is possible.

3 *Memorability*: The system should be easy to remember, so that the occasional user is able to return to the system after not having used it for some time, without having to re-learn everything.

4 *Errors*: The system should have a low error rate, so that when using it users make few errors, and if such errors occur they can be easily rectified. Further, there must be no catastrophic errors.

5 *Satisfaction*: The system should be pleasant to use so that when they use it users are subjectively satisfied.

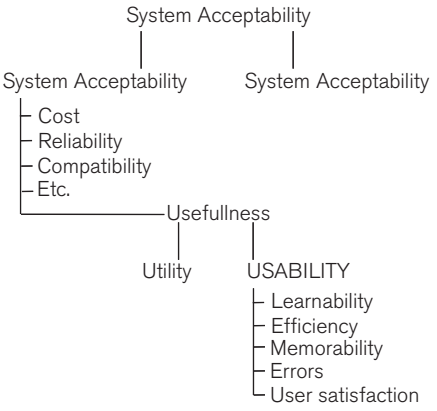

Figure 1.2 The context of usability in general (Nielsen, 1993, p.25).

Below other, and somewhat similar descriptions of usability are given:

Dix et. al. (1998) divide usability into categories according to principles which are then further described in sub-categories (not shown here):

1. Learnability – the ease with which new users can begin effective interaction with the system and maximize their performance.
2. Flexibility – the multiplicity of ways in which user and system can exchange information
3. Robustness – the level of support provided to the user for determining successful achievement and the assessment of goals.

It should be noted in the above that Dix et al does not employ any category referring to 'user satisfaction' or related notions.

The International Organization for Standardization (ISO) working group for human-system interaction has also concentrated on standards for usability. Their definition of usability is as follows:

> *"Usability of a product is the extent to which the product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use."* (ISO 9241-10:1996)

One final note, in relation to the general theme of this thesis, is that in all definitions of usability, the notion of 'user satisfaction' is either listed last, or not mentioned at all. This neglect of 'user satisfaction' poses a two-fold problem. (1) By listing 'user satisfaction' as one of the categories of usability, the HCI community safeguards itself against the criticism of neglecting its importance. However, the discussions concerning the techniques used to evaluate this category are in general brief, and comments such as 'hard to grasp', 'subjective data and therefore difficult to analyze' can often be found (c.f. Nielsen, 1993). (2) Further, the number of studies conducted concerning user satisfaction in HCI over the years is low, if compared, for instance, to those dealing with efficiency and numbers of errors. It is as if merely by including 'user satisfaction' as one category, the problem has been covered. It is seen as one possible category of usability aspects to be investigated. However, it seems that having said this nothing further is required.

This thesis argues for further investigation and development of the 'user satisfaction' category.

Evaluating usability

When designing evaluation experiments, it is important to first consider some more general aspects and to distinguish between: (1) *process*, (2) *purpose* and (3) *object* (Karat, 1997).

The process includes the choice of evaluation method and basic assumptions in data handling, e.g. are objective or subjective data required in the specific context. If subjective data are required in the context, how then are comments such as 'cool', 'nice' or 'attractive' to be interpreted? Further, the process carried out, could be more or less conscious. Unintentionally, the process could have been designed to be something completely different from what was intended. When thinking about the process, the aspect of who is conducting the evaluations should also be taken into account. Individual differences between evaluators cannot be avoided. Self-reflection in the process is vital. These are some examples of issues to take into account in relation to the process (c.f. Karat, 1997).

The purpose of the evaluation is crucial. What should be the outcome of the evaluation? Are navigation problems the biggest issue? Is the question how to maximize customer satisfaction? Are the designers asking for opinions about the graphical 'look and feel' of the system or is the main purpose to maximize economic outcome for target organization? Are we working in a project which is iterative, meaning that our findings are often reported as design flaws or problems, which have to be translated into design suggestions? All the different purposes and goals of the evaluation match with different types of actions when it comes to both the evaluation process and the object of study (c.f. Karat, 1997).

Finally it is also important to identify the object of study in the evaluation. Even if such an identification seems easy, for example a web site, a game or an application it might be considerably more complicated. Which part of the web site should be used in the evaluations? Should the use of structured tasks be considered in order to narrow the subject's choices? Moreover, are results from only one section in a game sufficient to provide guidelines for the whole game, i.e. could the data be considered generalizable? In addition, the context of the object of study must be recognized. Is the object of study evaluated to be taken out of context, or should the context also be taken into account? If the latter situation is the case, how can anything valid be said about the data? Which results refer to the object and which to the chosen context of the evaluation? (c.f. Karat, 1997)

Methods in usability evaluation

Usability engineering is a field where practitioners and researchers have been working with usability testing for more than 20 years. Within the fields of research and practice, a number of evaluation methods and techniques have been developed, such as Think-aloud protocol, Heuristic Evaluation and Task Analysis (c.f. Virzi, 1997; Nielsen, 1993). In general, a usability evaluation methods can be defined as:

> *"A Usability Evaluation Method is a process for producing a measurement of usability"* (Karat, 1997)

Usability evaluation methods are divided into two types, empirical usability methods and inspection methods (Nielsen, 1993; Virzi, 1997; Dix et. al. 1998; Karat, 1997). The two types will be described and discussed in more detail below.

Empirical usability evaluation

Empirical usability evaluations may be conducted in numerous ways, including everything from using one single technique to employing a whole repertoire of approaches. In general, all empirical usability evaluations involve users of some kind. The evaluations can be conducted in a variety of contexts but usually so-called usability labs are used. In the absence of a usability lab, other settings can be used. For instance, an office or conference room would be suitable. In other situations, if the tested application so requires, contextual inquiries are conducted. Here, the system or application is evaluated in a 'real-world' setting, usually with target users, i.e. the users who are working or living in the context.

It is important to be aware of what is to be measured. One common way of dividing up the measuring of usability is the following (Redmond-Pyle and Moore 1995):

1. *Performance tests* where users use the system to perform a task, and their effectiveness is measured. Common measures are speed, accuracy and/or errors.
2. *Attitude surveys* where user satisfaction and user perception of the software are evaluated. Common ways of securing data are questionnaires or interviews.

Table 1.1 presents an overview of examples of empirical usability methods.

Method	Measures	Generated data
Think-aloud (or verbal) protocol	Captured events from usage situations; problems, expectations etc.	Record of cognitive processes of users in system usage
Use data collection	Number of errors, types of errors, time to complete task	Record of statistics for errors, listings of types of occurring errors, time statistics
Clinical experiments	Eye gaze, heart rate, skin color, body heat	Statistics for measured clinical aspects
Surveys and Questionnaires	Accuracy regarding memory, learning etc.	Record of answers – quantitative or qualitative
Interviews	General information from users. Structured or unstructured.	Record of answers - qualitative

Table 1.1 Usability evaluation methods

Think-aloud (or verbal) protocol

In some disciplines, for instance experimental psychology, the impact of what people actually say in a test situation, might not be given much weight. The 'hard' data, e.g. number of errors or clinical data, such as heart rate, are often considered to be more valid or reliable. A well-established distrust of what people say exists in this type of discipline. The data from verbal protocols is also somewhat diffuse in its form, and cannot easily be managed and categorized. (Karat, 1997) However, useful results can be gained from such approaches if some methodological aspects are considered[2], for instance time of response, i.e. when is it fruitful to ask subjects for answers, the level of intervention from the evaluator and the process of finding suitable tasks or assignments for the subjects to carry out.

Another, related approach, is called cooperative evaluation. This could be seen as a method that lies somewhere between empirical usability evaluation and inspection methods. It relies on non-expert evaluators working together with a single user who thinks aloud. The level of intervention here might be considered to be very high. The data is definitely dependent on the evaluator, but that is the intention. The advantage here is that the think-aloud process is highly facilitated, i.e. it is more of a conversation between two collaborating peers (Virzi, 1997).

Use data collection

This group of methods and techniques all gather objective data about use of the system, such as, number and type of errors, time to complete task, requests for help and general logging data. There is a wide range of methods as well as types of data to collect. The advantage with these types of evaluations is that the data is easier to handle and structure. In order to give a bigger picture, it is common to combine such data with data from think-aloud protocols, questionnaires and interviews (Karat, 1997).

Clinical experiments

This type of empirical usability evaluation can also be considered as part of usage data collection. However, it is excluded because the types of data collected with these techniques are quite distinct from the types of data collected above. Clinical experiments originate from the discipline of experimental psychology and examples of data collected are eye-gaze, heart rate, skin color, body heat. The data are analyzed according to specific assumptions that they correspond to specific states of the subjects. For instance, a change in heart rate might be considered to be an indication of happiness, a change in skin color might communicate that the subject is aroused in some way, etc. The basis for this type of science might be questioned, but it is covered here as does exist and such experiments are conducted. For instance, eye gaze measuring is becoming increasingly popular in related literature. Overall, these types of methods could be used for the evaluation of systems and applications where the experiences and moods of the subjects are of interest.

Surveys and questionnaires

The questions in surveys and questionnaires could either be general or very specific. Users can be asked to give long and detailed answers or the structure of the survey or questionnaire could be very strict, where users check boxes for right alternatives to questions such as agreeing or disagreeing with statements or ratings on scales (Karat, 1997). An overview of possible styles of questions used in questionnaires might be as follows (Dix et. al, 1998):

- **General** – Used for information about background of user, such as age, gender and other types of general background information.
- **Open-ended** – Used to gather general subjective information
- **Scalar** – The respondent makes judgments about a specific statement on a numeric scale
- **Multi-choice** – The respondent is offered a set of choices of explicit responses to choose from
- **Ranking** – Used to order items, useful for indicating user preferences

The possibilities are many, but this does not mean it is easy to design surveys and questionnaires. As in all research where this type of data collecting is used, it is extremely important to conduct meta-tests of the material. This could be done by allowing colleagues to read and make comments on the questions, or by testing the questionnaire or survey on a small number of users, for further revision. (Dix et. al, 1998).

Interviews

For a novice, forming questions and conducting interviews often seem to be straightforward and simple. However, after conducting some interviews, the novice will probably realize that the perceived simplicity has produced frustration. There are many 'schools' of thought concerning the formatting of questions and the interpreting of data obtained. Postmodernism, Hermeneutics and Phenomenology are examples of these schools. Each wants to discover how knowledge should be viewed, and thus the process for each will be different (Kvale, 1997).

Interviews frequently use open-ended questions and the three different interview approaches influence the formatting of the questions. The three approaches are: (Patton, 2002):

- Informal conversation interview
- General interview guide approach
- Standardized open-ended interview

The main difference is the extent to which the questions are pre-determined and standardized before the interview.

The *informal conversation interview* is the freest approach relying mostly on conversation. This approach is frequently used when conducting field experiments. The interviewee might not even be aware that she/he is participating in an interview. The resulting range of data is extensive and varies with each interview, which might make interpretation difficult. However, one advantage is that such data may be rich as unexpected aspects emerge.

Second, the *general interview guide approach* uses an interview guide, which lists possible questions and/or topics, that the interviewer is free to explore. One advantage of this approach is that the guide ensures that the interview covers the chosen topics. This, however, might be counterbalanced by possible bias resulting from the subject's awareness of being in an interview situation.

Third, the basic idea behind the *standardized open-ended interview* is conformity and structure. The questions are carefully worked out and worded before the interview. It is important that each subject receives the same questions, or *stimuli*,

so as to produce the same conditions for subjects. This will give the data obtained, higher quality. One advantage of this method is that the interviews are transparent and open to external inspection. In addition, variation among different interviewers can be minimized and the time can also be used efficiently. Finally, analysis is easier when the data material is very homogeneous regarding both structure and stimuli. One of the disadvantages is that the interview situation is very different from a traditional conversation, which might affect the interviewees, resulting in answers that lack quality.

Figure 1.3 A view from control room in a typical setting of a stationary usability lab at the Rabobank (the Netherlands). (http://www.noldus.com)

Usability labs – an overview

There is a lot of equipment available on the market to make the work of conducting empirical usability evaluation easier, more structured and more efficient. The use of a proper usability lab, with equipment for data recording, timing, and analysis is a widely-used model in such work. Usability labs vary in appearance, depending on resources. Corporations, such as Ericsson, Nokia and Telia Sonera (Swedish telecom) are examples of companies with very high-quality usability labs. This type of lab can also be found in research settings.

In discussions of usability labs, it is often the specific type of physical setting that is of interest. Usually, one room is used to accommodate the test person and a second is used as control room for the evaluation staff. Sometimes there is a third room which serves as an observation room for other types of audiences, such as buyers of the application tested or students interested in the test or experiment being conducted. The rooms should preferably be physically connected but may also be separated by walls with two-way mirrors, in order to minimize any disturbance for the test person during the experiment. This type of setting requires a lot of space, as it is designed only for conducting experiments. The equipment, such as computers, cameras, scan converters, video mixers, two-way mirrors, microphones and wiring, is often stationary with little or no possibility of making it moveable.

Figure 1.4 Example of usability lab equipment showing a so-called 'semi-portable lab' setting. (http://www.noldus.com)

Usability labs can also be mobile, where the equipment used, cameras, DV-recorders, video mixers etc., is portable. These labs are used to test mobile artefacts or context sensitive applications.

Figure 1.5 A portable, or mobile, usability lab setting. (http://www.noldus.com)

In addition to the physical setting, another important aspect of usability labs is the manner in which they handle data. Some of the equipment, e.g. scan converters in combination with a video mixer, receive digital signals and transform them into analogue signals for further investigation and interpretation. A further type of lab handles data as a continuous string of digital signals, working with software-based recording and storing. The lab is then incorporated into computers instead of comprising specific hardware for each feature. This can be conceptualized as in Figure 1.6.

Once the data, in the form of video and audio signals, transcriptions and other types of added comments are recorded, the analysis phase can start. Various types of technology could also be brought in to support the analysis. One such example would be *The Observer*™, an application for analyzing different types of data material. The audio- and video materials are usually transformed into computer-based video files, sometimes also connected with screen capture from the user's computer. In *The Observer*™, these files can then be marked and combined with a text file with comments. The marked sections of the video file can then be used to select video clips to create a highlight video or to sort events into categories, such as occurring errors of various kinds. The original data is still stored in its original shape for further investigation. Such applications can be considered extremely valuable in the analysis phase, if used correctly.

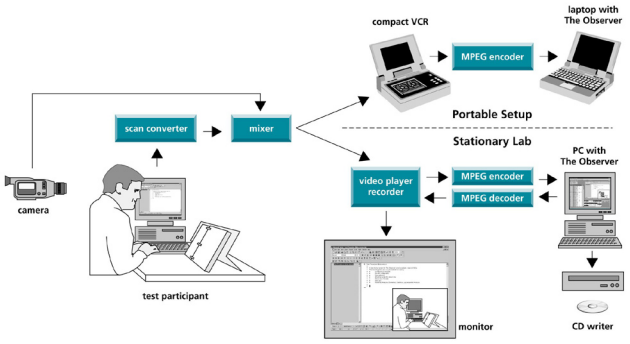

Figure 1.6 A conceptual schema showing possible set-ups for a portable and a stationary lab. (http://www.noldus.com)

Inspection methods

The term "Usability Inspection" was coined at a workshop at the CHI conference in 1992 (Mack & Nielsen, 1993), approximately two years after research papers on the topic first were published (Virzi 1997). Two of the earliest methods used were Heuristic Evaluation (Nielsen and Molich 1990) and Cognitive Walkthrough (Lewis and Polson, 1990) - two competing methods at that time in the HCI research community.

Usability evaluation methods designated inspection methods, all fulfill one of the following criteria: (1) they demand few resources in relation to the results gained, and (2) they identify potential usability problems. A third characteristic can also be added in that most of them minimize end-user involvement (Virzi, 1997).

There are, as so often in academic fields, many ways in which methods and techniques can be categorized and differentiated differently in different sources. However, regardless of how the methods are divided, the methods in themselves remain constant. On a higher level, all traditional inspection methods (IM) can be described as unique combinations of three dimensions (Virzi, 1997) namely:

1. Characteristics of the evaluators
2. Number of evaluators in a single session
3. The goals of the inspection

A list of commonly used inspection methods is presented in table 1.2. It should be noted that some approaches are designated differently in related literature. As far as possible, the division of methods below uses names from primary sources, i.e. original names. However, some methods have been developed further and others have, to some extent, merged. For instance, a change in the number of evaluators in an approach might in some literature be regarded as a completely new approach, while others consider the change to be only a change in the application of the same approach. For this reason, the picture of existing general inspection methods is somewhat diffuse.

The division of the methods below is taken from a tutorial given by Nielsen (1994), entitled 'Usability Inspection methods'. The descriptions of the methods are based on papers presented at the workshop on Inspection Methods, held at CHI'92 and also on the book published by Nielsen and Mack (1994), *Usability Inspection Methods*. Inspection methods found elsewhere, and not covered by Nielsen (1994a) have been added to the list[3]. These categories are then, in the more detailed descriptions of the methods, related to other sources, in order to provide as detailed an overview as possible.

Method	Characteristics of evaluators	Number of evaluators	Goals of inspection
Heuristic evaluation	Usability experts	One	Judge whether each element in interface follows heuristics
Cognitive Walkthrough	Cognitive Psychologists	One	Predictions of user behavior regarding learning
Formal usability inspections		One or group	Combines Heuristic Ev. with Cog. Walkthrough
Design (or pluralistic) Walkthroughs	Users, developers and Human Factor HCI experts	One or group	Walkthrough of each dialogue element by using scenarios
Feature inspection			Inspection of sequences of (complicated) features
Consistency inspection	(External) designers	Group	Comparison of different designs to check consistency
Standard inspections	Expert in (specific) standard(s)	One	Inspection of interface for compliance with standard(s)
Theory-based reviews	Experts in each method	One	Discover problems on a micro-level.

Table 1.2 Overview of inspection methods

Related inspection methods are also briefly described but for more descriptions of the methods as well as application examples the reader is recommended to consult related literature. The inspection methods applied in the evaluations in this thesis are discussed in more detail to provide a guide for further reading.

Heuristic Evaluation

The heuristics in Heuristic Evaluation were originally designed by Nielsen and Molich (1990) and the process of Heuristic Evaluation can be described as follows: A specific list of design guidelines, frequently called 'heuristics' is used as a basis for evaluation of a system or application. The evaluator, often a UI expert or designer, reviews the system and comments on usability problems in relation to each heuristic. The original list of heuristics is as follows:

1. Simple and natural dialog
   Simple means no irrelevant or rarely used information. Natural means an order that matches the task.
2. Speak the user's language
   Use concepts from the user's world
   Don't use system-specific engineering terms.
3. Minimize user memory load
   Don't make the user remember things from one action to the next. Leave information on the screen until it is no longer needed.
4. Be consistent
   Action sequences learned in one part of the system should apply in other parts.
5. Provide feedback
   Let users know what effect their actions have on the system.
6. Provide clearly marked exits.
   If users get into a part of the system that doesn't interest them, they should be able to get out quickly without damaging anything.
7. Provide short cuts
   Help experienced users avoid lengthy dialogs and informational messages they don't need.
8. Good error messages
   Let the user know what the problem is and how to correct it.
9. Prevent errors
   Whenever you discover an error message; ask if that error could have been prevented.

The original list of heuristics has later been revised, in order to make them more comprehensible. The new, and most commonly used list of heuristics in Heuristic Evaluation nowadays is as follows (Nielsen, 1993):

1. Visability of system status
   The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.
2. Match between system and 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. Allow users to tailor frequent actions.
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.

Cognitive Walkthrough

Cognitive Walkthrough is based on a theory of learning by exploring (Lewis & Polson , 1990), and for this reason the method focuses on 'ease of learning' (Virzi, 1997). It is based on the idea that users learn how to use a system or interface by exploring its functionality and that users learn new aspects of a system only when new tasks are conducted. The method in itself provides a formal framework for inspection. It is shown in Table 1.3.

Outline of a cognitive walkthrough	
Preparation	• Define assumed user background • Choose sample task • Specify correct action sequence(s) for task • Determine interfaced states along the sequence(s)
Analysis	• For each correct action: construct a success story that explains why a user would choose that action or use a failure story to indicate why a user would not choose that action • Record problems, reasons, and assumptions • Consider and record alternatives • Modify the interface design to eliminate problems
The table occurs in Lewis & Warton (1997) and is here only graphically revised.	

Table 1.3 Outline of cognitive walkthrough.

This method can be applied to different types of prototypes, from those that are paper-based to fully functional prototypes. Since its inception in 1990 it has been modified numerous times (Lewis & Warthon, 1997) but the basics shown are still valid.

Formal usability inspections

This method comprises a review of potential task performance when using a product. It was intended to be used by the designers or engineers of the system evaluated. One of the main purposes of the method is to help evaluators (designers) to find and organize large numbers of usability problems in big systems. The method employs a six-step procedure and broadly originates from prior usability and engineering inspection. (Kahn & Prail, 1994) In the context of inspection methods, it has similarities to Heuristic Evaluation in combination with Cognitive Walkthrough (Nielsen, 1994). The method has three characteristic features: (1) A defect detection (with six logistical steps) and description process, (2) an inspection team, and (3) a logical structure in the usability evaluation lifecycle.

Design (or pluralistic) walkthroughs

Generally, these types of evaluations are conducted in teams, even though some examples have only single evaluators (Karat, 1997). Usually, a team of designers and HCI experts 'walk through' the interface on the basis of some kind of more or less structured scenario (Nielsen, 1994b). Sometimes the approach is known as Pluralistic Walkthrough, and sometimes the term Design Walkthrough is used. Both of these can be regarded as almost the same, since the walkthrough method *in itself* is seen as one of the freest evaluation approaches of all inspection methods. It originates from traditional usability walkthroughs, which were carried out in the HCI research community long before the concept of inspection methods was even coined (Bias, 1994; Karat, 1997).

Design walkthrough is considered to be a fairly common-sense approach to usability evaluation (Karat, 1997). The approach is recommended for evaluations where no other inspection method seems suitable, i.e. when it is not known what is being sought or when the evaluated material is of a kind which is hard to encapsulate in heuristics, models or theoretical frameworks.

In discussions of the approach in the literature, *Pluralistic Walkthrough* has five defining characteristics: (1) Members of the evaluation team may include representative users, product developers and HCI professionals. (2) The prototype used in the walkthrough should be presented in the manner and the order in which it would appear if the system were up and running. (3) All participants should consider themselves to be end-users. (4) Every member should note down the action they would take in each situation, before any discussion begins. (5) During discussion, the representative users speak first followed by the rest (Bias, 1994).

Even if the above description seems to be very structured, it should once again be pointed out that the Design Walkthrough approach is very loose regarding its application. Even if it is said to be common to have a group of evaluators, single sessions where more or less structured scenarios guide the usage also occur. The general idea of Design Walkthrough is that it is intended as an approach where a thorough walkthrough of the system would capture many types of usage situations and not only mainstream actions (Karat, 1997).

Feature inspections

Feature Inspection originates from methods devised to design program languages, Programming Walkthrough (Bell, 1992 in Nielsen & Mack, 1994). It differs from other inspection methods in that it not only considers the usability of an interface or system, but also its functionality, i.e. utility. In general, other techniques do not cover utility, but mostly take it for granted. This type of programming walkthrough

is similar to Cognitive Walkthrough in that it is dependent on specific tasks. When using these methods, the evaluator needs (a set of) specified problems or tasks (Wharton et. al., 1994). In conducting evaluations with Feature Inspection, sequences of features used to accomplish typical tasks are listed, checks for long sequences are made and overall complicated tasks that the user would not carry out naturally are analyzed. Feature inspections are often conducted by members of the design team. (Nielsen, 1994a ; Nielsen & Mack, 1994)

Consistency inspections

This method is concerned with the demand for consistency of various kinds. The focus here is not necessarily on graphical consistency, which is addressed in other methods such as Heuristic Evaluation, but rather on technical consistency. The demand for the intergration of compilers, editors, debuggers, project management and end-user products necessitates component consistency. (Wixon et. al., 1994). Consistency inspections are often handled by designers, representing different parts of a larger project (Nielsen, 1994).

Standard inspections

There is a large and growing number of different types of standards in hardware and software. In many design situations it is important to ensure resemblance to or consistency with a specific standard. In this context, Standard inspections can be considered. An expert in a chosen standard inspects the interface or system for compliance. (Wixon et. al., 1994; Nielsen, 1994)

Choice of methods in relation to usability aspects

As mentioned, it is important to identify the aspect of usability that is to be evaluated. Different aspects of usability require different types of methods, since each has its advantages and disadvantages. In Figure 1.7 a model summarizing the choices of methods is shown. The aspects of usability are located in the center circle of the model[4]. The next circle outside the center illustrates a variety of examples of possible measures for each category of usability. The third circle illustrates empirical usability methods. Finally, in the outer circle expert methods for evaluating each category can be found. The model is descriptive in that it visualizes how things are or have been in the past in the area of HCI.

Another comment, in relation to the above model, is that inspection methods are not in general connected to a specific category of usability. Rather they are discussed with regard to how they relate to the process of design, i.e. where they

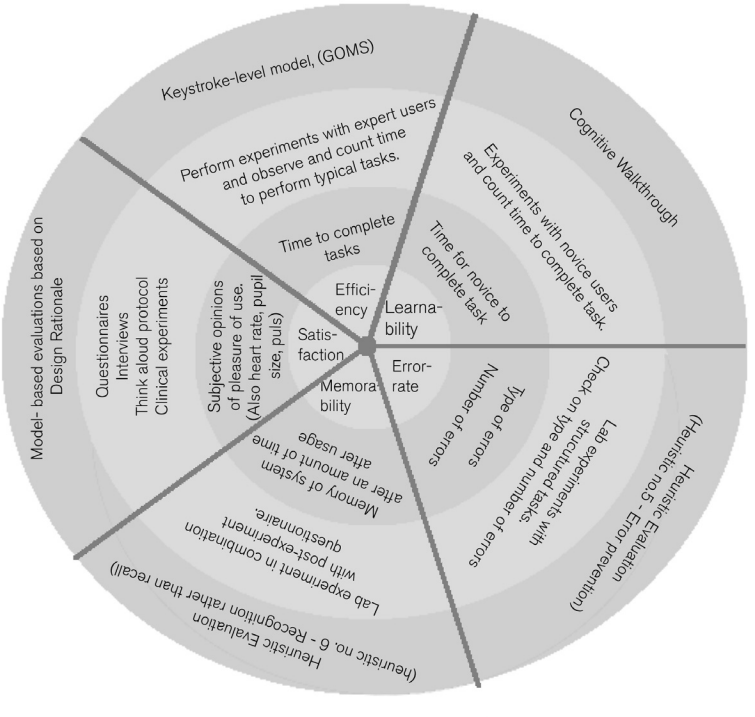

Figure 1.7 A model of how aspects of usability, possible measures and different methods relate in traditional HCI research and literature.[5]

are to be most suitably used in the process. The reason for this might be that their scope generally is wide; they cover a wide spectrum of aspects. Heuristic Evaluation, for example, includes heuristics that cover practically all aspects. Furthermore, different types of walkthroughs, such as Design Walkthroughs (c.f. Karat, 1997) rely on a very free approach, with no or only minor restrictions on the expert evaluator. It is difficult here, to connect a specific aspect of usability to a method in general.

Nevertheless, this is not the same thing as saying either that the mainstream inspection methods do not cover all the aspects of usability or that they could not be used for a specific aspect. If, for instance, the purpose of an expert evaluation is to discover efficiency or error prevention, the evaluator could be asked to check the artifact evaluated with regard to these particular aspects (Newman and Lamming 1995).

Efficiency. One common measure concerning this aspect is 'time to complete task'. The target users here are usually experts in the system (Nielsen, 1993). One empirical usability evaluation method recommended, is to carry out evaluations and experiments with expert users, where general observations as well as time-

counting in relation to performance of typical tasks are included (Nielsen, 1993). One example of an inspection method suitable for measuring this aspect of usability is Keystroke-level model (Dix et. al., 1997, p. 415). Another example might be GOMS. However, some argue that GOMS cannot be seen as a proper usability evaluation method (c.f. Virzi, 1997, p.708).

Learnability. This aspect is connected with measures regarding use of the systems by novices, for instance, 'time to complete task for novice users' (Nielsen, 1993). A natural choice of an empirical usability evaluation method would be experiments with novice users, where time to complete task is taken for selected tasks (Nielsen, 1993). Cognitive Walkthrough might be a good choice to measure this aspect. *"A cognitive walkthrough is a very specific type of usability inspection method that evaluates how easy an interface is to learn"* (Virzi, 1997, p.707)

Error rate. The usability aspect of error rate is usually measured in 'number of errors' or 'type of errors' and the empirical usability evaluation methods used are traditionally controlled experiments including structured tasks where the above measures are used (Nielsen, 1993). Here, it is worth mentioning that this aspect is something of a 'trade-off' in relation to the measuring of efficiency, i.e. if we put the subjects under pressure regarding time, the number of errors is likely to increase and vice versa. This trade-off has to be considered carefully when designing the experiments, i.e. the question is what are we actually measuring. It is not easy to find inspection methods that completely cover this aspect. However, one of the heuristics in Heuristic Evaluation includes this aspect, i.e. no. 5 'Error prevention'. It does not actually count or judge errors, but instead, concentrates on *prevention* of errors. (Nielsen, 1993; Newman & Lamming, 1995, p. 183)

Memorability The choice of a suitable inspection method for this aspect is not obvious. In related research, no recommendations are made regarding a specific inspection method for this aspect. However, the inspection method Heuristic Evaluation takes a related aspect into account – 'Recognition rather than recall' – even if this heuristic recommends, to some extent, the opposite, i.e. that users should not necessarily be forced to recall (memorize) but instead the interface should be based on recognition. Even if these aspects might be considered to be opposites, the Heuristic Evaluation method, could be regarded as more or less covering the aspect of memorability, as it at least lifts up the distinction between recognition and recall that is related to the aspect. (Nielsen, 1993; Newman & Lamming, 1995, p. 183)

Satisfaction A number of possible measures, which could be divided into qualitative and quantitative, could be applied to this aspect. The quantitative measures commonly used are heart rate, pupil size and body temperature. These are measured in clinical experiments, where the subject uses the system. Qualitative measures usually refer to various kinds of subjective opinions – often obtained in methods such as questionnaires, interviews and evaluations of systems including Think-aloud protocol (Nielsen, 1993). Inspection methods mentioned in the literature connected to user satisfaction are various model-based evaluations based on, for instance, Design Rationale (Dix et al. 1998).

Methodological considerations

As shown above, there are many methods and techniques available for evaluating the usability of a system. Each method requires both theoretical and practical skills, to ensure that the method chosen is the most suitable for the task. However, it is also very important to understand that a collection of methods is only a box of tools. As in the case of a carpenter, it is important not only to own the tools and have the skills to handle each tool properly. It is also essential to have the practical skill to know which tool to use in each specific situation to achieve the best result. This skill, in relation to usability, i.e. which method to choose in a certain situation, can to some extent be taught, but generally it comes with experience in working with usability evaluation. There are no hard and fast rules – each method has both strengths and weaknesses (Dix et al., 1998)

Some factors however can be generalized including: (Dix et. al., 1998)

- Stage in cycle at which to evaluate usability
- Style of evaluation
- Subjective vs. objective techniques
- Type of measures provided
- Information provided
- Time of response from subjects
- Level of intervention
- Required resources

These factors are discussed below in more detail.

Stage in cycle at which to evaluate usability

The purpose of the evaluation determines the time chosen for it to be conducted in the design process. If the design project uses an iterative process, where output from the evaluation will be used in further stages, an early evaluation is obviously

crucial. The evaluation is most effective if the system is tested at a prototype level. If this is the case, expert evaluations using a variety of inspection methods are the most suitable techniques. The reason for this is that it is difficult for target users to get an feeling for how the finished system will be, based only on paper prototypes. This will influence the quality of results. Experts, however, have the skills needed to imagine future systems based on mock-ups and prototypes and can report future problems without actually having to see them. Subjects in an empirical evaluation of prototypes will probably focus too much on problems related to the prototype itself rather than the future, running, system (c.f. Dix et al, 1998).

Even if an early evaluation is important it should not be seen as the only way to conduct usability tests. The value of end-user input should not be forgotten. Experts cannot predict everything and do not have full insight, for instance, into the context-related use of proper definitions in the system. In addition, the accomplishment of specific, commonly performed tasks in the system might also be difficult for experts who focus on general interaction models and guidelines in evaluations to understand. In this instance, subjects should be carefully chosen and tested regarding tasks and situations that are as authentic as possible in relation to future use of the system (c.f. Dix et. al., 1998).

Evaluation style

One big question is whether usability laboratories should be used or whether field studies should be conducted. In the laboratory controlled experiments can be designed which is useful for structured usability tests, for instance, where subjects are given designed tasks. Furthermore, different laboratory experiments have the advantage of being repeatable with a number of subjects under conditions that are as similar as possible. In a 'real' situation, such as in a field study, conditions are more difficult to control. (c.f. Dix et. al., 1998)

Yet, no matter how well-designed laboratory experiments are, they will never simulate future 'real' use of the system. Things happen, telephones ring, the system exists in parallel with other systems, which users choose among. If the designed system, a web site for instance, does not work as expected, a user is free to leave, and pick instead a competing system, site or service. It might be the case that 'real' use of a system is a *mix* of different systems, a fact that is easy to overlook in a laboratory situation. In these cases, and also others, it might be more suitable to use a field study type of evaluation. (c.f. Dix et. al., 1998)

Subjective vs. objective techniques

Evaluation techniques differ according to their objectivity. Some methods rely heavily on the interpretation of the evaluator. For instance, the Think Aloud technique relies on the expertise and experience of the evaluator who has to recognize problems and understand what the user is doing in the session. In such inspection methods as Cognitive Walkthrough or Heuristic Evaluation, the subjective interpretation of the experts' understanding of the future use of the system is also important in evaluating the system. Objective techniques, often in the form of controlled experiments of a quantitative kind, conducted in laboratories, are repeatable. These methods avoid biases and provide general and comparable results. However, they are deficient in that they might not reveal unexpected problems, resulting from user experiences. In an ideal situation, both subjective and objective approaches should be used when designing the evaluation of a system. (c.f. Dix et. al., 1998)

Type of measures provided

The measure of success differs, depending of the system evaluated. Sometimes the measures are quantifiable and sometimes they rely on a qualitative basis only. Thus the purpose of the evaluations should be defined. (c.f. Dix et. al., 1998) For instance, it should be known whether the system is to be used as effectively as possible by expert users, or mainly by novices or if the designed system is for leisure purposes. The system and the evaluations should vary according to the purpose for which it is to be used. Sometimes it might be fruitful to combine quantitative and qualitative measures, as for example in situations where it is important that the system is both effective and fun to use. In some cases, the measures are only quantitative, such as in the case of a booking system for travel agents. In other cases, for instance games, it might be argued that quantitative measures are of no importance and it is qualitative data that are required of a successful system[6].

Information provided

The type of information obtained from an evaluation might also vary. In the case where low-level information is important, for instance, the choice of font, the definitions used in navigation bars or the color used for different labels or buttons are significant. In other cases, higher-level information might be more appropriate. This would be true of more general discussions concerning such issues as whether the system is usable or the design pleasing or the game fun to play. The former type of low-level information can be gathered for instance by means of controlled and well-structured experiments, the higher level through questionnaires, interviews or by using the Think-aloud technique. (c.f. Dix et. al., 1998)

Time of response from subjects

One issue to consider in usability evaluation is when the subjects should be asked for the required information. Some methods or techniques for gathering information about usability record immediate user behavior at the time of the interaction itself, as does the Think-aloud protocol. The subject is asked to talk immediately about the problems as soon as they occur. Other techniques, such as post-task questionnaires or post-task interviews, allow the subjects to recall events (c.f Dix et. al., 1998). Some argue that the latter are not valid techniques because we should not ask users about what happened and try to get them to reconstruct *why* they did one thing or another. This will only result in excuses and artificial explanations, i.e. the subjects will try to find a reason for their behavior. This type of argumentation requires that evaluators should instead *interpret* user actions when they happen and not ask subjects about the *reasons* for their actions (c.f Nielsen, 1993).

There are numerous arguments on both sides about this matter. Therefore, being aware of this problem and knowing how to handle the data and the information required, and simultaneously exercising care and self-reflection in every situation, might be the only way to overcome the difficulty.

Level of intervention

Another problem, in relation to the above question of time of response, concerns the level of intervention the evaluators should use. Different methods require different levels and approaches. For instance, Think-aloud techniques are used in many types of methods where the level of intervention is elaborate. In some methods, the evaluator is instructed to be more of a 'fly on the wall' during the test session. In others, for instance the Collaborative Evaluation method, the evaluator works together with the subject in order to ease the process of getting the subject to talk about the interaction in a natural way. When the 'fly on the wall' approach is used the evaluator influences the subject less, and it is easier to compare the data from different subjects. The disadvantage here could be that the user talks less and the resulting data is not as rich as one might want, meaning that it could be more difficult to interpret.

Required resources

One other crucial consideration when choosing an evaluation method or technique is availability of resources. The resources required include time, money, subjects, experts and availability of a usability lab (Dix et. al., 1998). The level to which usability labs are equipped varies widely. There are arguments in favor of large, complex labs to permit the collection of valid data. This might be acceptable

where, for instance, techniques are used that require large amounts of quantitative data to be collected and analyzed. In such circumstances the availability of suitable equipment such as cameras, screen and time-capture equipment would be crucial. Other techniques are less demanding in this sense, more qualitative approaches for example, such as Think-aloud protocol and all the inspection methods involving experts and all kinds of context-related methods, e.g. field studies.

It is true regarding the level of resources in relation to usability evaluation overall, that *time* is the most important resource when it comes to obtaining valid data from tests. Evaluations of usability in design projects *should* be conducted (a lot of projects completely lack usability testing), it should be carried out continuously throughout the project and as many techniques as possible should be used, in order to obtain as much data as possible. A frequently discussed view is that 'tests only need 5 users', as stated by Jakob Nielsen in 2000 (Nielsen, 2000). The idea is that it is useless to test with a larger number of subjects, given that the conditions in the tests are identical. Some studies have shown that after testing five users all usability problems are found (Nielsen & Landauer, 1993; Nielsen, 2000). Other studies, however, show that this is not the case. For example in a study by Spool (2001) the author instead argues that when "testing web sites – five users is nowhere near enough" (Spool, 2001). In complicated situations, where the target user group is heterogeneous and the tested environment is extensive – for instance a corporate or e-commerce web site – things are too complicated to support the use of such a small number of subjects (Spool, 2001). Nielsen, however, may have a point when the reference is to test design in which it is better to use resources to test more frequently with fewer users than to spend all the resources on one test with a larger number of subjects.

Web Usability

The same or similar techniques to those described above have traditionally been used when evaluating usability in the context of the World Wide Web (www or the web). However, this new evaluation context requires new approaches (Borges, Morales et al. 1996; Borges, Morales et al, 1998; Bevan, 1998; Nielsen, 1999; Spool, Scanlon et al. 1999; Olsson, 2000a; Olsson, 2000b; Kaasgard, 2000)

In order to find proper ways to evaluate usability on the web it is important to know the characteristics of the web site, i.e. the object of study (Shneiderman 1997). The author presents a number of bases for categorization of web sites, and these bases are: (1) by originator's identity (2) by the number of web pages in the site (3) by the goals of the originators, as interpreted by the designers, and finally, (4) by measure of success. These bases for categorization are further investigated later in this thesis.

Other work done in this research area discusses how web sites have characteristics that differ from those of traditional interfaces (Laskowski & Downey, 1997). Gaines, Shaw et al. (1996) discussing various problems on the web and trying to categorize sites according to the concepts of usability, utility and likeability arrived at the idea of a layered framework. Ratner (1998) tries to come to some conclusions concerning novice and expert users in learning environments that use Netscape. The author stresses that even if the educators have a specific goal and the students seem to be a homogeneous group, they are most certainly not in reality. This must be taken into account in the design of web-based learning environments.

Borges et al (1996,1998) present one example of conducting performance tests on the web in their initial heuristic evaluation of a number of university sites. Some of these sites were redesigned and finally a task analysis was carried out where users were measured while performing tasks. The usability team then arrived at a list of guidelines as a result of their test. However, they state very clearly that these guidelines can only be applicable to a narrow spectrum of web sites as support for design.

Spool et. al. (1999) present an example of conducting attitude surveys. The research team conducted a huge usability test of big corporate sites with the main focus on e-commerce. The report, or rather book, covers the study of nine sites, and the tests were much broader than in the other examples above. Instead of using the clock for measurement, which is common in performance tests, the research team used interview forms before and after the test combined with observations of the use of the web sites. The users performed tasks, but interest was centered more on *ways* of finding information, rather than on how quickly the information was retrieved. In another example, (Grose, Forsythe et al. 1998), a two-fold study showed that web-style guides differ from traditional style guides and stresses the fact that this must be further investigated.

Usability and user satisfaction in HCI research

The traditional definitions of usability (Nielsen, 1993; Shneiderman, 1998) tend to focus on factors that consider user productivity and performance, in order to ensure time and cost effectiveness in those situations where the system is applied. The majority of the most central measurable human aspects of usability evaluation are defined in terms of time for performing specific tasks, speed of performance and number and rate of errors made by the user (Shneiderman 1998).

Subjective user satisfaction, an attribute of usability that is concerned with how pleasant it is to use a system (Nielsen 1993), may be the most important aspect of the traditional definition of usability, when it comes to evaluating experience focused web sites. In the HCI research community most of the methods suggested for measuring this aspect among users, depend on various kinds of post-experience questionnaires. These questionnaires are to be answered after some other kind of user test (Nielsen 1993). Apart from some psycho-physiological methods that may be very complicated to manage in a usability test situation (Nielsen 1993), these questionnaires seem to be the only available way of abstracting the users subjective experience of a system that traditional usability offers.

This approach to the users subjective satisfaction provides data abstracted after the actual interaction with the system. Since the post-interaction questionnaires (c.f. Dix et al., 1998) provide no data about the users real-time experiences and thoughts when in contact with the system, this method may fail to identify important thoughts and aspects generated by the user during the interaction.

Usability and user satisfaction in other research fields

In other research and practice contexts, different methods and techniques for collecting and measuring user satisfaction than those described earlier are used and discussed. For instance, Jordan (2000) describes a number of techniques that are valuable in other research fields than mainstream HCI. Examples of these techniques are:

Empirical techniques

- Private camera conversation
- Co-discovery
- Focus groups
- Think aloud protocols
- Experience diaries
- Reaction checklists
- Field observations
- Questionnaires
- Interviews
- Laddering
- Participative creation
- Controlled observation

Non-empirical techniques

- Immersion
- Expert appraisal
- Property checklists

Those techniques are summarized below:

Empirical techniques

The techniques covered are, as mentioned above, divided into empirical techniques and analytical techniques. Below, the empirical techniques are described.

Private camera conversation

This method can be used with one person only or more. The participant enters a private booth, alone or with another participant, to 'talk to' a video camera about a product or concept. Investigators can give the participant(s) a list of issues to talk about, or a free approach can be used where the participant chooses what to talk about in relation to the object evaluated. The object is also present in the booth, if possible practically. If it is impossible to include the object, the setting has to be re-arranged.

Advantages of this method are that it generates an understanding of the people using the product and reveals the benefits given by the product and the properties associated with these benefits. Another advantage is that, as the evaluator is not present in the booth, the level of intervention is minimized.

Disadvantages of the method are that it is difficult as an evaluator to control the session, and it depends much on the participant(s) whether results are achieved. Furthermore, the analysis phase might be difficult due to the unstructured nature of data (Jordan, 2000).

Co-discovery

This method includes two subjects, often already friends or at least acquaintances. This is important, as the subjects feel less inhibited talking in such circumstances. They are assigned to the task of exploring a product, and an investigator is also present in the setting. This investigator might give instructions or feedback or help. The whole procedure is video taped. More or less structured tasks might also be assigned to the subjects. Another approach is to leave the subjects alone with the camera, and with no investigator in the room.

One *advantage* with this method is that it reveals people's initial responses to a product. The material might also be used very convincingly in discussions with design teams, as they see for themselves how the product is discussed and used by initial users.

Disadvantages with this method relate to control and analysis aspects. The session might produce information that the investigators are not interested in and the unstructured data could be difficult to grasp and analyze (Jordan, 2000).

Focus groups

The focus group method– originating from the discipline of market research – uses an often heterogeneous group of people in some kind of meeting situation. They discuss, for instance, users' experiences of, and attitudes towards, a product. Anything can be covered – aesthetics, functional aspects, where the product could be used and much more. Group members can be of all kinds, researchers, designers, target users, programmers and others. Usually, one person functions as a kind of coordinator in discussions and this person has some sort of agenda which focuses the group meeting. There are also different kinds of techniques used to trigger the discussions, such as prompts or scenarios. These are seen as 'ice-breakers' rather than rules or truths about any future use of the product discussed.

One *advantage* with the method is that it can be used at any stage in the design process, as participants can discuss a concept as well as an existing product. Furthermore, the discussions, which are only loosely controlled, can lead to other issues than those initially expected by the planners.

The method has the *disadvantage* that bad group dynamics may lead to destructive argument or small fights. It can also happen that some members of the group become too dominant and give the impression that the opinions raised originates from the group as a whole and not just themselves. Being a coordinator in a focus group meeting can be a very demanding assignment. (Jordan, 2000)

Think-aloud protocols

This method has been described earlier, and is used when participants are asked to verbalize thoughts that arise in using a system or product.

The *advantages* are, that it may be possible to understand not only how people react in a certain situation, but also *why*. It is also an efficient method when used with a small number of subjects, because each subject can give a rich picture of the use of the system or product.

One *disadvantage* is that subjects construct reasons for their behavior. When things 'just happen' people have a way of finding arguments why they behaved in a certain way. Another disadvantage is that it is too easy for the evaluator to intervene, such as when the subject gets stuck in an action. At that point it may be impossible to say anything about what would have happened in an authentic situation. (Jordan, 2000)

Experience diaries

Participants are given diaries containing mini-questionnaires where they can make notes about their daily use of a product or system over a period of time. The questionnaire should be sufficiently easy, i.e. not too extensive, to encourage completion every day, but thorough enough to cover what the evaluators want to cover. Generally speaking one page is enough. Usually the diaries are filled in without the intervention of the evaluators. Since there is a trade-off between content and length here, it is important to be aware of the choice of questions to be included.

The method has *advantages* both in terms of time and effort for the evaluators as well as the fact that no specific evaluation setting is needed. The subjects can fill in the diary anywhere. Another advantage is that the method can give a picture of usage over time, which is difficult with many other methods.

The *disadvantages* are control aspects. There are no guarantees that the participants will fill in the diaries as planned. Participants may drop out, or fill in the diaries in such a way that the data might be useless. Such comments as 'I use the system because it is nice', for instance, create problems for the evaluators – what does 'nice' mean to this person. Finally, the method can only be used with complete systems or products. Prototypes should not be used over time in an authentic way, so the method is of more a help in future re-designs than as a base for iterative design. (Jordan, 2000)

Reaction checklists

In this method, a checklist of potential reactions is used, for instance, on the basis of *The four pleasures*[7]. The checklist is structured with regard to pleasures and statements are written out in documented form – a checklist. One example concerning psychological pleasure is 'The system is fun to use'. Participants are asked to mark their reactions, when using the system or what they believe their reactions would have been if they had used the system. The method can be used both for positive and negative reactions and statements. The method can also be extended to cover, for instance, possible features to be added into a future system. Here, future users can mark what features they would prefer to use in the system if they could choose. Advantages are that the method is cheap, since it is undemanding with regard to the evaluators' time and resources. One disadvantage with the method, compared to Think-aloud protocol, is that no guidance is provided as to why the responses appear as they do.

Field observations

To observe users of a product or a system contextually, i.e. in the environment where they usually use the system or product, has many advantages over a lab experiment. Participants are usually less tense, since they are in a context they recognize, and they do not have to feel that they *themselves* are being observed and tested, which is a common reaction in lab evaluations. Usually, the observations are conducted over a longer period of time, and participants are followed during, for instance, one working day. Field observations can continue for months. One major disadvantage is the extensive amount of field data that has to be analyzed from sources such as field notes or videotapes. Another *disadvantage* is that the data received from field observations might be measuring something other than, for instance, like or dislike of a system. The observed participants might respond to something in the environment instead of to the system or product in focus for the evaluators. (Jordan, 2000)

Questionnaires

A questionnaire is a printed list of questions, either open-ended or with fixed responses. The fixed responses can be of different types. The respondents should mark one or more alternative responses to a question, or are given a statement with which they have to agree or disagree – often using different numbers of choices on a scale. The scale can be numbered from 'I strongly disagree' to 'I strongly agree'. Usually, regardless of the scale used, in words or numbers, there is a 'neutral' choice in the middle. There are also standard questionnaires, for instance the 'System Usability Scale' (SUS), 'Task Load Index' (TLX) or 'Software Usability Measurement Inventory' (SUMI). These are used both in research and practice.

Questionnaires have the *advantages* that they can be checked in advance to avoid problems of validity and reliability. They are cheap in relation to the number of respondents that can be involved. The data are also quite easy to analyze in a structured way since the structure is given, especially in a questionnaire with fixed responses.

The *disadvantages* are low level of completion of the questionnaires, if filled in remotely by respondents. If questionnaires are sent out by mail, on average 25 percent of them are completed. One might think that this could be overcome by sending out four times as many questionnaires, but the biggest problem here is that the people who make the effort to return completed questionnaires cannot be regarded as representatives of the target group. Rather it is people with strong opinions in the matter who complete the questionnaire. (Jordan, 2000)

Interviews

This is an oral technique where an evaluator poses questions to respondents in various ways. Here, the number of techniques and their scientific or methodological bases differ a great deal. However, on a general level, three techniques are used – unstructured, semi-structured and structured techniques. The method has also been discussed earlier in this chapter. (Jordan, 2000)

Laddering

Laddering is a type of interview technique where *why* questions are repeated after every statement made by the subject. The method has its roots in marketing. The overall idea is to understand links between the formal and experimental properties of a product and a system, product benefits and the characteristics of people using the system. An interview session could go as follows:

Interviewer: Please tell me something that you like about the system (or product).
Participant: I like it because it is fun to use
Interviewer: Why is it important that it is fun?
Participant: It is a game – they should be fun.
Interviewer: Why should games be fun?
Participant: Games usually are.
Interviewer: Why is it important that this game is similar to others?
Participant: I want to know what to expect when I start to play a game.
Interviewer: Why is it important to know in advance what you are getting?
Participant: I want to feel safe.
Interviewer: Why is important to feel safe?
Participant: That is just the way it is.

The above example is completely fictitious and is presented only to give an example of what a session can be like. After the session a possible ladder of this example might be shown as in Figure 1.8.

The *advantages* are that the method enables investigators to gather information about the formal properties, experimental properties, desired benefits and characteristics of people that are targeted as potential users of the system or product. Furthermore, the method produces information about the relations between the above-mentioned aspects. The method can also be used at any stage in the design cycle, since the participants can be asked to comment on prototypes as well as completed systems.

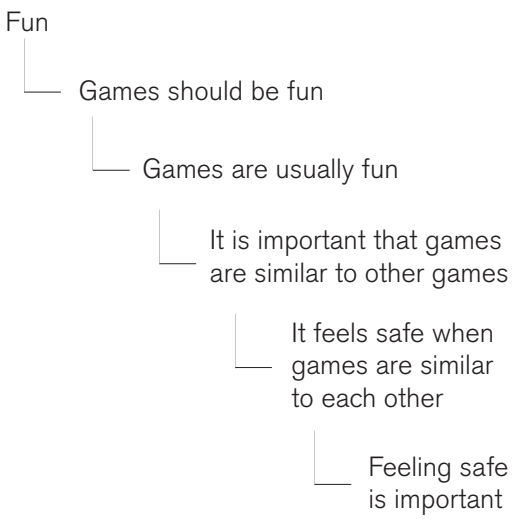

Figure 1.8 An example of the application of the technique 'laddering'.

One *disadvantage* is time - the method is time-consuming both for participants and investigators and furthermore can be rather demanding. The participants may feel badgered by the continuous stream of *why* questions, and they might feel compelled to adhere to their earlier statements, even if they come to realize during the questions that they would like to change their opinion in some way. It might happen that earlier statements do not fit with any rational explanations, which are acceptable from a methodological point of view, but since the questions are structured the way they are, this fact might not be apparent. One result of this might be that participants experience the discussion or interview as somewhat negative or pressuring, which is unfortunate even from an investigator's point of view. The method also puts a heavy pressure on the investigator, since she/he has to react immediately and produce proper follow-up questions – on the basis of earlier answers. Furthermore, as there is a risk of creating inconvenient situations for the participants, it is important not to put too much pressure on the respondent. The method requires a lot of experience on the part of the investigator for it to work as planned. (Jordan, 2000)

Participative creation

This method uses teams of designers, presumptive users and HCI specialists. The team discusses issues related to the design of a product or system. This could involve a list of requirements for instance of important features or benefits wanted from use of the product. It may also include aesthetic aspects. On a general level, the method seems to be similar to the *Focus Group* method. However, the *Participative Creation* involves team members in 'hands-on' design more than the other method does and demands that team members see themselves as part of the design process.

The *advantages* are that the method involves possible users of the system or product at earlier stages in a direct way, which might be very fruitful for the final result. Designers can reveal their ideas even before prototypes have taken shape and they get to understand the demands and requirements at an early stage in the design cycle.

The *disadvantages* are that the method is very demanding for the participants – both in terms of time and work required. Sessions usually last for three hours or more. The participants are required to make design judgments which only designers are trained and educated to do. In some sense this may be impossible since not everyone can think as designers do. On a communication level, the method also

has constraints. To make strong arguments directly to designers about their ideas can be embarrassing and difficult, both for HCI specialists and end-users. It could also be difficult for designers to accept criticism under these circumstances, no matter how constructive that criticism is. (Jordan, 2000)

Controlled observation

This method is a form of observation technique, where a use situation of a product or system is observed. It differs from field observation in that it does not cover contextual aspects of real usage. Instead, this method is a formally designed investigation with more control of the observed use session. Investigators strive to gain more 'noise-free' data in comparison to the *field observation*. The sessions are designed according to what aspects of a product or system evaluators want to cover. Subjects are required to complete tasks with varying levels of structure, i.e. the approach can be more or less free. Task orders are balanced between sessions and overall different pros and cons in different conditions are considered and changed between sessions. This is done to meet the main purpose of covering all possible events that might appear in authentic use. However, here the main theme is 'control'.

The *advantages* of the method are that it can deliver 'pure' data, with less noise, than for instance *field observation*. Furthermore, it might also be a suitable method for comparing different aspects of a system. If correct control and balance are achieved, data validity is high.

One *disadvantage* is that the level of control and balance may mean that no authentic use situation is obtained. The subjects may react according to an evaluation session instead of giving data that can be seen as examples of a real-use situation. If this occurs, no matter how well designed the evaluation is, the data is useless. (Jordan, 2000)

Non-empirical techniques

In the non-empirical techniques no subjects are used, but evaluations are instead conducted by experts. This is similar to inspection methods, described earlier in this chapter. The non-empirical techniques are described in more detail below:

Immersion

This method can be placed somewhere between empirical and non-empirical approaches. It is conducted by an investigator who experiences a product by himself or herself. The investigation of the product or system is initiated at the very start of the product cycle, i.e. in the shop where the product is displayed. The investigator buys the product and then takes it home, as if it was the first time he had seen or used it. The packaging, installation instructions and other manuals

are evaluated and then the investigator proceeds to use the product for a period of time, for instance one week. The investigator can use an external person, who video-records all interaction with the product. The main reason for using this technique is to discover how users or buyers of the product react and how the product is presented, in the store, in packaging and in the related material, such as manuals. In addition, initial use is in focus in this technique. The evaluator can give feedback to designers and management about changes that should be made. The advantage is that the technique is a 'real-life experience' approach, comparable to field observations. A disadvantage is that it might be difficult for evaluators to disregard the fact that they are not first-time buyers or users of the product. (Jordan, 2000)

Expert appraisal

This technique is used by evaluators considered to be experts in a target area related to the product evaluated, on the basis of education, profession or other experience. The expert evaluates the design of the product to see if it can be regarded as being pleasurable for target users. The expert may also have specific knowledge in, and experience of the target group of users. For instance, if the users have a handicap of some kind, the expert might have specific knowledge in this field. Sometimes this technique is used with a number of experts working separately or in groups, to evaluate the product. The advantages, as in all expert investigations, are that no user participants are needed and that the investigation or evaluation can be conducted on non-complete products, as experts may be judged to have a higher level of understanding of products in that stage of the product-cycle. (Jordan, 2000)

Property checklists

The technique using property checklists can be considered to be more structured than expert appraisal in that it is based on a checklist, specifically developed to describe required features or aspects of the specific product. However, it is much more difficult, or even impossible, to find a standard list of properties for complex systems or products measured mainly by pleasurability, compared to evaluation of traditional functional systems, where such methods as Heuristic Evaluation could be used, with a standard heuristic list. However, the checklist in the Property checklist method may be produced by the designers of the product. The advantages of the technique are that no subjects are required and that it can be used throughout the design process. As it is based on a checklist, it is also less person-dependent, regarding experts, i.e. the checklist offers some kind of support. One disadvantage can be that the design of the property checklist might be somewhat speculative. This might, in the end, result in less valid results. (Jordan, 2000)

Usability and fun: An overview of relevant research in Human Factors and HCI

The emerging focus on experiences in different media is highlighted in a number of related literatures, of which some are summarized below. Further, a large number of studies in relation to usability and fun have been published in the area of HCI is described more in detail.

Fun, entertainment, and IT product quality

Fun and entertainment are becoming increasingly important in almost all uses of information technology (IT) (Wolf, 1999; Pine II and Gilmore 1999; Monk, Hassenzahl et al. 2002). The entertainment industry is expanding rapidly, and the number of crossover activities between various media, such as movies, television, computer games, toys and the web, is exploding. No longer do we see only stand-alone entertainment products, such as video games and toys. Today entertainment products are sold as "packages" designed for use with different media (Wolf 1999; Bolter & Grusin, 2002). Historically, various types of media have been used in marketing of products of all kind. Now it is hardly possible to discern any boundaries between the marketing of the products and the products or services themselves (Pine II and Gilmore 1999). A movie nowadays comes with a web site with movie clips, and stories from the production process as well as clips from backstage. Images from the movie can also be downloaded; post cards with a movie-related content can be sent and games can be played on the web site based on the story of the movie. DVD of movies are also on sale which include additional material such as sequences or storylines not shown in the movie. There seem to be no limit for crossover features in the entertainment business today. What is happening has been termed a Technology Convergence (Pavlik, 2000; Pinhanez, Karat et al. 2001; Bittanti, 2002; Bolter and Grusin, 2002; Frank and Lundblad, 2002).

IT-based entertainment can take a variety of forms. Firstly, there are the more interactive types of entertainment: games of all kinds using different types of technology; *Virtual Reality*-chats for making new friends; applications for mixing one's own music, design and send virtual post cards etc. However, not all such entertainment is interactive, much is designed for passive use, such as music videos to watch, virtual museums and art galleries to visit, live images of Earth from a view of a satellite or space shuttle to be explored. Technology driven entertainment knows no bounds (Pine II and Gilmore 1999; Wolf 1999).

Yet, fun and entertainment as qualities are only rarely discussed and measured in the context of IT. HCI has focused mainly on work-related systems and their

measures of success (Blythe & Wright, 2003). There were some attempts to cover questions related to fun and entertainment in the HCI field in the early 1980's by Malone (1980; 1982), but this was not followed by other researchers in the field (Carroll & Thomas, 1988). The interest of the Human-Computer Interaction (HCI) community in pleasure and fun in relation to IT is now beginning to grow[8], but still the lack of a coherent understanding and theoretical base is emergent (Carroll and Thomas 1988; Monk, Hassenzahl et al. 2002; Blythe & Wright, 2003).

When fun and entertainment become the objects of study it raises further ontological questions such as what is enjoyment and fun, where does it appear, where and how can it be measured and what measures should we use? (Aboulafia et al., 2001, Monk, 2002; Monk et. al., 2002; Blythe & Wright, 2003). Many philosophers throughout history, for instance Plato, have raised these questions but no one to date has produced a definitive answer.

In the past, Human Factors (which in this context can be seen as equivalent to HCI) had a minor impact on 'design', as in product design. Jordan (2000) divides the role Human Factors have played in design into three historical phases: (1) Being ignored – fifteen to twenty years ago very few human-factors specialists were hired by industry, which did not consider human factors of any importance. (2) "Bolt-on" human factors – in this era some specialists were hired, but only to add a 'nice' interface to systems and products already structured and designed. (3) Integrated human factors – lately, human factors have come to be considered relevant throughout the design process and specialists are consulted as part of design projects from the very start (Jordan 2000). Even if the latest phase is referred to as integrated human factors, the optimal scenario with research about HCI occupying a central position in the design of pleasure products and technology still seems distant. One can speculate numerous reasons for this. Firstly, the entertainment industry as a whole traditionally assigns only a minor in design to audiences and users. There are exceptions to this rule[9] but in general products are launched without any input from potential end users, i.e. the audience. Second, and perhaps more worrying, the HCI community seems to neglect these types of objects of study (Monk, Hassenzahl et al. 2002). Even if some early work was conducted by Malone (1984) for example, who proposed heuristics for enjoyable interfaces, there is little interest in further research on the topic (Monk, Hassenzahl et al. 2002). There may be several reasons for this but the fact that cognitive psychology, which must be seen as an important discipline in HCI has no tradition of carrying out such studies is probably one. Another may be the subjective nature of pleasure as an object of study. It is hard to produce significant results in that type of research (Monk, Hassenzahl et al. 2002). A third

reason might be the process of funding research. Those responsible for funding research, worldwide, may have been reluctant to fund research into pleasure and fun, concentrating resources rather on workplace related technologies and systems (Monk, Hassenzahl et al. 2002)

Currently, there are three basic perspectives on enjoyment and fun in HCI research: (1) *Usability reductionism*, where enjoyment is simply seen as a result of ease of use. (2) *Design reductionism*, where enjoyment and fun are features to be added on by graphical and industrial designers. Finally, (3) *market reductionism* where the concept of fun is seen only as an advertising tool. These perspectives provide almost no support in the extensive work of designing and evaluating entertainment, enjoyment and fun. Research on this topic is beginning to clarify the complexity of users' needs in such contexts (Monk et. al., 2001). Yet, in many cases understanding the concepts of pleasure, entertainment and fun is neglected, as is the case in usability evaluation (Thomas and Macredie 2002). What does usability mean in the context of fun and entertainment, and how can we evaluate this type of technology? These are key questions for future HCI research.

Usability, entertainment, experience, and the web

The current state in research related to entertainment and fun, within the area of HCI and other disciplines, can be summarized as follows:

Recently, discussions have surfaced about the need for new types of measures. A more holistic view, compared to the cognitive and physical view of products, is emerging and different types of human-product relationships are being explored. Jordan (2000) constructs an explanatory framework and discusses four different types of pleasures; Physio-Pleasure, Socio-Pleasure, Psycho-Pleasure and Ideo-Pleasure. These are of more general types of pleasure and are not confined to the web. However, they may well be used as a support in designing tests and analyzing data. Pleasure in relation to product design and system design has been covered in recent research (c.f. Bonner 2002; Creusen and Snelders 2002; Popovic 2002; Reinmoeller 2002; Ruecker 2002), as also has pleasure in relation to usability (c.f. De Angeli, Lynch et al. 2002; Noyes and Littledale 2002; Overbeeke, Djadjadiningrat et al. 2002).

In the area of *Affective Computing* (Picard 1998; Höök, Persson et al. 2000; Picard, 1997), research is increasingly focusing on issues related to interfaces and systems that imitate human behavior, such as robots, interface agents and assistants etc. confining around such questions as how they should be constructed and how people react to them. Some studies have been conducted where designed affective systems were evaluated with users, giving rise to interesting discussions (c.f De Angeli 2001; Wiberg and Wiberg 2001). User experiences and emotions

and how they are affected by information technology of different kinds have also been covered in recent research (c.f. Bates, 1994; Marcus, 2002) in both theoretical and concrete terms (c.f. Aboulafia, Bannon et al. 2001).

As esthetics play an important role in our object of study, i.e. entertainment web sites, this is obviously a domain that should be covered here. Esthetics and beauty are taken up in the research and are discussed in the context of IT and IT use (c.f. Schenkman and Jönsson 2000; Tractinsky, Katz et al. 2000; Karvonen, 2000). The questions generally asked in this context are how can esthetic values be judged and how does that relate to usability. Schenkman & Jönsson (2000) explored first impressions of web pages when presented to subjects in order to discover the kind of web pages they preferred and what subjective factors determined their overall impression of the web page. They found that of four important factors, beauty, i.e. esthetics, the best predictor of the overall impression of the web page. Tractinsky et al (2000) discussed the relation between aesthetics and perceived usability and proposed a correlation between the two. After an empirical experiment with an ATM machine they were able to show that the hypothesis appeared to be correct, i.e. the esthetics of the system affected the post-use perception of both aesthetics and usability. Karvonen (2000) emphasized the need for knowledge concerning aesthetic theory and history when discussing interface design, instead of inventing new frameworks for every new type of interface. In the usability field Nielsen (1999) argues that simplicity should be the guideline when designing usable web pages – an argument that has been familiar in aesthetics for three hundred years. Karvonen also argued that aesthetic values might not be individual and unique for every person. Instead preferences might be more generally based on styles, trends or fashion – and can therefore also, at least to some extent, be categorized.

The number of research publications concerning computers and information technology in relation to entertainment and fun, is increasing rapidly (c.f. Thomas and Macredie 1994; Draper 1999; Agarwal and Karahanna 2000; Amant and Young 2001; Pinhanez, Karat et al. 2001a; Pinhanez, Karat et al. 2001b; Pinhanez, Karat et al. 2001c)

Experience and flow

Pine II and Gilmore (2000) provide some guidelines for exploring the concept of experience, and present a number of theoretical frameworks for shedding light on important aspects of experiences in general. For instance, one fruitful way to categorize experiences would be to use two dimensions, absorption vs. immersion and participation vs. non-participation. Pine II and Gilmore present a framework based on these dimensions described further in Chapter 2. However, this work is on a more general level of abstraction, which includes all experiences and not

just those related to the use of IT. In the context of IT research, mainly in HCI, a number of attempts to come to grips with the idea of IT-related experiences and affect have been presented. This includes both empirical and theoretical work. The main purpose of the empirical work can be said to be the sharing of results concerning the evaluation and measurement of human experiences of using various types of IT systems (c.f. Höök et al, 2000; Höök & Svensson, 1999; Scheirer, J., 2002). Theoretical work on the other hand often focuses on how experiences and affect in the context of IT can be understood on a more general level and the results often include frameworks and theories of different kind (c.f. Aboulafia et al, 2001; Huang, 2003; Forlizzi & Ford, 2000; Norman, 2002; Norman, forthcoming).

The concept of 'flow' is sometimes also used when investigating experiences. This concept was invented by Csíjszentmihályi (1990) and describes the state of a mind when it is experiencing something. Some empirical work based on this concept has been conducted in the context of web use (c.f. Chen et al, 1999; Novak et al, 1998).

User satisfaction

When discussing fun and entertainment in the context of usability, the most closely related notion, as already mentioned in Chapter 1, is 'user satisfaction'. Usability evaluation has traditionally focused on aspects related to function, such as efficiency, number of errors etc. (cf. Nielsen 1993). However, research into the aspect of user satisfaction has so far been neglected in the research discipline of HCI (Lindgaard & Dudek, 2003). Similarly the concept of user satisfaction been fully explored, nor have methods for evaluating this aspect been developed. Data regarding user satisfaction are mainly subjective, which may be one explanation for the lack of research in the area. In addition, one of the most influential research disciplines regarding usability evaluation in HCI research, has in the past been cognitive psychology and no tradition exists in this field of investigating subjective aspects such as user satisfaction, either conceptually or methodologically. Thus this may constitute another explanation for the past absence of research into user satisfaction. However, the subject has been covered in more recent research literature (c.f. Evans, 1993; Harrison & Rainer, 1996; Mahmood et al, 2000; Chin & Lee, 2000; Lindgaard & Dudek, 2003).

Fun and entertainment

Fun and enjoyment are to some extent related to user satisfaction. There is a marked trend towards entertainment on the web. Evaluation of entertainment web sites, specifically designed to be affective vis-à-vis the user, challenges traditional

evaluation frameworks in the area of Usability Engineering (Olsson, 2000a, ibid. 2000b). Recent studies of Internet use or "surfing" show that when people are enjoying themselves, time parses unnoticed and they focus mainly on the current activity (Agarwal & Karahana, 2000).

Various views and aspects of entertainment on the web are covered in research. Some authors mainly consider the traditional view of entertainment, where there is a sender – receiver situation, with no or only a little interaction, 'webTV' for instance. One large study of such web pages questioned whether web entertainment could be passive and the results showed that this could in fact be the case – users preferred less clicking and more watching (Pinhanez et al., 2001a; ibid. 2001b; ibid 2001c). Similar studies, from the same research group produced similar results in that entertainment on the web is interactive, but not exclusively so – entertainment that is only watchable also resulted in user satisfaction (Karat & Karat, 2003). Other types of research focus on high level interaction and the social aspects of entertainment (c.f. Amant & Young, 2001). Quite a large number of studies deal with entertainment by focusing on games – both web-based and stand-alone games - (c.f. Thomas & Macredie, 1994; Draper, 1999; Fabricatore et al, 2002), where for example, the 'playability' of games is investigated and measured Research needs to highlight the salience and possibilities of user studies and other evaluations of usability-related aspects as, with few exceptions, the games industry has little or no past experience in this field (c.f. Federoff, 2003). Playability is one aspect that is further developed in the context of this study. To analysis fun as a software requirement can be complex, as exemplified by Draper et al. (1999).

Discussion

The above discussion of related work indicates that there is a certain lack of focus on methodological considerations regarding usability evaluation in general and web usability in particular. There are few studies on how users react to entertainment technology (e.g., Pagulayan et al, 2003, Karat & Karat, 2003, Desmet, 2003). Some researchers argue that completely new methods are needed to deal with fun and pleasure in the are of HCI (c.f. Thomas and Macredie 2002). This might well be true, but as we have so little knowledge about how traditional usability evaluation works in the context of fun and entertainment work, it is difficult to argue for new approaches. Further studies are much needed (c.f. Carroll & Thomas, 1988; Thomas & Macredie, 2002; Pagulayan et al, 2003, Karat & Karat, 2003, Desmet, 2003; Nielsen, 2003, Monk et al, 2002). Arguably, usability evaluation methods can have a substantial impact on designing pleasurable and enjoyable systems and web sites (c.f. Pagulayan et al, 2003, Nielsen, 2003).

Therefore, even though extending traditional usability to include evaluation of fun an entertainment appears to be a sensible research objective in the context of existing HCI/Human Factors research, this research provides little guidance on how this objective can be accomplished. This conclusion constituted a point of departure for this thesis and resulted in extending the underlying framework of analysis from immediately relevant areas (that is, HCI and Human Factors) to a more general perspective, which is reflected in the sections comprising the rest of this chapter.

The next chapter discusses entertainment and fun in general as well as in the context of web usability. Central concepts such as pleasure, experience, flow, fun and entertainment, will be defined and further developed. This is intended to give the reader an understanding of how entertainment web sites are defined and categorized in this study so that they will be able to better judge the findings and conclusions of the study conducted in this thesis.

Footnotes

[1] For further information please see pp. 24-25 in Nielsen (1993).
[2] For further discussion about methodological considerations, see the section entitled 'Methodological considerations', and more specifically the discussions around time of response and level of intervention.
[3] The approach added to the list by Nielsen (1994) is *Theory-based reviews*.
[4] Here, the categories of usability from Nielsen (1993) are chosen, although there are other ways to categorize usability. Nielsen (1993) was chosen because of its wide application within the HCI community and the literature.
[5] References to the model: The three inner circles are based on Nielsen (1993, pp. 26-37). Discussions about the placing of Inspection Methods are found in: Cognitive Walkthrough (Virzi, 1997, p. 707); Keystroke-level model (Dix et. al., 1997, p.415); GOMS (put in brackets as some do not consider GOMS to be a usability evaluation technique (c.f. Virzi, 1997, p. 708))(Dix et. al. , 1998, p.415); Model-based evaluation based on Design Rationale (Dix et. al., 1998). Heuristic Evaluation (for Errors and Memorability) (Newman & Lamming, 1995, p. 183).
[6] Note here that this thesis investigates the last type of system and that this case will be further investigated, developed and discussed.
[7] For further description – see Chapter 3.
[8] The number of related research conferences which include the areas of pleasure and fun is growing (c.f. "Computers and Fun" and "International Conference for Affective Human Factors Design"). Respected research journals are publishing special issues on the subject (c.f. Monk & Frohlich, 1999)
[9] A documentary, shown in Sweden in 2001, pointed out that the world famous and award-winning television show 'Friends' is partly tested on live audiences to see, for instance, if the audience correctly understands the jokes.

Chapter 2

Entertainment and fun as aspects of web usability

There are a number of conceptual problems associated with including aspects of fun and entertainment in web usability. One of the main challenges is the *operationalization* of fun in relation to usability. Users of web sites may have a wide range of phenomena that can be classified as fun, excitement, pleasure, etc. Arguably, not all of them are relevant to web site evaluation and many of them in themselves have nothing or very little to do with web sites. They can be caused by idiosyncratic conditions, the general context, or memories, which may be impossible to anticipate and generalize.

To develop or re-design standardized usability methods that aim to take into account fun and usability it is necessary to specify the object of our study, that is, what exactly is to be evaluated. In other words, a critically important precondition for research into the issue is an operational definition of "evaluation of entertainment or fun": a definition which would make it possible to design and conduct a concrete, objective study that would also make sense in the general context of web design. One of the aims of this chapter is to deal with that challenge. The chapter identifies aspects of fun and entertainment that are relevant to usability, mainly with reference the web, and provide an operational approach to evaluating fun, organized as set out below.

The chapter is organized as follows. The first section looks into definitions of entertainment and fun, both in dictionaries and related literature in HCI and art. It is concluded that although the definitions clarify differences and similarities between these two concepts and highlight some important issues, they do not provide enough guidance for their operationalization for usability on the web. In the second section a selected set of theories of fun is presented. As in the case of definitions, the theories emphasize important conceptual distinctions relevant for

understanding the nature of fun in general but are not directly applicable to web usability. The third section, based on the analyses in sections 1 and 2, introduces the approaches to operationalization of evaluation of fun and entertainment adopted in this thesis, that is, *evaluation of entertainment web sites*. The section discusses the rationale behind this approach, distinctive features of entertainment web sites (EWSs) compared to web sites in general, and the relationship between the concept of web sites and their design form. The forth and final section discusses the advantages and limitations of adopting the above approach and provides a logical link to the next chapter dedicated to a detailed exposition of methods and ways in which to evaluate, judge and re-design methods which is, in fact, the main aim in this thesis.

Entertainment and fun: Definitions

'Entertainment' as a general expression has many interpretations but most people have an idea of what entertainment is. It is not easy to define just simply because everyone seems to know what it is. Furthermore, it is a somewhat common-sense idea (Dyer, 1992). However, in order to pinpoint what to evaluate in this study, i.e. entertainment in the context of IT use, it is necessary to formulate a closer definition. Below, more general definitions of entertainment are presented starting with that of the Oxford English Dictionary:

> *"The action of occupying a person's attention agreeably; amusement"*
> (The new shorter Oxford English Dictionary)

As mentioned earlier, entertainment in this study is explored in relation to the use of IT. This type of situation is often understood to be interactive, but the level of interaction differs in the context of entertaining IT products or systems. In some cases, such as WebTV, interaction is minimal. However, in other cases, for instance in games, the level of user-system interaction is obviously high. Even if the level of interaction varies in the context of entertainment IT use it still has to be considered. This, however, will be problematic if only the above definition of *entertainment* is used, as it provides no or very little input about whether the audience is considered to be active or passive. In this definition the focus lies on the activity of the supplier of the entertainment, which may imply that the audience is passive. An American dictionary, *Webster's,* defines entertainment as:

> *"1:Provision for guests especially in public places (as hotels and inns) 2: Amusement 3: Recreation 4: a means of amusement or recreation; esp: a public performance."*
> (Webster's New Encyclopedic Dictionary, 1995, p. 335)

This second definition indicates that entertainment, here may be regarded as a public performance. This indeed implies a passive audience. As mentioned, this is a problematic definition to use in the context of entertainment IT. In order to find a more easily applicable definition to serve as a basis for considering how to regard entertainment in the context of IT use, the concept of *interactive entertainment* is the notion closest related to this thesis. It is defined as:

> *"All types of amusement in which the involved persons could change the course of events"*
> (Fjellman & Sjögren, 2000)

However, if this definition is used, all types of games, such as Monopoly etc. are included. This is not exactly what would be considered interactive entertainment in the context of IT use, which was the above authors' intention. This is why the authors Fjellman & Sjögren (2000) give two further criteria for defining interactive entertainment.

> *"The interactive entertainment should be experienced with some kind of digital technology…the second requirement is that this digitally delivered entertainment should offer active engagement."*
> (Fjellman & Sjögren, 2000)

The above definition of interactive entertainment with the additional requirements is the definition of entertainment most applicable in the context of IT use found in the related literature. This definition will be further investigated and developed later in this chapter in relation to the object of study in this thesis. It might also be of importance to investigate the relation between entertainment and fun and to this end, standard dictionaries as well as dictionaries of synonyms were consulted. First, the American Heritage Dictionary of English[1] defines 'fun' as:

1. *A source of enjoyment, amusement, or pleasure.*
2. *Enjoyment; amusement: have fun at the beach.*
3. *Playful, often noisy, activity.*

In Table 2.1 the information gained from a dictionary of synonyms is presented, showing the two concepts in parallel[2]:

	fun	entertainment
Function:	noun	noun
Definition:	amusement	amusement
Synonyms:	absurdity, ball, big time, blast, buffoonery, celebration, cheer, clowning, distraction, diversion, enjoyment, entertainment, escapade, festivity, foolery, frolic, gaiety, gambol, game, good time, grins, high jinks, holiday, horseplay, jesting, jocularity, joke, joking, jollity, joy, junketing, laughter, merriment, merrymaking, mirth, nonsense, pastime, picnic, play, playfulness, pleasure, recreation, rejoicing, relaxation, riot, romp, romping, solace, sport, tomfoolery, treat, whoopee	ball, bash, big time, blow out, celebration, cheer, clambake, delight, dissipation, distraction, diversion, divertissement, enjoyment, feast, frolic, fun, gaiety, game, good time, grins, high time, laughs, leisure activity, merriment, merrymaking, party, pastime, picnic, play, pleasure, recreation, regalement, relaxation, relief, revelry, satisfaction, shindig, sport, spree, surprise, treat, wingding
Concept:	social action	social action

Table 2.1 A comparison between the concepts fun and entertainment.

As these definitions of entertainment and fun show, the two concepts somewhat overlap in meaning. For example in this dictionary there are twenty-two that are common to both words. This is 42% of the total of 52 synonyms for fun and 51% of the total of 43 synonyms for entertainment. These findings indicate a general correlation between the two ideas of 40-50%. However, it is also important to recognize that in some cases these notions differ in meaning. Entertainment for instance could be differentiated from fun or amusement as described by Langer (1977):

> *"But…entertainment is not essentially frivolous, like amusement. The latter is a temporary stimulus, the "lift" of vital feeling that normally issues in laughter. It is generally pleasant, and sometimes erroneously sought as a cure for depression. But entertainment is any activity without direct practical aim, anything people attend to simply because it interests them. Interest, not amusement nor even pleasure, is its watchword."*
> (Langer, 1977, p.404)

To further highlight the difference between these two concepts of fun and entertainment, it is worth mentioning entertainment in relation to tragedy and comedy – both understood as types of entertainment, but not necessarily situations where fun and entertainment be seen as equals. Here, tragedy and comedy are considered as entertainment, but tragedy cannot be regarded as fun. Langer (1977) discuss this:

> *"Shakespeare's tragedies were written for an entertainment theater in which people sought not amusement but the exhilaration of artistic experience, overwhelming drama."*
> (Langer, 1977, p. 404)

The ways in which the ideas of fun and entertainment correspond or differ could cause a great deal of confusion. This may be an argument for a thorough analysis of the relation between the two, resulting in a conceptual model of the relationship. On the other hand, if such a thorough analysis and conceptualization had been made in the context of this study, the participants in each session in the study would have needed to know about this conceptual framework, in order to give feedback about web sites in accordance with it. However, such an approach must be considered as an intervention in the sessions, i.e. it would probably undermine the aim of providing as natural and authentic a setting as possible for the users of the web sites. For this reason, no conceptual model concerning the relation between fun and entertainment was used in the study. Arising from this it was rather difficult to interpret participants' ideas of concepts such as fun and entertainment. When participants used this type of word, evaluators tried to overcome the interpretation problem by using follow-up questions about meaning. Throughout this thesis the words entertainment and fun are used largely interchangeably to refer to one concept, but most often entertainment is used. On the basis of the arguments of Langer (1977), as described above, this interchangeable approach might be problematic with reference to entertainment in the form of drama or tragedy. However, since the form of entertainment presented in the object of study in this thesis, i.e. entertainment web sites, includes no or only minor elements of tragedy or drama this can be seen as an insignificant problem.

Related theories and frameworks

In HCI and related disciplines, a number of related theories and frameworks have been used when investigating experiences, pleasure, user satisfaction and fun. Below, some commonly used and frequently referred to examples of such theories are presented. This is done to investigate whether these theories can provide any guidance for how to operationalize entertainment in relation to usability evaluation, i.e. can any of these theories indicate which direction usability evaluation of entertainment-related IT use might take. More specifically, the related theories and frameworks covered below are:

- Computers as theatre (Laurel, 1993)
- The four pleasures (Jordan, 2000)
- The experience realms (Pine II & Gilmore, 2000)
- The notion of flow (Csíjszentmihályi, 1990)
- Emotion and design (Norman, forthcoming)

Computers as theatre

This theory examines how to consider the relation between systems and humans, and what metaphor to use for this purpose. The book *Computers as Theatre* by Brenda Laurel (1993), in which this idea is presented, was seen as groundbreaking when it was published in the early 1990s. In the book, the author describes computers in general, and criticizes earlier models of interfaces. These models concerning interfaces in HCI, according to Laurel, developed from a first stage in which the pre-cognitive view of interfaces ruled. In this first stage, there was no intent to include human aspects in IT design. In the next phase, the mental-models view of models of interfaces, it was argued that the user had a mental model concerning the system and the system should have an 'understanding' of the user. This is where the problems began, Laurel continues, such as the problem of how to include this view in interfaces. Finally, the model of interfaces in which the interface was supposed to mediate the mutual goals of the system and the user was produced. As the 'actors', i.e. the user and the system, seldom actually have these mutual goals, this model proved to be inefficient[3]. Laurel want to argue in favor of a new metaphor where the interface and the interaction would be seen as a theatre, with a stage containing performing actors and an audience watching the performance. What is interesting, but also somewhat contradictory, in the

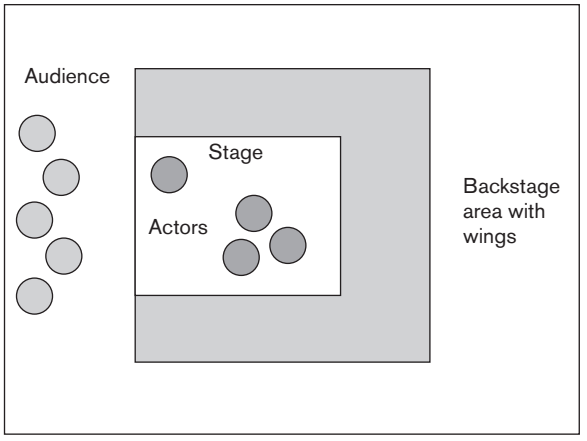

Figure 2.1 A suggestion of a theatre metaphor as a sender – receiver situation (Laurel, 1993)

relation to the above discussions about entertainment is that Laurel also rejects the idea of the computer as a stage with the user as a viewer, or audience only. In the context of computers, the users are active, which makes a one-way communication metaphor impossible. This scenario is shown in Figure 2.1.

The metaphor is developed further, and it is suggested that the audience be put on the stage. The author concludes that this is a confusing situation, because it is not a natural place for an audience to be, either for the audience or for the actors on stage. This is shown in Figure 2.2

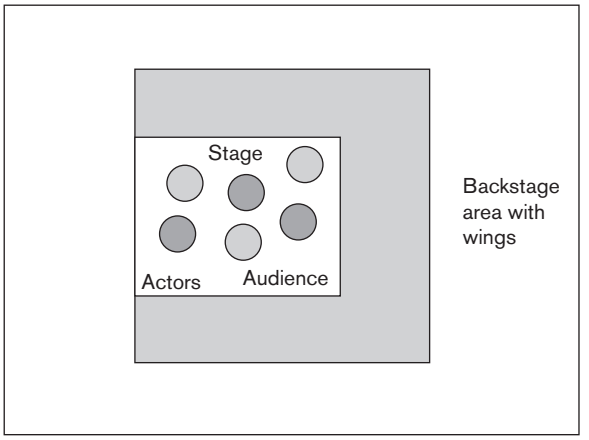

Figure 2.2 A development of the theatre metaphor where the audience is active – on stage. (Laurel, 1993)

Laurel's solution to finding a proper metaphor for this is to view human-computer interaction as a theatrical approach where 'the representation is all there is' (Laurel, 1993). This is shown in Figure 2.3. Stars in the figure represent agents - physical or virtual - circles are physical users.

This may look and sound confusing, but Laurel explains:

> *"In a theatrical view of human-computer activity, the stage is a virtual world. It is populated by agents, both human and computer-generated, and other elements of the representational context (windows, teacups, desktops, or what-have-you). The technical magic that supports the representation, as in the theatre, is behind the scenes. Whether the magic is created by hardware, software, or wetware is of no consequence; its only value is in what it produces on the "stage". In other words, the representation is all there is[..]Think of it as existential WYSIWYG[4]"*
> (Laurel, 1993, p. 17)

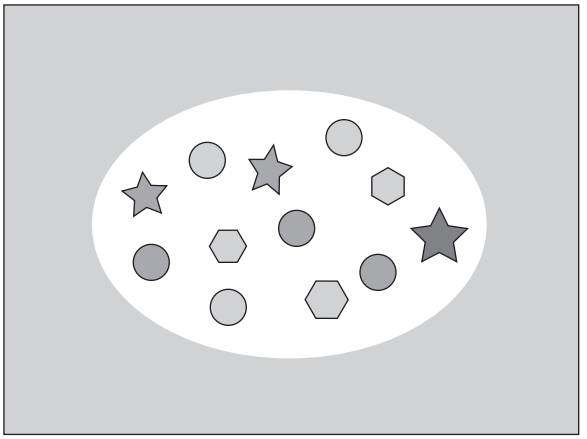

Figure 2.3 A solution to using a theatre metaphor, where the human-computer interaction is seen as a virtual stage, shaped like a spotlight beam, to suggest that all the things that matter in the context are 'illuminated'. (Laurel, 1993)

In the context of entertainment, this theatre metaphor is valuable as an argument against viewing entertainment only as a sender – receiver situation, as would be the case if general definitions of 'entertainment' were used in attempting to understand entertainment web sites. Neither are all users of entertainment IT to be considered as only receivers, nor is the entertainment IT system *per se* to be viewed as the only actor in the use situation. Based on Laurel's models and argumentation, entertainment is what happens in the *interaction.*

The four pleasures

As in the case of entertainment, people might share a general perception of what pleasure is, but again to define it is rather difficult. The new shorter Oxford English Dictionary defines pleasure as:

> *"The condition of consciousness of sensation induced by the enjoyment or anticipation of what is felt or viewed as good or desirable; enjoyment, delight, gratification. The opposite of pain."*

If this definition is developed further and it appears that it can be broken down into components; a subject, who is somehow conscious of sensation; an object which induces enjoyment or anticipation in the subject; the enjoyment or anticipation is viewed as good or desirable – thus implying that the subject has some opinions or feelings. This can be interpreted to mean that it is the judgment of the subject, and not the object itself, that determines whether a thing is pleasurable. If this discussion about pleasure is related to entertainment and fun, the focus moves away from the object, i.e. the system, towards the subject, i.e. the users of the system and their feelings and/or opinions about the system.

The above definition considers 'pleasure' in general, but Jordan (1999) defines pleasure in a product or system as:

> *"The emotional, hedonic and practical benefits associated with products."*

In this definition, 'practical benefits' are related to the user's accomplishment of a task (Jordan, 2000). For instance, it is important for an information retrieval web site to be easily navigable and for information to be searched and found quickly. Emotional benefits relate to a person's mood when using the system. It might be exciting, fun or confidence enhancing to use the system at hand (Jordan, 2000). A game, for instance, might fill the user with joy when it is played. Hedonic benefits, finally, are benefits connected to the sensory and aesthetic side of using a system (Jordan, 2000). A person might experience a product or system as beautiful or an artifact 'nice to handle', such as a joystick, a cup or a PDA or a laptop computer.

This definition provides more input compared to the former given by Webster's, in that it seems that the 'pleasure of products' is not only about opinions and/or feelings, emotional and hedonic benefits, but it is also about more practical benefits, not included in Webster's definition.

The pleasure in products (systems for instance) occurs to some extent in the *relationship between* a product and a user. The pleasurability, i.e the ability of an object to spread pleasure in a subject, Jordan defines further as:

> *"..not simply a property of a product but of the interaction between a product and a person"*(Jordan, 2000)

Again, the relation between the two nodes in the interaction is highlighted. This is interesting as interaction is a prior object of investigation in usability evaluations, which might imply at least an initial potential for these methods to be used in the context of entertainment, fun and pleasure, in order to give feedback about 'pleasurability'.

Tiger (1992) and Jordan (2000) divide the concept of pleasure into four categories:

- Physio-pleasures
- Socio-pleasures
- Phsycho-pleasures
- Ideo-pleasures

All four types are described in more detail below:

Physio-pleasures

This category of pleasure is related to the body and the pleasures derived from the sensory organs. Examples here are the smell of a new car, kissing and touching an artifact, for instance a keyboard, a vase or the handle on a teapot. (Tiger, 1992; Jordan, 2000)

Socio-pleasures

This pleasure derives from relationships between people, as in meetings with loved ones, colleagues and friends. It also includes the relationship with society as a whole. Examples here are a person's image and status. Products in this category would be anything facilitating social interaction, such as a nice sports car at a gas station – it is easy to make contact with people who would like to comment on the car. Products may also indicate the type of person you are, or want to be. For instance, a *Harley Davidson* cap or leather vest might show that the owner rides that brand of motorcycle. This adds a lot to the personal image of the rider. Relationships with products may in this sense become – or at least extend – the owner's social identity. (Tiger, 1992; Jordan, 2000)

Psycho-pleasures

The idea of psycho-pleasure is derived from individual activities. For instance, if a piece of editing software can be used to create and edit images and produces an advanced and attractive result, this would provide a higher level of psycho-pleasure than software which was either too difficult for the user to obtain the same good

result as above or was too simple so that the only thing it can do is display images with no editing possibilities. Another example of psycho-pleasure is when someone uses his or her skills to perform something which may imply emotional satisfaction. (Tiger,1992; Jordan, 2000)

Ideo-pleasures

This type of pleasure refers to people's values, and examples are books, newspapers, films and so on. For instance, a film might be loaded with patriotic undertones, or a book might take a standpoint against a race or a religion. The gender perspective could also come in here. Another aspect of ideo-pleasure is when a product is considered an art form. Any functional product, such as a knife, a water boiler or a car, can be differentiated according to its functional aspect. At the same time the product *in itself* can be seen as art – the water boiler could also be seen as furniture and the knife as a symbol of a high status kitchen. (Tiger, 1992; Jordan, 2000)

The importance of these kinds of pleasures depends on of the purpose behind the evaluations of entertainment IT. In general perhaps the two most critical aspects of pleasure in this context are Psycho-pleasure and Socio-pleasure. The former is important in single-user situations, such as games, exploring features of the web sites etc. The latter is important in activities such as chats, which are a common feature of entertainment web sites. However, in some cases such aspects of entertainment IT artifacts as tactile feedback, might be the main object of study and then Physio-pleasure would be the most relevant aspect. Overall, however, the framework of the four pleasures, gives little or no guidance in endeavors to operationalize entertainment and fun in relation to the evaluation of usability regarding the use of entertainment IT systems.

Emotion and design

Related to the concept of pleasure is that of emotions. Emotions may also be divided into different aspects similar to the division of pleasure into four types, as presented above. Donald Norman, an influencial researcher in the HCI community, presents in a forthcoming book an interesting discussion about emotions in relation to design, mainly of IT artifacts. Three different aspects of design are described, visceral, behavioral and reflective.

(1)Visceral design is concerned with appearances – how things look – an attractive teapot for example.

(2) Behavioral design concerns pleasure and effectiveness of use. If the teapot is nice in that it pours without spilling this is an example of this aspect of design.

(3) Reflective design, finally, considers rationalization and intellectualization. The owner of the teapot can tell a story about how it was bought in Italy. The teapot makes the owner proud. These are two examples of when reflective aspects are present in a design. (Norman, 2002; Norman, Forthcoming).

One possible way to use Norman's theoretical concepts is to link measures of success for use of entertainment IT to the types of designs presented within this framework. This may be done to guide the design to some extent and also to evaluate such technology. An example is shown below of how such a division of various measures of success in entertainment IT systems could fit into Norman's framework as shown in Table 2.2.

Type of design	Description of the type of design	Examples of 'measures of success' for entertainment technology
Visceral design	Appearance	Aesthetic, contemporary, classical or cartoon-like design, etc.
Behavioral design	Pleasure and effectiveness of use	Fun, entertaining, effective, efficient, free from errors, easy to learn, easy to remember, high level of game play, etc.
Reflective design	Rationalization and intellectualization	Extend feeling of, for instance, self-confidence, self-esteem, independence, personal image, etc.

The three types of design originate from Norman, Forthcoming and Norman, 2002. The diagramatic presentation of the links between typical features of these types of designs was made in this study.

Table 2.2 Three types of design.

The implications for use of entertainment IT systems, based on these aspects of emotional design are perhaps that when emotions are involved there are other aspects to consider other than what appears on the screen, i.e. the framework reflects both aspects which are included and those which are excluded in traditional usability. Norman's discussions also highlight the need to extend both the concept of usability and the corresponding evaluation methods. However, overall the theoretical framework presented by Norman is rather too broad to serve as a tool for distinguishing among designs, and thus gives little or no help in solving the problem of how to operationalize entertainment IT use. The measures of success, for instance, might easily be found without using the framework.

The four experience realms

Experience is a broad concept and has both positive and negative associations. The idea of experience plays a role, which could be in terms of users of entertaining IT artifacts experiencing some kind of entertainment. In some way, for instance, when user studies are conducted, what is actually observed is how users or subjects experience the entertainment web sites, and that is also what is reported. This supports the argument that experiences play a big part in the study. However, because experience, as a concept, can be considered to be quite neutral, or positive or negative it might be rather pointless to use it as measurement of success when evaluating entertainment and fun. Fun, entertaining, pleasurable, enjoyable, etc. are all examples of ideas that have a positive value. For this reason they are more closely studied and used in the context of this study. Nevertheless, experiences are still closely related to entertainment and fun, and are therefore still mentioned.
Experience is defined by Websters as:

> *"1 a: the usually conscious perception or understanding of reality or of an event b: the sum total of the conscious events that make up an individual life or the past of a community, nation, or humankind generally 2a: the actual living through an event or series of events [learn by* experience*] b: something that one has actually done or lived through [a soldier's* experiences *in war] 3a: the skill or knowledge gained by actually doing or feeling a thing [a job that requires* experience*] b: the amount or kind of work one has done or the time during which work has been done [a person with five years'* experience*] [Middle French, from Latin* experientia *"act of trying", from* experiri *"to try". "*
> (Webster's New Encyclopedic Dictionary, 1995, p.353)

Pine II & Gilmore (1999) discuss various types of experiences by presenting four realms of *experience*. The four realms are; entertainment, education, escapism and esthetics, as shown in Figure 2.4.
The authors describe the different concepts in the figure as follows:

Entertainment: This type of experience occurs when people passively absorb the experience through their senses, e.g. when listening to music.

Education: Also a type of experience where the user absorbs the experience. However, education is more of a participatory type of activity.

Escape: This realm involves much greater immersion than entertainment or education experience, and this type is often a popular opposite of the above-mentioned pair of experiences. Games, theme parks, virtual chat rooms and a room for paintball are all examples of the escape type of environments.

Estheticism; Here, the users immerse themselves in the experience. However,

they have little or no effect on the experience itself. Art galleries are a good example of such experiences.

In relation to the work conducted and discussed in this thesis, it is worth highlighting the definition of *entertainment* by Pine II and Gilmore, as something where the audience is non-participatory and non-immersed. Possibly this cannot be fully applied in the context of computer-based entertainment, as defined here, as the majority of the entertainment dealt with in this context is *interactive entertainment,* where at least some of the situations, such as games, must be considered immersive. Despite this somewhat confusing difference in the view of entertainment, the framework of the 'Experience Realms' might still be fruitful in the context of this thesis. With regard to the main object of empirical study in this thesis – entertainment web sites – the framework functions as a base for categorizing different types of features within the web sites, as appears below.

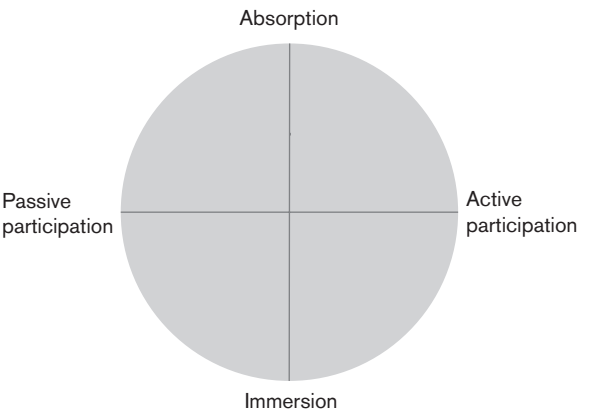

Figure 2.4 The Experience Realms (Pine II & Gilmore, 1999)

The concept of flow

Flow, as a related concept when discussing experiences was introduced by Csíjszentmihályi (1990). It is defined as:

> *"the state in which people are so intensely involved in an activity that nothing else seems to matter; the experience itself is so enjoyable that people will do it even at great cost, for the sheer sake of doing it."*
> (Csíjszentmihályi, 1990).

When we are in the state of flow, according Csíjszentmihályi, (1990):

> *"we feel "in control of our actions, masters of our own fate…we feel a sense of exhiliration, a deep sense of enjoyment."*
> (Csíjszentmihályi, 1990).

These are two examples among many of definitions of flow. However, even if definitions and descriptions of flow differ slightly in the literature, they are fairly similar. Simplified, flow is a combination of perfect conditions, where a balance between challenge and skill manifests itself in a situation where a person is

performing an activity. This is described in Figure 2.5.

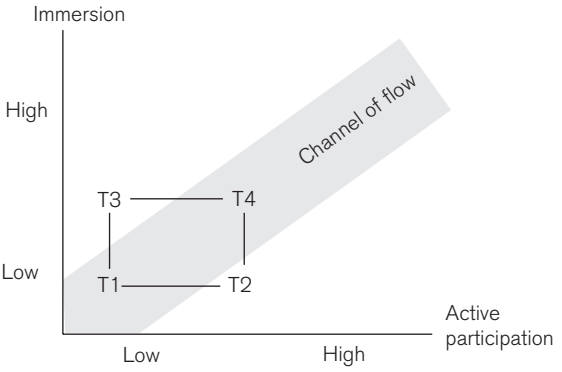

Figure 2.5 A model of the relation between challenge and skill for flow experiences.(T1-T4 = different times) (Csíjszentmihályi, 1990).

The general model above describes how the level of a person's skill and the challenge can influence the flow experience in an activity at four different points in time (T1-T4). For instance, consider the situation when a person starts to play tennis. At the beginning the level of skill and challenge is low – the person just has to get the ball over the net. The relation between skill and challenge is balanced, and the person is in a flow state of mind. However, this is not a stable situation, as two things will eventually happen. (1) The skill of the player will increase, and if the challenge does not also increase, the player will be bored (T2). (2) Another situation is that the challenge will increase too quickly in relation to the player's skill if, for instance, the opponent is more skilful at the game. Then, the player will become anxious. In order to get back to a flow state the player can either (1) search for greater challenges, or (2) practice to achieve greater skill. If the player is successful in this s/he will once again be in a flow state of mind (T4). It is also worth mentioning that the levels of skill and challenge in this model are the perceived levels of the player. (Csíjszentmihályi, 1990).

The notion of flow highlights important aspects of human motivation in general and motivation in relation to entertainment was mentioned earlier in this chapter as being 'for itself' (Langer, 1977). Csíjszentmihályi outlines the concept of "autotelic experience" which is similar to the one Langer describes (Csíjszentmihályi, 1990, p.67ff). The author describes it as:

> *"The term "autotelic" derives from two Greek words, auto meaning self, and telos meaning goal, It refers to a self-contained activity, one that is done not with the expectation of some future benefit, but simply because the doing itself is the reward"*
> (Csíjszentmihályi, 1990, p.67)

In the light of Csíjszentmihályi's description of autotelic experience, which seems to correlate to some extent with entertainment and motivation, it might be argued that the notion of flow is relevant for exploring the use of entertainment IT. However, the main reason the related theories presented in this chapter are explored in this thesis is to see to what extent they could serve as a guide in operationalizing entertainment and fun in the context of IT use. The idea of

flow does not offer a clear solution and was therefore not explored further in the context of this thesis. Nevertheless, it is interesting and may well be investigated and used in research into IT use.

The overall theories and definitions mentioned above emphasize distinctions important for an understanding of fun and entertainment on a general level, but they are not necessarily applicable to entertainment and fun in the context of web usability. Rather than providing practically applicable frameworks that could be used in the operationalization of fun and entertainment in the context of web usability, these definitions and theories provide guidance on other levels of abstraction. These levels of abstraction can also, in fact, be of major importance in this context, and the understanding gained from many of the theories and definitions was drawn on in the proposed methodology for evaluating fun and entertainment in the context of web usability. However, as the theories provided no or little insights into how to *practically* operationalize fun and entertainment in the context of web usability, another way to achieve this had to be found.

The solution in this thesis was to circumvent the problem through a detalied analysis and conceptualization of one empirical phenomenon, which includes fun and entertainment aspects on the web, i.e. entertainment web sites (EWSs). This was done in four steps: (1) The concept of EWSs was clearly defined and categorized in order to show which web sites are included in the concept of EWSs and which are excluded. (2) In addition, the EWS as an entity was further developed and characteristic aspects were identified. (3) On the basis of the knowledge and awareness of these aspects of EWSs, a conceptual framework for how to evaluate an EWS was developed. (4) This framework could then be used in the operationalization of fun and entertainment in evaluation of each EWS. These four steps are further developed below.

Categorizing web sites in general

Before categorizing entertainment web sites, it might be fruitful to demonstrate how web content can be generally categorized. In the paper *Designing Information-Abundant Web sites: Issues and Recommendations,* Shneiderman gives four examples of bases for categorization of web sites (Shneiderman, 1997):

- *By originator's identity.* The originator can be an individual, a group, a university, a corporation, a non-profit organization or a government agency.
- *By the number of web pages in the site.* The number of web pages in a site can vary from one to millions. A similar way is to look at the amount of information on the site.

- *By the goals of the originators, as interpreted by the designers.* Here the spectrum is wide ranging from a personal file with chaotically structured information to impressive annual reports from organizations. As commercial sites start to grow, elegant product catalogs and lively newsletters will also become the norm. Web-zines - magazines on the web, digital libraries and much more, all bring with them different kinds of criteria, as well as special usability needs.
- *By measure of success.* For individuals, the measure of success for an on-line resumé may be getting a job or making a friend. For many corporate sites, the number of visits measures the publicity. For others, the value lies in the number of articles sold from the site. Other measures of success are diversity in hits or hours spent on site.

All these examples of bases for categorization can, of course, also be applied to entertainment web sites. These web sites have different kinds of *originators*, such as corporations, individuals, groups etc. These originators have different *kinds of goals*, such as gaming, promotion of products or services, edutainment etc. The *number of web pages* in entertainment web sites also differs. Some are considered small, e.g. support web sites for events; others are very large, e.g. the big corporate web sites in the music, television or movie industries. As most of these corporate web sites are loaded with entertainment features and content they must be regarded as entertainment web sites. Finally, *measure of success* could also be considered a valid basis for categorizing entertainment web sites. Number of visits, length of stay, downloaded items or articles sold are examples of measures of success which could be used on many entertainment web sites.

When categorizing or defining any phenomenon, the use of classifications of the kind described above is rather problematic, as it is difficult to know if and when the list is complete or who or what guarantees the significance of the list presented. It is also important to consider the abstraction level, i.e. whether such a list is too specific and thus considered less flexible for use in categorizations.

The web continuously changes its nature – new actors, new types of web sites, and new needs for categories such as those above emerge all the time. This gives rise to a situation where such lists are rather difficult to design. Even if Shneiderman did a good job with the list above, it is important to be aware of possible problems that may occur when using such a list as a basis for categorization. On the other hand, in some situations such lists may be applicable, in order to get to grips with a phenomenon, as in this case web sites. However, it is important to always keep in mind the above discussed issues linked with such categorizing lists, when using them.

Categorizing entertainment web sites

In order to be able to operationalize fun and entertainment in the context of web usability through evaluation of EWSs, it is necessary to define and categorize EWSs in the study. Various approaches to this were employed. The bases for categorization of web sites in general as presented by Shneiderman were further explored in order to differentiate EWSs from other web sites. Some of these bases for categorization were also combined with other dimensions of web sites in order to further develop the notion of EWSs. Here, EWSs were examined on the basis of the concepts of *form* and *content*, with the intention of pinpointing important aspects for the evaluation of fun and entertainment in relation to web usability on EWSs. It might be somewhat problematic to divide EWSs into form and content, as it may not be completely clear what to include in the two concepts. For this reason, EWSs were further explored in relation to this problem. The general aim is to present a definition and categorization of EWSs that is clear and understandable as possible.

Firstly, the idea of EWSs is explored on the basis of Shneiderman's categorization list. One of the bases for categorization – the number of web pages included in the ewb site – is of less relevance in the context of defining which web sites are to be considered EWSs.

Originator's identity

In order to be considered an EWS, it may be asked whether a web site do have to be delivered by an originator who is usually understood to be a traditional entertainment provider. Another question is whether all web sites originating from traditional entertainment providers to be regarded as EWSs. These questions have to be answered in relation to this base for categorization.

To answer the first question: All types of originators can deliver EWSs. Whether the web site should be regarded as an EWS has more to do with the framing of the message than the message itself. Edutainment is a good example of this, where content providers, i.e. originators, frame a course or other educational material in a game-like environment with the objective of making it more fun to learn the material. Another example is when providers of products or services, not necessarily entertaining in themselves, present these products in an entertaining way, for different reasons, e.g. selling or marketing. When this is done on the web the web sites should be regarded as EWSs.

With reference to the second question: The originators traditionally seen as entertainment providers do not deliver EWSs only, but also other types of web sites. It might seem strange in this context, but in fact there are numerous other types

of web sites than EWSs delivered by such originators. For instance, entertainment providers often like to be visible on the web but not necessarily framed in an entertaining way. Those web sites are frequently, so-called, information retrieval web sites, where the originators provide information about themselves and their products, for the visitors to search and navigate, but without having any further intention of entertaining the visitor. In some sense it might be seen as *entertaining* in that the information might be considered as such by some people. However, as the activity supported is mainly an information search, it can also be regarded as an information retrieval web site when these web sites are designed and evaluated. It is worth mentioning, however, that what is *evaluated* is the information retrieval process, and not whether or not the entertainment content (information) is fun.

Measure of success

This basis for categorization, as presented by Shneiderman is also relevant in the context of EWSs. In fact, this aspect is of great importance in relation to the operationalization of *any* type of web site. In order to be clear about the required measure of success for the EWS, the evaluation of entertainment in relation to usability is almost designed. Whenever a system is designed it is important to include measures of success in the overall purpose of the system, and this is also true for the design of EWSs. The problem, however, is that it is somewhat more complicated than that. The choice of measure of success is critical, since a stated measure of success like 'entertaining' provides no guidance for measurement of the EWS. More specific measures are needed which are also, if possible, connected to specific parts, features or aspects of the EWS. For instance, an EWS that includes small stand-alone games, downloadable items and traditional information about the theme cannot have 'high playability' as an overall measure of success. Such a measure can only be used for the game included in the EWSs. The concept of measure of success is relevant in operationalization of entertainment and fun in relation to web usability, but it does not completely solve the problem of operationalization of fun and entertainment in the context of web usability in EWSs.

Before proceeding with Shneiderman's last basis for categorization, some central concepts need to be clarified if we are to benefit fully from if with reference to understanding the concept of EWSs. The above discussion attempts to categorize EWSs and excludes what is not covered in this concept. It also reveals high-level goals, such as the originators' identity as well as measure of success. To some extent the above can be seen as aspects of the *process* of design. However, these discussions do not necessarily cover the *product* in itself, i.e. the included attributes, qualities and characteristics. The concepts discussed below are *form* and *content*.

These will be presented, explored and further developed before being combined into a framework together with the remaining basis of categorization – *goal of the originator as interpreted by designers*.

Form and content of EWSs

In order to be able to discuss the product perspective of web sites in general and entertainment web sites more specifically, the framework of *content* and *form* might be of value. In the past, this has been an important framework when discussing various types of artifacts (c.f. McLuhan, 1994; Langer, 1977). In order to highlight important aspects, distinctions, or perhaps difficulties in making distinctions by using this framework in the context of web sites, a broad distinction between content and form might be: *content* answers the question *what,* as in what is the main message of the web site; *form*, on the other hand, describes *how* the message is being delivered[5]. One example is a news web site, where the content includes all the news delivered and the form how the news is visualized and delivered. The former includes the information and the latter, the structure, navigation, graphic layout etc. So far, content and form seems a reasonable distinction to make regarding attributes on a web site. In order to similarly categorize all web sites so as to pinpoint EWSs, a matrix is presented in Figure 2.6. The two-by-two matrix includes the two dimensions – form and content – where content is subdivided into two main categories, entertainment and other, and form into traditional form and high standard form. Based on these dimensions and categories an overview of web sites might appear as follows: there are three possible alternatives which could be regarded as EWSs. The only one not included in these is the alternative where content is other than entertainment and that the form is regarded as traditional form, i.e. traditional structure, navigation and graphical layout. This category of web sites could for instance include traditional IR (information retrieval) web sites with non-entertaining type of content – corporate, organizational, products or other. This gives a situation where the rest of the options are possible to categorize as EWSs. However, it will be shown that this is not the case.

As the purpose in this thesis is seen to be to highlight those web sites where traditional usability evaluation methods are insufficient in one way or another, there is another square which could also be excluded from this problem area. This is the square in which the content is entertainment and the form is traditional. Examples of this combination for instance are eonline.com or IMDB.com, both of them IR *about* entertainment such as movies, TV and celebrities. These web sites are nothing other than IR web sites and can be evaluated as such (functional). Simply because the content is entertainment it does not change the requirements for evaluation methods. This gives a two-by-two matrix shown in Figure 2.6.

The problem, however, with ideas of form and content in the context of EWSs is that it might sometimes be difficult to clearly distinguish the two. The EWS of the *Eurovision Song Contest*[6](ESC), for instance, – a support web site for the annual event – can be used as an example of a web site where it might be difficult to distinguish between form and content. The content, or message of the ESC web site is the contest itself and information about it. But, the web site also includes games, downloads and features where, for instance, the user is given the chance to mix his or her own re-mix of the song 'Waterloo'[7]. Furthermore, on the night the event took place it was possible to follow the contest on the web with additional camera views from back-stage and other places. Overall, the graphic design of the web site could be regarded as high standard quality. Where should the boundary between form and content be drawn in this example? How should the downloadables, games, re-mix features and the extended camera views be seen – as form or content?

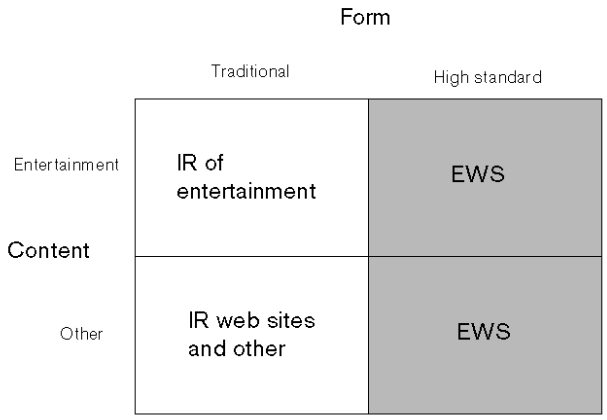

Figure 2.6 Web sites in dimensions of form and content

The difficult or impossible situation of separating form from content has, for instance, been discussed in the context of art where Langer (1977) discuss the relationship between the two as:

> *"Our scientific convention of abreacting mathematical forms, which do not involve quality, and fitting them to experience, always makes qualitative factors "content"; and as scientific conventions rule our academic thinking, it has usually been taken for granted that in understanding art, too, one should think of form as opposed to qualitative "content". But on this uncritical assumption the whole conception of form and content comes to grief, and analysis ends in the confused assertion that art is "formed content"; form and content are one.*
> (Langer, 1977)

Langer (1977) discusses *form* and *content* in relation to art, which is somewhat different to design. In most cases, the only purpose of art is the intrinsic motivation of the artist, i.e. the only purpose of the art is the art *itself* – no other external purpose exists. Designers, on the other hand, have some kind of customer or client, who has a purpose for ordering the artifact to be designed (Nelson &

Stolterman, 2003). In the case of design, the idea of content, therefore, can quite naturally be linked to the intention or purpose behind the artifact, and what surrounds that can be called form. However, this is still not very satisfactory in the context of evaluation of EWSs, as it would be fruitful to have a more specific theoretical framework for describing the different entities included in the web site in order to gain an awareness of what can possibly be measured and also how this can be done.

If the idea of content answers the question 'what' – what is the message of the EWS – and the form includes the answer to the question 'how' – how is the message presented – a third question can also be added – 'by/through what' – by or through what is the message presented in the way it is. As will be argued below, this third questions also relates to the idea of form in this study.

Another important aspect to discuss in the context of this thesis is to what extent the designers can change the entities of EWSs as the design is the process the evaluations are supporting. The content, as defined above, is seldom or never under the control of designers of EWSs. Instead this is given in advance. So what is evaluated primarily is the form of the EWSs. This could be seen as important when separating form and content in EWS. Content is the part of the EWS the designers do not control and form is whatever is under their control. In this way the *division* of the two concepts is facilitated to some extent. However, in operationalization of entertainment and fun in the context of web usability, this facilitated division of form and function provides little help. As discussed above, form includes a number of aspects of the EWS, such as graphic form, navigation, structure and different kinds of 'added value'. It is reasonable to believe that these aspects differ considerably regarding how they are measured in relation to fun and entertainment. Therefore, a more specific framework is needed. Based on aspects included above, the model of an EWSs was developed, where the EWS includes aspects of content, added value, structure and graphic form.

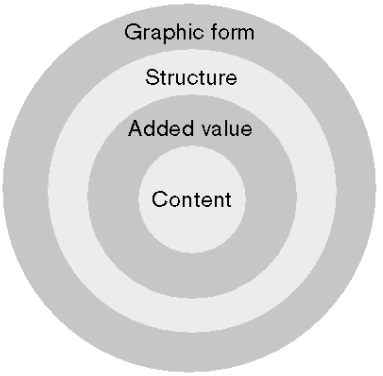

Figure 2.7 Included entities in entertainment web sites

In order to describe the included entities in the model, an entertainment web site is used as an example, i.e. the web site of Eurovision Song Contest. In the case of ESC the three entities can be identified as:

- *Content* – the event of ESC in itself as well as information about it.
- *Form* – the add-on values, features such as the Waterloo re-mix, postcards, downloads etc., the graphic layout, structure and navigation of the web site.

Goals of the originator, as interpreted by designers

This basis for categorization originating from Shneiderman, is divided into 'goals' and 'dimension of web site'. The former are subdivided into two types of goals, which could be entertainment or non-entertainment. The latter, i.e. the dimension of the web site, covers which of the dimensions form and content, in which the goals are to be included. This gives four possible options, which are visualized in Figure 2.8.

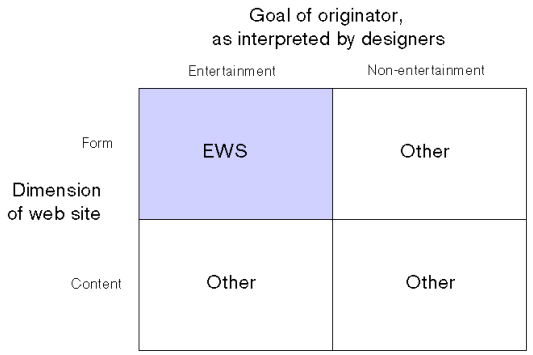

Figure 2.8: Web sites regarded in dimensions of form vs. content and goal of originator, as interpreted by designers.

Only where the goal of the originator was to entertain, and where this was to be regarded as constituting form could the web site be regarded as an EWS. The rationale for this is that only in the case where the entertainment was a goal of the originator, interpreted by the designers, could the results from evaluation have any influence. All other aspects of entertainment, jokes for instance, were excluded in this study.

Operationalization of entertainment and fun

All the frameworks presented above provide guidelines for how to view EWSs. It might, however, be argued that they do not categorize EWSs sufficiently to provide any guidance in how to evaluate different types of EWSs, as in this thesis.

It might perhaps be possible or fruitful to make narrower distinctions and categories among EWSs for example presenting listings of types of EWSs, such as 'game web sites', 'event web sites', 'advertisement web sites' etc. A methodology based on this type of thinking can be presented as in Figure 2.9.

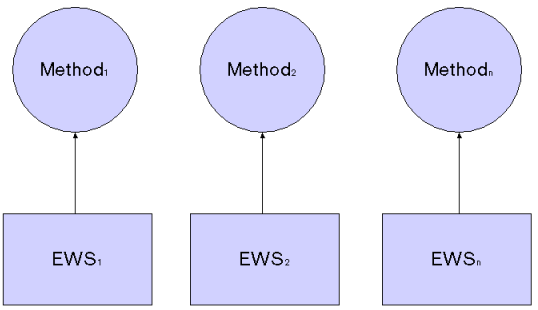

Figure 2.9 A possible methodology for evaluating entertainment web sites.

In the above case, one method would be suitable for each type of EWS, i.e. there would be one specific list of heuristics or one standard approach for how to conduct user tests for each type of EWS. This is not the approach chosen in this study, and there are two main reasons: (1) the rapid development of this type of web site, i.e. as new types of EWSs are released all the time such a list most would certainly very soon be out of date; (2) many EWSs include a large number of features, creating a situation where a specific EWS would probably be difficult to fit into just one of the categories defined. However, these problems can be solved methodologically if a *flexible*

approach is developed, where a number of possible alternatives are given together with guidelines for how to handle each situation. This type of methodological approach, which is also the approach chosen in this thesis, can be visualized as in Figure 2.10. This broad and flexible approach is necessary, as the very nature of EWSs makes it difficult or even impossible to place EWSs in boxes or categories. This approach will be discussed later in further detail.

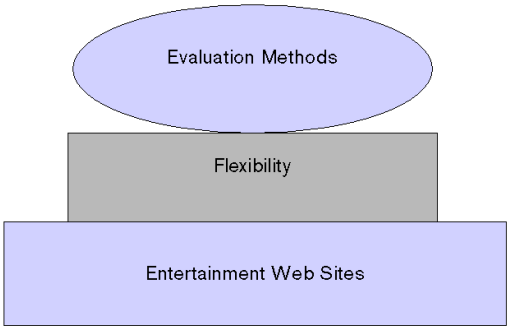

Figure 2.10 A flexible methodology for evaluating entertainment web sites.

Features of entertainment web sites

In this study a conceptual framework was necessary in order to describe the chosen EWSs, mainly to visualize the rationale behind the choices of web sites to be included in the study. This was done by identifying some common features, often found in entertainment web sites[8]. These features were then positioned in different types of conceptual frameworks in order to reach a conceptual view of features included in the selected EWSs.

Features commonly found in entertainment web sites are presented below:

1. *Entertainment information* – information about the theme of the web site, jokes etc.
2. *Downloadable items* – screensavers, pictures etc.
3. *Small 'stand-alone' games* – 'Memory' or such.
4. *Other features dependent on plug-in technology* – Re-mixing of music etc.
5. *High quality graphic design*
6. *Edutainment content*
7. *Communication with others* – chats, virtual meeting rooms etc.

These features were placed in the framework of the *Experience Realms* in order to see what type of experience might be typical of the specific feature, as shown in Figure 2.11.

Entertainment, as viewed in the context of entertainment web sites, can include all aspects in the *Experience Realms* framework – education, escapist, esthetics and entertainment (as defined by the Pine II & Gilmore). However, in most cases a web

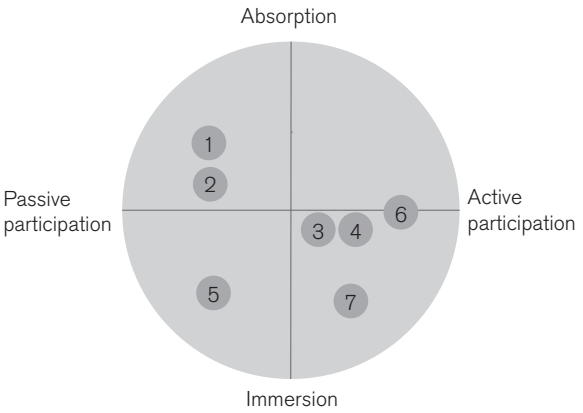

Figure 2.11 Positioning of typical features of entertainment web sites in the model of the Experience Realms by Pine II and Gilmore (1999). The numbers in the model correspond to the numbers of the features listed above.

site cannot be considered particularly entertaining if it *only* contains the experience education or *only* escapist. It also has to involve some kind of amusement, fun or enjoyment. To be esthetic alone is not necessarily to be entertainment, but combined with other aspects it might enhance the experience of fun at an entertainment web site.

Discussion

This chapter included some of the possible alternative theories that could be used when operationalizing entertainment and fun in the context of web usability. Some of them show potential to be useful, however, unfortunately they provided little or no guidance for how this could be done. Nevertheless, many of the theories highlighted the situated, subjective nature of these qualities, which it is important to be aware of in evaluations of this kind. This in itself may be a reason why no general theory is able to deliver specific guidance in such an operationalization. Instead of a theoretically based operationalization of fun and entertainment in general, a specific type of technology where fun and entertainment are included was chosen, i.e. entertainment web sites, in order to advance further. By defining, categorizing and analyzing entertainment web sites, partly based on the theories discussed earlier in this chapter, an understanding was reached. This understanding is used in study design and analysis throughout the study, as will be shown. How this was done is presented in the Chapter 4, which covers the overall strategy and structure. However, first, the concept of 'methods' is discussed and this is included in Chapter 3.

Footnotes

[1] The American Heritage® Dictionary of the English Language, Fourth Edition Copyright © 2000 by Houghton Mifflin Company. Published by Houghton Mifflin Company.
[2] Roget's Interactive Thesaurus, First Edition (v 1.0.0) Copyright © 2003 by Lexico Publishing Group, LLC. (http://thesaurus.reference.com/)
[3] These models of the interface are not further investigated in the context of this thesis. For further descriptions of these views, see Laurel, 1993, (pp.12-15).
[4] WYSIWYG is an acronym for What You See Is What You Get. It was coined by Warren Teitelman at Xerox PARC and has become an important paradigm for direct-manipulation interfaces. (Laurel, 1993, p.17)
[5] Note, however, that this straightforward distinction will be further discussed below and that it should not be seen as final or complete.
[6] For further description and visualization of the web site of Eurovision Song Contest see Chapter 5.
[7] Which was the winning song of ESC'74 performed by the Swedish group ABBA.
[8] This list of features was developed in collaboration with designers of entertainment web sites at Paregos Mediadesign AB.

Chapter 3

Usability evaluation methods as objects of study

When evaluating usability a large number of evaluation methods are available, the majority of which originates from the field of Usability Engineering. The choice of method in each case is crucial and depends very much on the purpose of the particular evaluation. Different methods also cover different aspects of usability and are moreover based on various kinds of *rationale* such as whether the method is process- or product-oriented. Process oriented evaluation methods describe the steps to take and tasks to complete in the process of evaluation. The underlying assumption behind these methods is that a proper evaluation can be achieved by following a pre-specified procedure. In product-oriented evaluation methods, the focus is on normative information about the *product* to be evaluated by the provision, for instance, of checklists or guidelines for how to devise a usable system.

When developing, or re-designing, standardized usability methods it is important to consider the rationale behind the particular method in order to specify the proper judgments to be used in the process of development or re-design. It is also important to consider applicable criteria, measures, and indicators of success on which to base the judgments. Finally, when developing or re-designing methods, it is crucial to design the strategy of the process of re-designing or developing the specific method. In usability evaluation, methods are regarded and used as tools. When developing or re-designing methods their role is different, they becomes the object of study. This different status of methods has important consequences for the strategy of evaluation and re-design of methods. This chapter deals with how to develop a strategy for re-designing traditional usability evaluation methods in the context of entertainment and fun. More specifically, it highlights aspects of how to consider methods as the object of study. It focuses on how methods should be

regarded in general, the context in which methods are employed and the rationale behind the methods. First, this chapter investigates some definitions of methods in general as they appear in dictionaries, to see how methods are described. The contexts in which usability evaluation methods are primarily used are presented, and the role of such methods in each context is discussed. In the next section of the chapter the process- vs. product-oriented rationale of usability evaluation methods is explored, as this is an important consideration when re-designing these methods. Finally, the chapter deals with some related work in the context of HCI regarding exploration and development of methods. The chapter concludes with some overall challenges regarding methods as objects of study, which is the approach to methods in this thesis. This forms an introduction to the next chapter, which deals with the overall strategy for re-designing traditional usability evaluation methods in the context of entertainment and fun in this study.

Methods defined

In order to gain knowledge about usability evaluation methods and how these should be regarded as objects of study, it is important to pinpoint what methods are and how they could be described. As a first step towards this understanding, dictionaries were consulted for definitions of method as a concept.

'Method' originate from the Greek word *Methodos (meta-* + *hodos* way – more at). Method is also described in Webster's dictionary as:

> *"a procedure or process for attaining an object"*
> (Webster's third new dictionary, 1422)

Since this definition is very general, it could be applied to most methods, including evaluation methods. It highlights the fact that 'method' traditionally conveys a *process* of some kind. Webster's dictionary further describes a method as follows:

> *"Method can apply to any plan or procedure but usually implies an orderly, logical, effective plan or procedure, connoting also regularity"*
> (Webster's third new dictionary, 1423)

This more detailed definition implies a general view of 'method' as containing a high level of structure within the process, with little freedom of action for the person using the method. This is probably the most widely accepted view of methods in the context of HCI and Usability Engineering. As will be shown

below, however, the latter part of Webster's definition does not fully apply to all usability evaluation methods. Other definitions of method are related to methods used in research. In this field the notion of method is central since research method is often the basis for obtaining reliable and valid results. In the context of academia, Webster's describes methods as:

> *"a systematic procedure, technique, or mode of inquiry employed by or proper to a particular science, art or discipline"*
> (Webster's third new dictionary, 1423)

As this description considers a more *specific* type of method, it also provides a more detailed and clearer view of method as a concept. As this thesis deals with various types of methods, of which the overall research method in the thesis itself is one, it is important to be aware of how academic methods in general are described. Overall, three types of methods can be found in the context of this thesis: (1) usability evaluation methods, i.e. the object of study in the thesis, (2) method for inquiring and developing the object of study, and finally (3) the overall research method used in the thesis. In any discussions concerning method in a specific case in the text it is essential to know what kind of method is being referred to. Hopefully, this is made clear throughout the thesis.

The next section focuses more specifically on usability evaluation methods. The fact that the evaluation methods originating from the field of Usability Engineering are used to evaluate usability in *research* and in *practice*, might be found somewhat confusing. Because of this, these methods and how they are used in research as well as in practice are discussed in more detail below.

Usability Engineering methods in research and practice

Usability Engineering is a field employing a number of usability evaluation methods (c.f. Nielsen, 1993), as mentioned in Chapter 1. The field of Usability Engineering can be regarded both as a research field and as a collection of methods used in practice, i.e. systems development. It is important to highlight the different types of contexts where these methods are usually put into practice and developed. In general, the methods in the field were developed through extensive research efforts, i.e. research *about* usability evaluation methods. Further, these methods are used as a part of a research strategy, i.e. research *with* usability evaluation methods. For instance, in the HCI research discipline, a research procedure may be presented as follows:

"Isn't it standard best practice in HCI to interview users to understand their needs, develop a system to meet these needs, and evaluate the system to determine whether it meets their needs?"
(Whittaker et al., 2001)

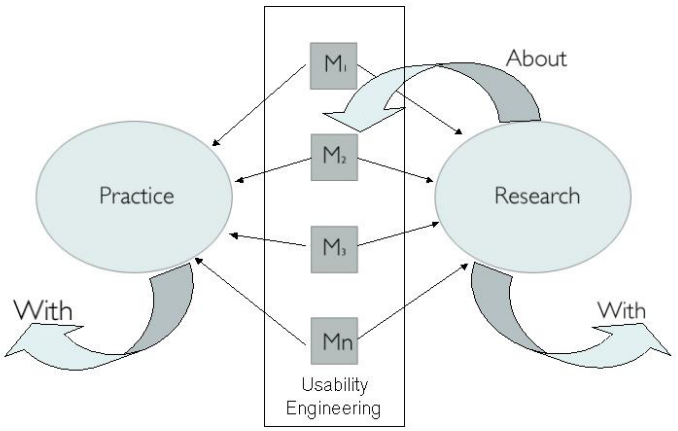

Figure 3.1 Relation between the usability evaluation methods within Usability Engineering, research and practice.

Finally, the methods included in the field of Usability Engineering are also commonly used in practice when evaluating system usability, i.e. practice *with* usability evaluation methods. In the two cases of research and practice *with* evaluation methods they can be seen as being used primarily as a tool. In this thesis the research process should be understood as research *about* methods, where the role of the evaluation methods is the object of study. The relation between usability evaluation methods within Usability Engineering, practice and research is shown in Figure 3.1.

Process- vs. product-oriented methods

Methods are based on some kind of *rationale*, as mentioned earlier. This rationale can be viewed from a number of angles and includes a number of aspects. One such angle is to consider the extent to which specific methods are process- or product-oriented.

When developing methods it is important to ask whether the rationale concerns the process, i.e. the performance of using the method, or the product, i.e. where the focus in on the final product. A process-oriented method is described by suggestions and rules for how the process should be designed. Here, a basic standpoint is, that if we correctly design and specify the process, the product will maintain a proper standard. Many of the existing methods in information systems design are based upon the process-oriented approach. The International Standard Organization (ISO) also uses this type of approach in its programs 'ISO 9000' and 'ISO 14000', i.e. by specifying (and standardizing) the process, high product quality will be achieved. In the words of ISO:

> *"Both ISO 9000 and ISO 14000 concern the way an organization goes about its work, and not directly the result of this work. In other words, they both concern processes, and not products - at least, not directly. Nevertheless, the way in which the organization manages its processes is obviously going to affect its final product."*
> (ISO.org web site[1])

In a product-oriented method, on the other hand, the purpose of the method is to specify the right requirements and standards for the *product* – this guarantees the result. Architecture is an example of a discipline where the product-oriented approach to methods is commonly used. An architect learns how the product, a building for instance, should be constructed in order to be of a high quality, but the process by means of which this is to be achieved is not specified. It is even believed that specifying the process rather than product can impose excessive constraints of the designer (in this case, an architect) and, therefore, be an obstacle to designer's creativity (Stolterman, 1991).

The methods used in this thesis, as well as in usability research in general, include both process- and product-orientation methods. However, in the case of evaluation methods, it important to first decide what the 'product' is – is it the product that is evaluated or the product of the design process, i.e. the IT artifact, or is it the product of the evaluation process, i.e. an account of usability problems. A further aspect to consider is what should be defined as process-oriented. In the examples described above the processes were described in detail, i.e. they seemed to control and structure nature of the method. In the context of Usability Engineering there are evaluation methods that are considered to be quite 'free', such as Design Walkthrough[2]. Should this then be regarded as a product-oriented method, as the product is not described in this method, not in either of the two senses of 'product' given above, i.e. the IT artifact or the evaluation process. This makes deciding whether the approach is a process- or a product-oriented method problematic.

Regarding the kind of product in question, there are two aspects to be considered: (1) Categorization of evaluation methods that makes more sense in this context (2) to what extent is evaluation considered a part of a greater whole, i.e. a design process. In very few examples, are the requirements or structure of evaluation reports *per se* specified in definitions and descriptions of usability evaluation methods. Instead, dependent of each situation, reports are designed on the basis of requirements of receiver of the report and the specific purpose of evaluations. This leads to the conclusion that it makes more sense to regard the IT artifact as the product. With

reference to the second aspect, whether the usability evaluation should be regarded as a part of the design process, opinions may differ and every situation is unique, but often a usability evaluation *is* a part of a design or a research process. This is another argument for seeing the IT artifact as a product of the evaluation process.

Considering a loosely structured evaluation process as a process-oriented method can be less problematic. Even if Design Walkthrough, for instance, does not include a list of procedures to follow, the process is at least to some extent described. For instance, the developers of the method give a recommended number of evaluators to be present in each evaluation session, guidelines for a particular evaluation, i.e. that the evaluator should be open-minded, etc. The fact that the focus is on process rather than product implies that such evaluation methods should be regarded as process-oriented.

One example of product oriented usability evaluation methods used in the study in this thesis is Heuristic Evaluation. No description of the evaluation process *per se* is given in the presentation of the method. What is provided, instead, is a list of guidelines, or heuristics, concerning the properties a correctly designed system in general should have. This is a typical example of a product-oriented method. Table 3.1 presents a number of common usability methods.

Method	Process-oriented	Product-oriented
Think-aloud	Yes	No
Interviews	Yes	No
Structured tasks – experiments	Yes	No
Heuristic Evaluation	No	Yes
Cognitive Walkthrough	Yes	No
Design Walkthrough	Yes	No

Table 3.1 Process- vs. product-oriented usability methods.

The table above shows that the majority of usability evaluation methods in Usability Engineering, at least those described above, are process-oriented. The few product-oriented methods that exist are all parts of inspection methods. One example is Heuristic Evaluation, as already mentioned. Other methods, similar to Heuristic Evaluation but based on other types of listings of design guidelines, can also be considered product-oriented.

Related work

The purpose of this thesis can be considered as in some way to *measure* the applicability of methods and to revise and re-design them on the basis of those measurements. This highlights the crucial question of when a method is to be considered successful.

In relation to this, one has to consider whether the standpoint taken towards methods is process- or product-oriented – is the purpose to judge the process or the product. In the situation where the focus is on the *process*, reference can once more be made to the ISO quality assurance program. The ISO provides checklists with guidelines about different types of processes, which are then used step by step as a basis for inquiry into the evaluated process, in order to ultimately ensure the quality of the results, i.e. systems, products or services, of the evaluated process.

In other words, if the quality of the *product* is in focus, the result, i.e. the product, has to be judged in relation to the initial purpose, and if the product fulfils the purpose a high quality product is obtained.

There might be different ways of conducting inquiries into the methods used in the context of Usability Engineering. One example of related work conducted in the HCI research area, is a framework of guidelines or heuristics for selection or design of usability evaluation methods (Khan & Prail, 1994). This framework highlights some potential measures for success of usability methods, with the main focus on *Inspection Methods*[3]. The included heuristics for judging or evaluating usability evaluation methods are:

1. Have a meaningful number (or ratio) of potential defects been found?
2. Are defects valid (i.e., users would have had problems)?
3. Have quality solutions been found?
4. Do engineers or designers perceive enough value in the method to warrant participation?
5. Do engineers enjoy using the method? (This represents an emotional component.)
6. Do engineers become more effective designers after taking part?

The form of a method can be shaped by a combination of characterized user needs, experience with methods in general and compliance with methodology design heuristics (Kahn & Prail, 1994). In this example of judging evaluation methods, the discussion of the quality of methods is primarily related to concerns of engineers and designers. This would imply that the authors are referring to procedures used in practice. As mentioned earlier in this chapter, since Usability Engineering methods are used both in research and in practice, the above list of heuristics can be seen as also generalizable to the use of methods in research. It

might, however, be worthwhile to explore the possibility of using each heuristic ineither research or practice, since one or two of the aspects listed above might be excluded when considering usability evaluation methods as part of research process. Other heuristics may also be included instead.

Each of the above heuristics can also be considered more or less applicable depending on whether they are used to evaluate process- or product-oriented methods. Some of the heuristics can be considered more related to the view of evaluation methods as process-oriented and some to the view of evaluation methods as product-oriented. In Table 3.2, the six heuristics are matched to these two types of methods.

Heuristic	Process-oriented	Product-oriented
Has a meaningful number (or ratio) of potential defects been found?	No	Yes
Are defects valid (i.e., users would have had problems)?	No	Yes
Have quality solutions been found?	No	Yes
Do engineers or designers perceive enough value in the method to warrant participation?	Yes	Yes
Do engineers enjoy using the method? (This represents an emotional component.)	Yes	No
Do engineers become more effective designers after taking part?	Yes	No

Table 3.2 Process- and product-oriented methods in relation to heuristics developed by Khan and Prail (1994).

As shown in the table, five of the six heuristics are linked to one or the other type of method and one heuristic, "Do engineers or designers perceive enough value in the method to warrant participation?", to both types (. This heuristic is seen as process-oriented in that the value of *participation* in the process is highlighted, and this could be both the participation *per se* as well as the *output* of the participation, i.e. the results of the evaluation.

Other researchers discussing the effectiveness of usability evaluation methods refer to such issues as the number of problems found, the level of severity of problems found, cost-effectiveness in relation to results obtained and finally the type of human resources required (Jeffries et al., 1991)

Another way to consider the suitability of methods is to compare different techniques or methods. Numerous researchers have conducted this type of comparison (Jeffries, Miller et al. 1991; Desurvire, Kondziela et al. 1992; Karat, Campbell et al. 1992; Nielsen & Philips,1993; Olson & Moran, 1998; Gray & Salzman, 1998; Karat et al, 1998). Typical measures of success used in these types

of studies are various resemblances in usability problems found between methods, sometimes called the *impact ratio* (Sawyer et. al., 1996). This way of evaluating methods is based on a product-oriented perspective of methods, i.e. it is the outcome or result of the use of the method that is judged. Some of the above studies comparing methods (i.e. Jeffries, Miller et al. 1991; Desurvire, Kondziela et al. 1992; Karat, Campbell et al. 1992) were the subject of a meta-study, conducted by Muller, Dayton et al. (1993). The authors present five criteria by which to compare usability methods. They are:

- *Raw yield* – The number of unique classes of usability problems found by each method.
- *Raw yield weighted by opportunity* – The raw yield per participant hour, i.e. per hour of opportunity to discover problem found for each method.
- *Refined yield* – The proportion of severe problems found by each method.
- *Benefit-Cost* – The average cost, in terms of total human hours involved, to find each problem, for each method.
- *Uniqueness* – The likeliness of finding problems, undiscovered by other methods, for each method.

The above criteria are comparable with the heuristics by Khan & Prail (1994) in how they can be used in the context of the study in this thesis.

Summary

This chapter presented the concept of methods from a variety of perspectives. It was argued that sometimes the same methods, for instance usability evaluation methods, could be used both in *research* and in *practice*. The difference between research and practice often lies in the purposes of the evaluations. In practice, the evaluation methods are used as analytical tools, whilst in research they can also be an object of study, i.e. the method *per se* is investigated.

When studying methods, it is important to consider whether it is the process or the product of the method that is of interest. Some usability evaluation methods are designed as being product-oriented and others process-oriented. In general, in the context of usability evaluation, process orientation can be regarded as more common. Finally, some other potential conceptual frameworks for use in analysis of data about methods were presented in this chapter, i.e. the heuristic list by Khan & Prail (1994) and the criteria for comparisons of methods by Muller, Dayton et al. (1993).

Discussion

In this study some of the above aspects of methods are applied to contribute to our knowledge of how traditional usability evaluation methods should be re-designed and revised to become more suitable for evaluation of entertainment web sites. The distinction between *methods as tools* and *methods as objects of study* is used in the analysis of the data. This distinction is highly relevant in the context of this thesis, as these two concepts are very easily confused. It is important, in every situation, to be aware of whether the method discussed is to be considered as a tool used to evaluate a product, for instance, an entertainment web site, or whether it should be seen as the object of study.

In addition, all usability evaluation methods included are investigated on the basis of whether they are considered to be process- or product-oriented. This is important for a number of reasons including how they should be judged. The basis for judging process-oriented methods differs from that of judging product-oriented methods. Two sets of general heuristics, initially designed by Prail & Kahn (1999) and Muller et.al, (1993), for judging methods were further described in this chapter. These sets of heuristics can be seen as valuable and some of them are also, to some extent, used in the context of this study. However, some of these heuristics seem to be designed for situations where evaluations are conducted in relation to a design process. They are highly focused on the perceptions, enjoyment and needs of engineers and designers. Since not all of the evaluations in the study in this thesis were conducted in collaboration with designers in an ongoing design process of the web sites, it was difficult to consult designers in order to get judgments of methods and their results, on the basis of the heuristics. However, in some cases it was possible, and the outcome of these judgments is discussed in Chapter 14.

A description of the structure and strategy of the study is given in Chapter 4 "Strategy and structure" (Part 2). The process of refinement and re-design of the evaluation methods is described in chapters comprising Part 3. The theoretical concepts described above are used as a tool for analysis of the results of the overall study in Chapter 12 (Part 5).

This is the end of Part 1 dealing with the underlying theoretical issues. In the next part of the thesis – Part 2 - the overall research strategy is presented and the first phase of the empirical study is reported.

Footnotes

[1] http://www.iso.org/iso/en/iso9000-14000/basics/general/basics_4.html (2003-10-07)

[2] For further description of this evaluation method, see Chapter 1.

[3] For a further description of inspection methods, see Chapter 1.

Part II

Evaluation of entertainment web sites using traditional methods

Part 2 presents the first phase of the study conducted within this thesis. It begins with a discussion of the overall research strategy and structure of the thesis. This discussion is followed by a presentation of a series of studies in which traditional empirical usability evaluation methods were applied to entertainment web sites. Finally, a series of studies employing traditional inspection methods is discussed. Part 2 includes the following chapters:

- *Chapter 4* – Strategy and structure
- *Chapter 5* – Evaluations using traditional empirical usability evaluation methods
- *Chapter 6* – Evaluations using traditional inspection methods – experts
- *Chapter 7* – Evaluations using traditional inspection methods – novices

Chapter 4

Strategy and structure

This chapter presents the main strategy and structure of the study reported in the thesis, including a discussion of the approach used to analyze and assess evaluation methods. The chapter also discusses the sources of empirical evidence and actors involved in the evaluations. An overview of the structure of the study, including its three main phases, is presented and overall methodological considerations are discussed. The chapter includes a discussion of the abductive approach used in the study, the materials used, and the subjects participated in the study. The types of empirical evidence and the process of analysis are also described. The aim of this chapter is to provide an overview of the complete research process presented in the thesis. Details and specific information about the different phases of the study are further developed in the following chapters.

Strategy

The overall strategy in the study could be described as follows: Common usability methods were used in the study for evaluation of entertainment web sites to assess their suitability for elucidating relevant information about EWSs. A special focus, as already mentioned in Chapter 2, was on the *form* of the web sites. The aim of applying traditional usability methods was to establish whether they needed further revision and re-design. The findings of the study indicated that the methods needed to be further revised and re-designed to become more applicable. Therefore, the methods were revised and re-designed accordingly. The re-designed methods were subsequently applied in evaluation of the same, or additional, entertainment web sites, to establish whether the re-design resulted in any differences in applicability. In other words, the aim of the new application was to find out if the changes in the methods resulted in changes in the outcome of the evaluation. The methods

were judged on the basis of the applicability, i.e. to what extent the methods could inform design of EWSs. This was compared to earlier steps in the study. Finally, on the basis of the results from the study, an improved methodology for evaluating entertainment web sites was presented.

Two types of evaluation methods explored

In this study two types of traditional usability evaluation methods were explored, namely, empirical usability evaluation methods and inspection methods. The use of two different types of methods required designing the study so that the most important aspects of each type of methods could be adequately evaluated. To meet this requirement, design of the study had to address a number of questions, such as: What are the crucial aspects of empirical evaluation? For instance, what are the conditions to be investigated using empirical evaluation methods in the context of evaluating EWSs? How should traditional expert evaluation methods, i.e. inspection methods, be judged and developed when evaluating fun in EWSs? What roles do the actors play and what possible sources of empirical evidence can be used?

Method for studying evaluation methods

In Chapter 3, three different processes including usability evaluation methods are discussed, i.e. practice *with* methods, research *with* methods, and finally research *about* methods. In both research and practice *with* usability evaluation methods, the process is similar – an object is studied in order to obtain knowledge about it. In practice this knowledge is mostly used as an input in developing a better product or system. In research the knowledge is usually further analysed in some way to achieve a better understanding of a larger context, that is, in order to obtain generalizable knowledge. Where research *about* methods is conducted, the process is somewhat different: a method is used to study methods, in this case evaluation methods. Findings of that type of research generate knowledge about evaluation methods, which knowledge can provide input into the design of better methods. This can be visualized as in Figure 4.1.

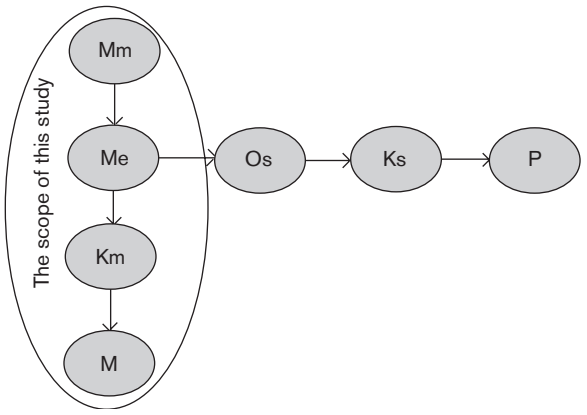

Figure 4.1 Two types of evaluation processes showing the evaluation method. Me= evaluation method, Os= object of study (system), Ks (knowledge about this system), P= a new (and better) product, Mm (method for studying methods), Km = knowledge about the method studied and M= a new (and better) method

This thesis is concerned with research *about* methods. In the HCI research field this type of research is quite frequently published in journals and at conferences (c.f. Jeffries, Miller et al. 1991; Desurvire, Kondziela et al. 1992; Karat, Campbell et al. 1992; Nielsen & Philips, 1993; Olson & Moran, 1998; Gray & Salzman, 1998; Karat et al, 1998). Therefore, it may appear that knowledge about how to analyse, evaluate, and design methods have already been developed over the years in this research discipline. However, the published research in HCI focuses mainly on the *product* of the research, i.e. the developed methods. Little effort is made to discuss or reflect on the *process* of this research. Often, the process of evaluating methods is only briefly mentioned in publications. It is seldom in focus in this type of research. Instead, the emphasis is on describing the advantages and disadvantages of the developed or re-designed methods. As no standard procedure could be found for analysis of usability evaluation methods, it is important to describe how this analysis was done in the context of this study.

When evaluation methods are used only as tools, it is the purposes and goals of the *product* that is the main measure of success, that is, to what extent the product could be said to fulfill its purposes and goals. The method is typically described and explained in that case with the main or sole purpose of showing that the results are valid and reliable. On the other hand, when evaluation methods themselves are the object of study, i.e. when research *about* methods is conducted, other aspects must also be considered, which requires additional work in designing a research strategy. In order to understand methods as 'objects of study' it also seems reasonable to use them as tools, as only through their application they can be fully understood and evaluated.

The research strategy employed in this thesis could be briefly described as follows: First, in order to obtain knowledge about evaluation methods an evaluation of entertainment web sites was conducted. Second, findings of this evaluation, together with evidence about the applicability of methods in each case, gained from the evaluation process *itself,* were used to assess the strengths and limitations of the evaluation methods. Third, this assessment resulted in re-designed methods. Fourth and finally, the whole procedure was then repeated with the re-designed methods used to evaluate new EWSs. The approach used in this study can be described as shown in Figure 4.2.

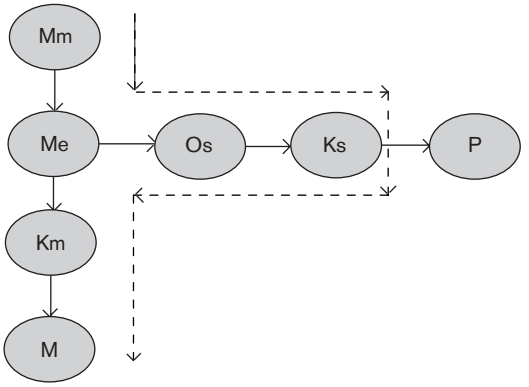

Figure 4.2 The research process in inquiry into evaluation methods, as performed in the study in this thesis.

The structure of the empirical study in the thesis is presented in more detail below.

Structure

The overall study in the thesis consists of three main phases – (1) use of traditional usability evaluation methods, (2) refinement and re-design of these methods, and (3) use of refined and re-designed usability evaluation methods. All three phases will be further developed and described in detail below. An overview of the complete study is shown in Figure 4.3.

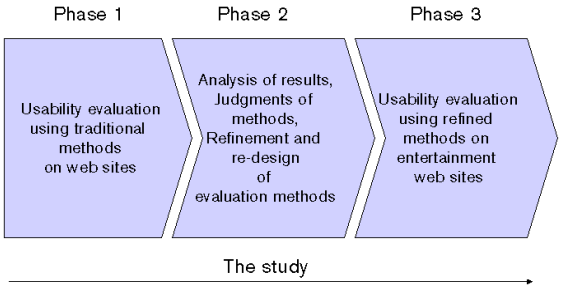

Figure 4.3 An overview of the three phases of the study.

Overall design of the study

Studies reported in this thesis can be divided into two groups: (1) those using empirical usability evaluation methods and (2) those using inspection methods. These two groups of studies had different designs. In the case of empirical usability evaluation methods, that is, the first group, a *between-subject design* was used. Briefly, this type of design can be described as follows:

> *"[..] two or more subjects are treated to different conditions, one of which may serve as a control condition. Contrasts are made between the results of the treatment."*
> (Solso, 1998, p.28)

In the case of expert evaluation, that is, inspection methods, a *within-subject design was employed*. This type of experiment design can be described as follows:

> *"[..] each subject undergoes two or more experimental conditions. The experimenter observes the results obtained after one treatment as contrasted with the results obtained after another treatment or treatments."*
> (Solso, 1998, p.28)

The choice of different designs was based on an assumption that sequence effects could potentially be more pronounced in empirical evaluation studies, compared with expert evaluation. In principle, within-subject design was considered more preferable, because it is typically more sensitive. However, the subjects participated in empirical evaluation studies did not have prior experience with evaluating web sites. Therefore, they were likely to change their behaviour significantly if participated in a sequence of web evaluation sessions (for instance, as a result of learning). Accordingly, a between-subject design was considered

most suitable for empirical evaluation studies. The experts applying inspection methods, on the other hand, were more familiar with usability evaluation tasks and procedures, so it was assumed that they were not likely to change their behaviour dramatically from session to session. That is why a within-subject design was used in these cases.

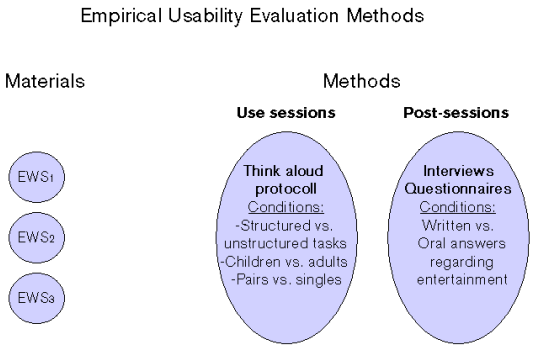

Figure 4.4 Exploration of traditional empirical usability evaluation methods. (EWS = Entertainment Web Site)

Phase 1: Usability evaluation using traditional methods

The first phase of the study comprised studies of empirical usability evaluations and inspection methods. The studies in which empirical usability evaluation methods were used were conducted on three entertainment web sites (W). The technique used was Think-aloud protocol. The design of the study is further described in more detail in Chapter 5. After the use sessions, the subjects were interviewed or asked to answer a questionnaire. The phase of study that includes the empirical usability evaluation methods can be described as shown in Figure 4.4.

In the studies using inspection methods, two parallel expert groups were employed to evaluate two web sites, one entertainment web site and one so-called information retrieval web site (IRWS). The second web site was used as a control site. The two groups differed in that the first group had ten (10) so-called experienced experts, i.e. HCI researchers and lecturers and the second group included twenty (20) novice experts, i.e. higher-level undergraduate students of informatics, with a specific focus on HCI. The students had a theoretical knowledge about usability evaluation but seldom or never used inspection methods before. To compensate to some extent for the novice nature of the experts in this group, the students were allowed to work in pairs. A total of ten (10) student groups were given the task of conducting a Design Walkthrough and a Heuristic Evaluation of the two web sites. As a help, they received handouts containing full descriptions of what they were to do together with forms on which to report problems and make other comments. The experts were subsequently asked to propose new heuristics,

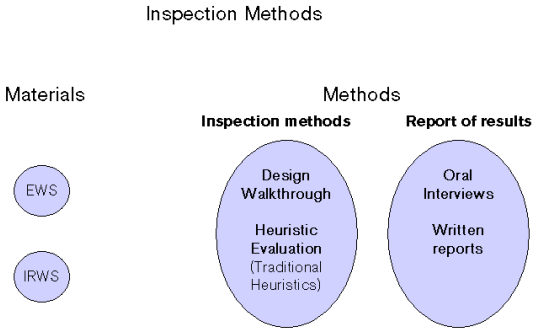

Figure 4.5 Exploration of traditional inspection methods. (EWS = Entertainment Web Site, IRWS = Information Retrieval Web Site)

suitable for evaluating fun, which were also included in the documentation. Finally, while the experienced experts were interviewed about the sites and the evaluations, the student experts were given the task of writing a report of their findings and suggestions and completing the handouts. The phase of the study containing these expert evaluations using traditional inspection methods is presented in Figure 4.5.

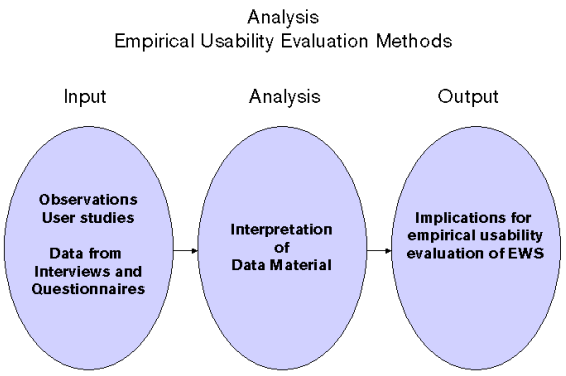

Figure 4.6 Analysis of results from use of traditional empirical usability evaluation methods.

Phase 2: Refinement and re-design of evaluation methods

The results from the first phase of the overall study were used as input for the second phase – the refinement and re-design phase. This phase of the study is presented below:

In the empirical usability evaluation methods part, the received data was in the form of observations from the user studies and the answers to questions in interviews and questionnaires. Each tested condition was evaluated in relation to the data. Revised methodological considerations regarding the conditions in focus were constructed in light of the analysis. These were later further revised and developed in the third phase. A graphical view of the analysis process of empirical usability evaluation methods is presented in Figure 4.6.

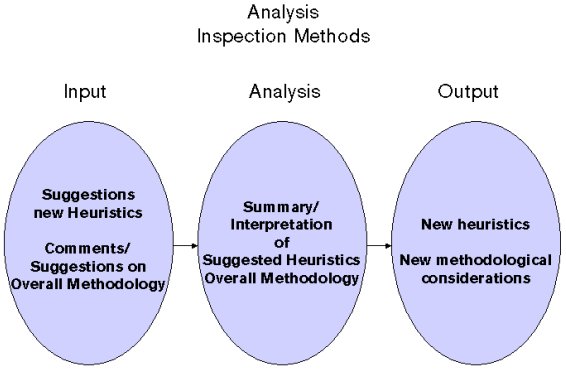

Figure 4.7 Analysis of results from use of traditional inspection methods.

In the inspection method part, the data consisted of completed handouts, with reported problems and other comments about the web sites together with experts' suggestions for new heuristics. The written reports from the student experts and the answers from interviews with experienced experts provided further input about more general methodological aspects. All this data material constitutes input for the interpretation and analysis in this second part of the study, conducted by evaluators in the research project. The data were interpreted on the basis of theoretical frameworks of EWSs and evaluation methods, described earlier. This analysis and interpretation generated a new output in the form of a new set of heuristics for Heuristic Evaluation for further exploration in the last phase of the study, as well as more general suggestions concerning how to evaluate fun using Inspection Methods. This process of analysis described above is presented in Figure 4.7

Phase 3: Usability evaluation of EWSs using new and refined evaluation methods

The third phase of the study was again empirical, i.e. it contained evaluations of EWSs. A number of studies were conducted based on findings from earlier phases of the study. The new set of heuristics for Heuristic Evaluation was implemented and the results regarding the conditions focused on in this study for empirical usability evaluation methods were further explored. The main purpose of this third phase was to show how applicable these new methods were with reference to evaluation of EWSs.

Two entertainment web sites were evaluated using empirical usability evaluation methods, which were on the basis of earlier findings. Ten (10) subjects evaluated all web sites, i.e. the total number of evaluations is twenty (20). After the use sessions, the subjects were interviewed about the web sites as well as about the evaluation session. The data were analyzed and further revisions of the evaluation methodology were conducted. The final proposal for a usability evaluation methodology for empirical usability evaluation methods for evaluating EWSs is presented as a conclusion, as shown in Figure 4.8.

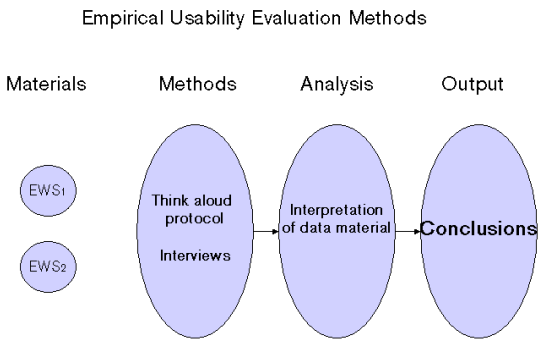

Figure 4.8 Exploration of new and revised methodology for evaluation of entertainment. (EWS = entertainment web site)

Regarding inspection method evaluations, the process was somewhat more complex. First, two entertainment web sites were evaluated by the same experienced experts as in the first phase of the study. They used an approach called 'free surf' –which replaces Design Walkthrough. The experts then conducted Heuristic Evaluation with the new set of eight (8) heuristics. Further, experts conducted a 'meta-evaluation' of the suitability of each heuristic for the evaluated web site. Experts were given a set of handouts similar to those in the first phase, which they used to report findings and suggestions for further revision of the methodology. Finally, the experts were interviewed about both their results from the evaluations of the web sites and their proposals for further developments of the evaluation methodology. This data material, i.e. the handouts and the answers from the interviews, were then subjected to further interpretation and analysis, and evaluators make new changes in the evaluation methodology. A further round of inspection method evaluations of EWSs was conducted by experts. Once again, two entertainment web sites were evaluated using the 'free surf' methodology approach and Heuristic Evaluation, this time with a revised set of heuristics. The number of heuristics was now ten

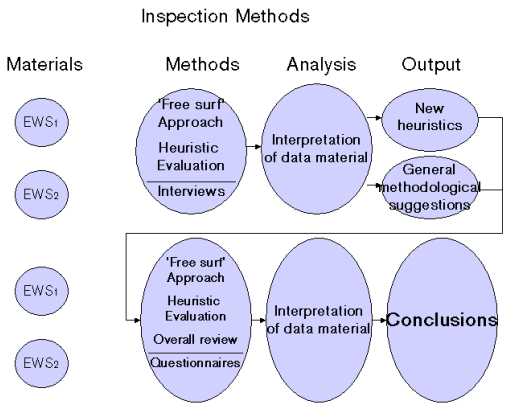

Figure 4.9 Exploration of new and revised methodology for evaluation of entertainment. (EWS = entertainment web site)

(10). In addition, a meta-evaluation of the suitability of the heuristics for each of the web sites was also used in this final round. An overall review of the complete web site was also added to the methodology to comply with the demand made by experts in the earlier phases. Finally, when the evaluations were complete, the experts answered a questionnaire in free text. The data from all the methods included in the revised methodology were then analyzed and conclusions were drawn from this third phase of the study. This is shown in Figure 4.9.

Materials used in the study

The choice of materials in any study is critical for its outcome. Choices must be properly considered and documented, as they have a great influence on what will be found later in the analysis of the data material. In this study three types of material were considered to be very crucial, i.e. the choice of methods to be explored, web sites to be evaluated, and subjects and experts to be included. General aspects of these choices are developed below and further details of the materials in each case in the study are presented.

Choice of methods

As presented in Chapter 1, there is a large number of usability evaluation methods already in existence, used in both research and practice. The choice of methods when designing the study was based on certain assumptions: (1) the chosen methods should be in common use and known to the research community; (2) as the concepts of entertainment and fun are mainly subjective in nature, the methods chosen should to some extent reflect this; (3) most of the well-known and commonly used usability evaluation methods in the HCI research area focus more on aspects of usability such as 'efficiency', 'errors', 'memorability' and 'learnability' rather then the aspect of 'user satisfaction'. As the first assumption, as shown above, of 'commonly used and known' was important in this study, no specific requirement for methods specifically designed for the aspect of 'user satisfaction' was specified. Another argument in relation to this is that as the concept of 'user-satisfaction' is a term that encompasses user satisfaction in *all* types of IT systems, it was not obvious that methods designed to evaluate this aspect would necessarily be better than any other technique when evaluating EWSs.

The traditional usability evaluation methods chosen were:

- Design Walkthrough,
- Heuristic Evaluation,
- Think-aloud Protocol,
- Interviews
- Questionnaires.

The different types of walkthroughs in HCI are numerous and are frequently used in research (Nielsen, 1994a; ibid. 1994b; Virzi, 1997, Karat, 1997). For instance, Karat wrote 1997 that "*design walkthroughs have been conducted for more than 20 years*", which today implies at least 25 years of experience of using the method in the HCI research community. For this reason, the method qualified for further exploration in this study.

The Heuristic Evaluation method has rather come to be known as '*the* inspection method', perhaps because of its inventor – Jakob Nielsen. Nielsen is a well-known personality in HCI research and is nowadays also a spokesman for applying usability to contexts other than research, worldwide. Some even call him 'the most famous web usability guru'[1]. Heuristic Evaluation was first developed in 1991, and since then has been used frequently in research and, also importantly, in practice. It was chosen as one of the methods for this study largely because of this popularity. The method is said to include a number of advantages, however, the question is how applicable is it when evaluating fun and entertainment.

Think aloud protocol, interviews and questionnaires are all very frequently used methods for empirical usability evaluation (c.f. Nielsen, 1993; Karat, 1997; Dix et. al, 1998, Shneiderman, 1998). This was the reason these methods were chosen for further exploration in this study.

A number of methods presented in Chapter 1 are described by Jordan (2000) as being suitable specifically for evaluating user satisfaction. As discussed above, the aspect of 'user satisfaction' does not correlate with fun and entertainment. However, 'user satisfaction' as well as fun and entertainment to a large extent include subjective values, hence the methods Jordan discusses can be of importance in the context of this thesis. The majority of the methods described by Jordan (2000) cannot, however, be regarded as being well-established in the HCI research community (yet) and as that was a criteria for inclusion in this study these methods were excluded. However, a closer look at the methods chosen in the study in this thesis reveals that some of them *are* mentioned as being suitable for evaluating user satisfaction by Jordan. Think aloud protocol, interviews and questionnaires are included in these methods. Furthermore, many of the other approaches mentioned resemble other chosen methods. For instance, Private

camera conversation may be seen as being similar to Think-aloud protocol, co-discovery bears strong similarities to the pair sessions conducted with the Think-aloud protocol in the study and controlled observations could be compared to the structured vs. unstructured condition in the empirical evaluation part of the study. As regards the inspection methods mentioned by Jordan (2000), Expert appraisal is somewhat similar to Design Walkthrough. Because of these similarities between methodological approaches used in the context of the study in this thesis, the results may also inform future research concerning the methods presented by Jordan (2000).

Choice of web sites

On the basis of the theoretical frameworks, as presented in Chapter 2, the evaluators identified the domain of EWSs as a type of web site which was interesting from a methodological viewpoint. The members of the research team decided that a number of web sites should be chosen and evaluated. The two possible choices were to evaluate web sites found anywhere on the web which matched the criteria for an EWS, designed by anyone, or to contact designers of what were seen as EWSs and to ask for a collaboration regarding ongoing or completed EWSs. Both alternatives had pros and cons but the second one was chosen, i.e. to contact designers of EWSs, mainly for practical reasons in the sense that the evaluators had access to all background information, such as target group of users, the goals of the originators, measures of success etc. This was important information to have when evaluating the EWSs, as it provided important knowledge about *who* to test the EWSs on, what *measures of success* to focus on in evaluations and what the goals and purposes of the originator of the EWS had been, which should be considered as important when defining what should be regarded as the *form* and the *content* of the EWS to be evaluated.

The web sites in the study were chosen from the completed or ongoing design projects of the collaborators, Paregos Mediadesign, and particularly those which included one or more of the typical features of entertainment web sites, as discussed in Chapter 2 were chosen. As mentioned earlier, a list of typical features in EWSs was formulated in collaboration with staff from Paregos Mediadesign. Close investigations concerning the fit with typical features of entertainment, preceded the choosing of the web sites for the study. If a web site included only one of the typical features, the type of feature included was crucial, as some are not sufficient on their own to justify classifying the web site as an entertainment web site. The most significant feature in this sense was the fifth – high quality graphic design – which could not by itself be considered sufficient to label a web site an entertainment web site. Here, a parallel can be drawn with the framework including *form* and *content,* where the form includes graphic form, structure and

navigation and 'added value'. If a web site's included entities connected with *form* contained graphical form and structure and navigation but no 'added value', it was not considered an EWS. A table of the chosen entertainment web sites in relation to the typical features is shown in Table 4.1.

	1	2	3	4	5	6	7
Eurovision Song Contest	✓		✓	✓	✓		
Mosquito	✓	✓	✓	✓	✓		
Total defence	✓				✓	✓	
Skyscraper	✓	✓	✓	✓	✓		✓
Vodafone – 'How are you?'			✓		✓		✓
Stadium Activity Town	✓	✓			✓		✓
Bad guys monkeys			✓		✓		
Jernkontoret 'Captain Steel'			✓		✓		

Table 4.1 An overview of the entertainment web sites chosen for inclusion in the study.The numbers in the table indicate the typical features, found on entertainment web sites, as described above. 1= Entertainment information, 2 = Downloadables, 3 = Small 'stand-alone' games, 4 = Plug-in-technology dependent features, 5 = High quality graphic design, 6 = Edutainment content, 7 = Communication with others.

Selection of suitable subjects and experts

The process of finding suitable subjects and experts to carry out evaluations of any kind is in general a very demanding procedure. Proper descriptions of targeted audiences for different systems or web sites, given by designers and/or target organization, are crucial. Furthermore, even if target users can be specified there can be many reasons to reconsider as some types of users fit better than others into experiment situations. For instance, if the target users are children, does one have to use only children, or can other groups be considered instead? The choices should be carefully considered, and the options should be evaluated for each situation.

The problem is similar regarding the choices of experts. Expertise in HCI in general and in expert evaluation in particular are two examples of factors that need to be considered. When finding experts for a project such as the one in this thesis it is also important to consider how well they fit with the target group of the included EWSs. Are they included in the target audience of the system, and if not, what might the consequences be? If the evaluation project is included in a research process, as in this case, it might be worthwhile considering using experts with knowledge and/or with personal experience of such processes. The reason for this is that expert evaluation processes in the field of research often include more reflective discussion than does expert evaluation in industrial type projects.

Experts

The experts in the study were asked personally if they could consider participating as experts in this study. Certain factors influenced the process of selecting those asked: (1) All those asked were on the staff of the Department of Informatics, Umeå University. The reason for this was accessibility, as the experts were contacted and employed in numerous stages in the research process. (2) Those chosen as experts had to have specific experience in either lecturing or research or both in the field of HCI. Informatics as a department, and, to some extent, discipline includes many international research fields, such as HCI (Human Computer Interaction), CSCW (Computer Supported Cooperative Work), IS (Information Systems), Virtual Communities, etc. Primarily those sought were people who had practical experience of inspection method usability evaluation, and if this proved impossible they should at least have extensive knowledge about methods for usability evaluation, from their own research or from lecturing on the subject.

All of the ten (10) people asked agreed to participate in the study. Furthermore, the group of experts remained consistent, i.e. the same experts participated in all the expert evaluations in the study. Only in the final evaluation of the third expert tests did one expert who had to leave, need to be replaced with a new expert. The substitute had the same background as the original expert, but less experience in evaluating entertainment web sites, which had to be taken into consideration when analyzing data material.

Another group of experts were also included in the study, the so-called 'novices'. These were undergraduate students, taking a course in HCI towards the end of their degree course in Informatics. They were twenty (20) in number. They were divided into pairs (2), to compensate in some measure for their lack of experience in the subject. The task they were assigned was part of the examination for the HCI course, and as such was compulsory.

Subjects

As mentioned above, the choice of subjects for experiments of any kind is difficult and many aspects must be considered. The subjects chosen in this research project were selected according to the target audience of the different web sites, as presented by the design team. The specific choices are further discussed in each section describing each phase of the study.

Equipment used in this study

No usability lab was available during the time this study was conducted. Therefore all the empirical parts of the study were carried out with other, more stand-alone, technical solutions. A digital video (DV) camera was set to record the screen as well as the user. In interviews, the sound was recorded with a mini-disc (MD) recorder or a more traditional analogue audio recorder. The physical settings, i.e. the rooms where the experiments took place, were offices, computer laboratories and other available rooms at Umeå University. Two examples of such settings are shown below. In Figure 4.10 an example of a single-user setting is shown, and Figure 4.11 shows an example of a setting where users work in pairs[2].

Picture 4.10 An example of a test setting with a single subject and one evaluator.

As described in later sections, other settings were also used, located in the vicinity of tested subjects, for instance meeting rooms at schools where students participated as subjects in the experiment. The physical settings were chosen to maximize suitability and accessibility for the subjects.

Picture 4.11 An example of a test setting with two subjects collaborating and one evaluator.

Understanding and generalization of qualitative results

One thing worth discussing when doing research - qualitative or quantitative - is the point from where to decide to start the interpretations. One approach is to let the standpoint initially be theoretical and from there look at the empirical data. This can also be called *deduction* (Alvesson & Sköldberg, 1994; Patton, 2002; Wallén, 1996).

Another approach is to let the empirical data, solely, drive the process of interpretation. The data never reaches any theoretical level, but the researcher let

the data speak for itself. This, often highly questioned, approach is called *induction* (Alvesson & Sköldberg, 1994; Patton, 2002; Wallén, 1996).

Further, a third choice is to start in empirical data, try to bring in up to a theoretical level, in order to, for instance, put it into existing theories, or to create useful theories or frameworks. After this is done, the theory, already existing or created from empirical data, is used to analyze other empirical data. This third approach can to some extent be seen as a combination of the two approaches above, and it is usually called *abduction* (Alvesson & Sköldberg, 1994; Patton, 2002; Wallén, 1996). A model of the three approaches deduction, induction and abduction is shown in Figure 4.12

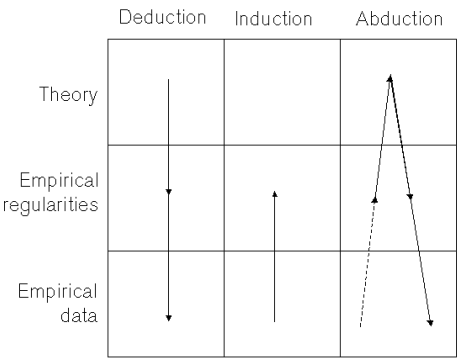

Figure 4.12 Description of deduction, induction and abduction (Alvesson & Sköldberg, 1994, p.45)

In this thesis, the research process should be understood to be *abductive*. As one of the main obstacles is to find a way to operationalize fun and entertainment in the context of web usability, a basis for doing this was necessary. In search of a way to do this, theories were consulted. This is a common first step before conducting an abductive process.

> *"The analysis of the empirical data can very well be combined with, or preceded by, studies of earlier theory in literature: not as mechanical application on singular cases, but as a source of inspiration to find patterns that provides understanding."*
> (Alvesson & Sköldberg, 1994, p.42, my translation)

A number of related theories to entertainment and related concepts were approached to see to what extent they could guide in this operationalization. As it turned out, the theories gave fruitful input in how to understand the covered concepts on a general level, but were less suitable in operationalization of fun in particular. Because of this, empirical investigations were conducted without specific guidance from theories in operationalization of fun and entertainment.

Types of empirical evidence

When conducting studies of this magnitude the amount of data material becomes extensive. In total, the study included the following material: In the empirical usability evaluations, eighty (80) empirical evaluations of five (5) web sites were conducted, each comprising pre-test questionnaires, Think-aloud sessions and post-task interviews and/or questionnaires. In the expert evaluations thirty (30)

experts produced evaluations of five (5) individual web sites together with follow-up interviews and/or questionnaires. Finally, the expert group called 'novice experts', i.e. the students, conducted the same expert evaluation of ten (10) individual entertainment web sites of their own choice. The student experts group also conducted inspection method evaluations using a different inspection method from the inspection methods designated in the handouts, i.e. Design Walkthrough and Heuristic Evaluation. This additional method was used on three web sites, the information retrieval web site, the assigned entertainment web site and their own choice of web site, with the aim of providing feedback about the suitability of the chosen method for use on entertainment web sites. However, these two steps – the evaluation of the third web site as well as the evaluations with the individually chosen technique are of somewhat secondary importance in the study in this thesis. The existence of the additional entertainment web sites and inspection methods is worth mentioning, however, as they may have influenced the students proposals for new heuristics and general methodological guidelines for inspection method evaluation of entertainment web sites.

There were approximately one hundred (100) hours of tape recordings, both video and audio. All interviews were transcribed producing approximately six hundred (600) pages of text. In all sessions the evaluators took extensive individual notes, which were discussed among them after each session where more than one evaluator had been present. The evaluators were all members of the research group 'Entertainment Services' at CDB (Center for Digital Business) at the Department of Informatics, Umeå University. In total four (4) evaluators participated including the author of the thesis.

Data collection

Data collection and analysis in a large study, such as the one presented in this thesis, is extensive and a complete analysis of *all* aspects is, of course, impossible. However, being present in most of the data collection sessions helps in the analysis part of the process. Some researchers have argued that in user testing every problematic usability aspect will have been found after five users have been tested (Nielsen, Alertbox, March, 19, 2000[3]: Nielsen & Landauer, 1993). This some researchers argue that this is not always true (c.f. Spool, 2002), especially if large web sites are evaluated and the approach used is quite 'free', as opposed to 'structured'. Incidents *do* get repeated, the same comments *do* recur, if, as in this case, twenty (20) subjects are included in an evaluation, using the same test design on the same web site. After such a series of tests, evaluators have a lot of time to make their analysis at the evaluation setting. In this case, video- and audio-tapes can be used more to refresh memory and/or exemplify results to other people. In

other cases, where a less structured approach is used, and as the evaluator has to focus and concentrate on what is happening in the test setting, analysis has to be conducted by viewing the videotapes. In both cases, the video recording of user sessions is crucial, for instance, for gaining external insights into the data material, especially when conducting qualitative research, as interpretation is both difficult and critical for the outcome of the study.

Interviews

Using interviews as a method for obtaining empirical evidence is rather complicated. The quality of the output of the interview, the data material, is heavily reliant on the interviewer (Patton, 2002, pp.341-348). The three main variations in qualitative interviewing are (Patton, 2002)[4]**;**

- Informal conversation interview
- General interview guide approach
- Standardized open-ended interview

In the case of the expert evaluations in the study in this thesis, the majority of the questions in the interviews were decided on before the interview process began. The number of questions differed but was generally around ten (10) to fifteen (15). The questions were asked exactly as they were worded on the interview form. The interviewer also used follow-up questions, when interesting aspects arose in the interviews. Furthermore, suggestions that appeared in the earlier interviews, were sometimes discussed in subsequent interviews, i.e. earlier interviews were more *standardized* regarding pre-set questions in comparison to later interviews. This ultimately created a situation where suggestions were not only given by one expert but were further reflected on before finally being included as a suggestion. Earlier comments and suggestions were also discussed with other experts, interviewed later in the series of interviews, and were sometimes further developed. This could be seen as a part of the analysis or at least as a preliminary stage. It was done to bridge the gap between the extensive amount of data and a detailed and valid analysis of the data. This flexible approach, where earlier discussions were included in later interviews, meant that the group of experienced experts was included to some extent in the analysis and interpretation process. This was deemed relevant and important, as it was this group who were the prime source of proposals for heuristics and general guidelines for evaluating EWSs. The order in which the experts were interviewed differed in each round, mainly determined by the amount of time the expert had, but also by the flexible approach. This balanced the impact of individual experts, as every expert had a random number of possible comments from others to consider. The same interviewer, i.e. the author of the thesis, conducted all of the interviews with the experts. All three

of the interview approaches discussed above, were used. There was a prepared set of questions – as in the *standardized open-ended interview*. As the series of interviews progressed these initial questions were supplemented with proposed methodological considerations or topics discussed earlier – as in the *general interview guide approach*. Finally, sometimes, a high level of interactivity between the expert and the interviewer was achieved when the interviewer used follow-up questions – as in *informal conversational interviews*.

A more standardized type of interview approach was used in the interview sessions of the subjects in the empirical usability evaluation. A ready-prepared interview form was used, and the questions were asked as they were worded on the form. Only in situations where the subject either gave very short answers or when the answer was not clearly understood did the interviewer ask follow-up questions. This was done to make the conditions for all the subjects as similar as possible. These interviews were also as standardized as possible because the interviews throughout the study were conducted by four different people.

Written reports

A written report of their findings was requested of evaluators in the case of inspection method evaluations conducted by the group of so-called 'novice experts', i.e. the students. These students also presented their work and reports as a part of their degree course. The reports were treated as the equivalent of the interviews conducted with the other expert group. Quotations were abstracted from the written reports in those cases where interesting information was presented. The reports complemented the completed handouts.

Process of designing new guidelines and frameworks for usability evaluation

A variety of approaches were used in the creation of new guidelines for evaluation from input from the evaluations conducted. In order to obtain useful output from the empirical evaluations conducted, discussions were held between the evaluators involved. Suggestions for guidelines were made which were further discussed and revised.

The input concerning guidelines for expert evaluations came mainly from the experts themselves. The transformation process from suggestions to guidelines is made visible in the thesis in form of a complete presentation of suggestions, tables showing overviews of the connection between the new heuristics and the expert sources. If suggestions were not further developed into heuristics, the reader is presented with a discussion of the reason for this.

The interviews with the experts also produced numerous suggestions on a more general level regarding how to evaluate entertainment. Many of these, often more methodological suggestions, served as input in the following phases in the study. Quotations are presented in the thesis, to give the reader a chance to evaluate their validity.

Throughout the thesis, there is an effort to make visible as much of the interpretation process as possible, to allow the reader to judge the validity of the results.

In the next chapter, the first phase of the study is presented in more detail. This phase includes the empirical evaluation of three entertainment web sites. The methods used are: interviews, the Think-aloud protocol and questionnaires. A number of conditions, such as 'structured vs. unstructured tasks', 'testing children vs. adults' etc. are elaborated, in order to demonstrate how these conditions may be used to produce high quality results in evaluations of entertainment web sites.

Footnotes

[1] C.f. Crawford Killian's column in 'Content Spotlight' at http://www.content-exchange.com/cx/html/newsletter/1-21/ck1-21.htm (2003-09-27)

[2] The subjects in the pictures were not subjects in the study. The purpose of the pictures is only to show physical settings *per se*. Any images or recordings from the sessions in the study are confidential, seen only by evaluators and transcribers in the research group.

[3] http://www.useit.com/alertbox/20000319.html (2003-09-30)

[4] For a more thorough description of the three variations, see Chapter 1.

Chapter 5

Using traditional empirical usability evaluation methods

Background

In this phase of the study the strategy was to use traditional usability evaluation methods to investigate their applicability to EWSs. Based on this knowledge, the methods were re-designed and used in further evaluations, in order to examine any changes in outcome of these evaluations. The evaluations were carried out using with both empirical evaluation methods and inspection methods. In this chapter the first attempts to evaluate EWSs using empirical usability evaluation methods are presented.

Methods included in this part of the study are think-aloud protocol, questionnaires and interviews. In these evaluations, a number of pre-conditions were chosen for the tests, with the aim of providing more specific feedback in relation to the stated purpose. These conditions were investigated in the tests, in order to show whether changes within conditions had any affect on the results of evaluations.

Subjects were recruited according to intended target group of users for each web site included, as stated by designers.

The entertainment web sites used in this part of the study were chosen according to the categories of EWSs described earlier. The approach is to include web sites that vary in profile regarding typical entertainment features described earlier. This first empirical study includes an EWS intended for edutainment, an EWS which was a support web site for an event, and finally an EWS which was seen as an extension of a TV-show. The web sites differ as regards intended target

group of users. The users of the first EWS are children and teenagers, of the second people aged between 20 and 50 with an interest for the specific event that is spot-lighted, and of the last EWS the TV-audience of the show, mainly was considered to be 7 to 30 years old.

Methodology

Various aspects of the methodology in this part of the study are presented, described and discussed below. Aspects covered are the subjects, the material, i.e. the web sites, the design of the study and the procedure.

Subjects

As mentioned above, subjects were recruited according to intended target group of users for each web site included in the study. Information about target audiences was provided by the designers of the web sites. The practical selection of the subjects to participate in the study within these groups of target users varies with each web site.

Eurovision Song Contest web site: The subjects in this study were contacted through e-mails, sent to students and staff at Department of Informatics at Umeå University. The entire staff received invitations to participate. The students selected were upper level undergraduate students taking an HCI course in informatics, given by the author. Some randomly chosen acquaintances of the test team, who matched the target group of the web site, were also included. Approximately 80 people received the e-mail, 20 responded positively and all of these were tested.

Mosquito web site: Two groups of subjects were used in these evaluations; (1) 'Adults', who were recruited within the age limits 20 and 30 years. The tested group were mostly students. (2) Children were recruited within the age limits 7 to 14 years, which was the target group of the site. This group was recruited from a local school. The majority of these children were aged between 9 and 10 years.

Total Defence web site: This site had a similar target group to the Mosquito web site, and the children were recruited from the same school as above. However, different subjects were used in the two tests. The age of the children was 9 to 10 years.

The table below shows data from the tests in relation to different sites. ESC corresponds to Eurovision Song Contest, M to Mosquito and T to Total Defence.

	ESC	M	T
Contacted	e-mail	A:Personally/ C:through teacher	35:Through teacher
Positive/total	20/80	A:10/15 C:11/35	13 /35
Group(s)	Adults, ESC interest	C=Children, aged 7-14	Children, aged 7-14
Group(s)		A=Adults; aged 20-30	

Table 5.1 An overview of selection and grouping of the subjects in tests

In order to collect background information about the subjects, a questionnaire was delivered to them before each use session. This questionnaire included items on sex, age, whether the subject was mainly a PC or MAC user and on what previous experience the subjects had had. The aspects investigated included how they considered themselves as experienced web surfers, how frequently they used the web, how much they used computers in their daily work, their interest in popular dance music, whether they had visited the evaluated web sites earlier and finally, whether they had any previous experience of participating in user studies. This background information is presented in detail in Appendix II.

Each subject received a ticket to the cinema in appreciation of their time and effort.

Materials-the web sites

This study included three web sites, as mentioned, and these are further described below[1].

The Eurovision Song Contest web site

This was an event site for an annual television event – the Eurovision Song Contest – which is an annual contest involving a number of European countries. In 2000 the contest was held in Sweden and Paregos Mediadesign AB designed the web site that supported the contest. The target group of this site was people interested in modern dance music (in Swedish called 'schlager music') in general and more specifically in the Eurovision Song Contest. The estimated age of the target group was 20 to 50 years. The purpose of the event site, as stated by originators through

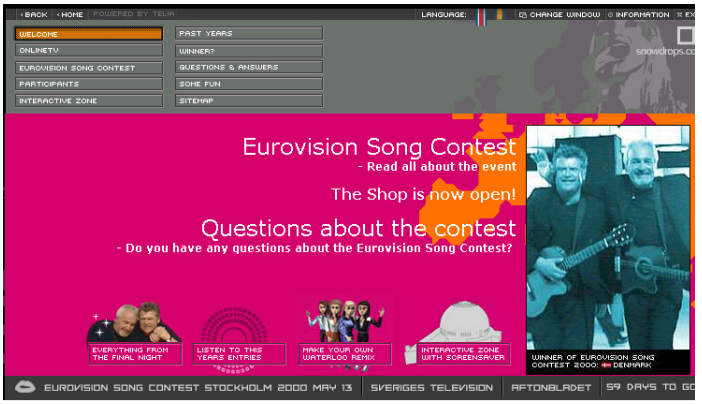

Figure 5.1 The ESC home page

designers was:

> *"Swedish Television and Aftonbladet [a Swedish evening tabloid newspaper] wanted a web site for the Eurovision Song Contest that was not just a pale copy of the television show and they wanted it to present the sponsors in a sensible way. The site was steadily the most visited for the weeks before and after the competition. The visitor can compete in a Song Quiz (with other visitors) and be his/her own DJ by mixing his/her own version of ABBA's Waterloo, and so on."*
> (http://www.paregos.com)

Figure 5.2 The Mosquito home page

The Mosquito web site

The Mosquito (M) web site was created by Paregos Mediadesign AB, as a support site of the Swedish television show with the same name. The target group for this site was children 7-15 years old and people interested in design and technology in general. The audience was probably to be found in the target group of the television show, but not necessarily. Quoting the description of the web site, from the corporate site of Paregos, the purpose of the production was as follows:

> *"Paregos accepted Swedish Television's challenge to create the web version of Mosquito as an extension of the TV-show and as a meeting place for those who like the program. The result is a flash site that has been awarded several prizes in the media business, was chosen site of the summer by the magazine Resumé and won the Prix Italia prize for "the best innovative solution". But mostly, it has been a high-octane, crazy, wonderful meeting place for all the "Mosquitoes".*
> (http://www.paregos.com)

The Total defence web site

This edutainment web site was released in 1999 and was given a lot of attention. The designers received Swedish as well as international design awards, such as, 'best information site' in Sweden. The intended target group of users was children and teenagers aged 7 to 15 years. The main purpose of the site, as stated on the web site of Paregos Media design AB is:

Important information does not have to be boring. Or rather, it MUST NOT be boring. The Total Defence needed a new web site that could be used in teaching students about total defence and security policies. Everything concerning total defence, such as UN rights, military defence, civilian defence, etc., would be found here. The visitor should remember everything he/she had learnt and not just consume facts - that was the most important issue.
(http://www.paregos.se)

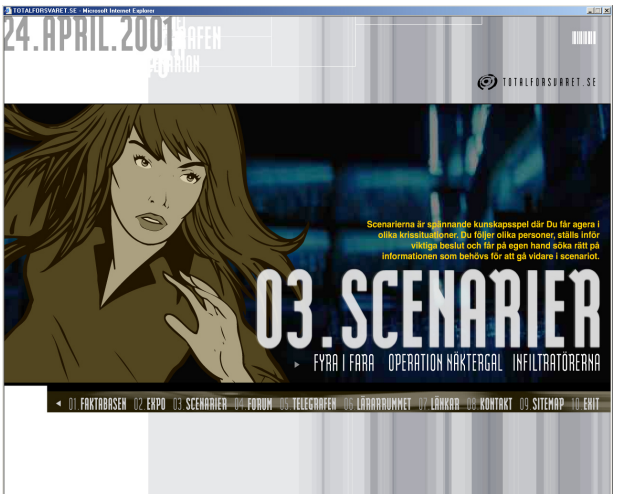

Figure 5.3 The Total Defence home page

Design of the study

The test settings

The tests were conducted in a variety of settings, such as computer labs, meeting rooms etc. The physical test settings included a desk with a computer, a digital video camera on a tripod, pointing towards subject and computer screen placed beside the subject. Slightly behind the subject, the coordinator of the test sat on a chair. Finally, a second experimenter was present in the setting.

Two experimenters were present in all of the sessions, one as a coordinator, running the test and making all the decisions related to test strategy in each session and a second experimenter taking notes, mainly about the test session *per se*. A digital video camera was used to record all the tests. The camera was mainly focused on the computer screen, but if possible, the subject was also recorded. It is difficult to use a video camera to record actions on screen and, in order to have a clear understanding of what happened on the computer screen, the evaluator used follow-up questions about on-screen actions.

Conditions investigated

The study was designed to provide an investigation of a number of conditions. These conditions were:

- Pairs vs. individuals
- Structured vs. unstructured tasks
- Testing children vs. adults
- Written vs. oral answers about entertainment aspects

These specific conditions were arrived at after discussions within the research group responsible for the evaluations and were based on earlier usability evaluations of other web sites, entertainment and others. The decisions were also

based on earlier discussions about the design of this study. For instance, whether it would be appropriate to test children, regarded in some cases as difficult; whether the questionnaires regarding entertainment aspects would be answered orally or in writing; whether structured tasks would be an appropriate choice in the context of evaluating entertainment web sites, which are often explorative in it's nature; and finally how the process of thinking out loud could be facilitated in the context of entertainment and whether pair sessions would be the answer. These and other aspects were discussed within the group and the above list of conditions was the result. The conditions are further developed below:

Pairs vs. Individuals

The intention was to see whether the number of simultaneous subjects tested was of significance. The questions posed were: How would the people tested react when being tested with a friend? Would the think-aloud protocol work better in pairs, or is web entertainment better experienced and tested alone?

	ESC	M	T
Individuals	14	17	1
Pairs	3 (6 subjects)	2 (4 subjects)	6 (12 subjects)
Total tests	17	9	7
Total subjects	20	21	13

Table 5.2 An overview of the subjects in the tests

Structured vs. unstructured user tasks

The use of structured tasks is seen as essential when carrying out user evaluations. The subject is supplied with a couple of tasks and the experimenter evaluates the result of the test – how the subject solved the task, the time required, mistakes committed, subject's comments etc. However, it is important to ask what happens in the case of evaluating entertainment. Are structured tasks a proper choice in this context? How does a more elaborate technique work where the subject is encouraged to be exploratory? Approximately one third of all tests were conducted using structured user tasks where users were given assignments to complete. One third used a mixed approach with both structured and unstructured tasks, where the latter could be described as using 'free surf'. For the remainder of the tests unstructured user tasks were used and 'free surf' was used throughout the whole test.

Testing children vs. adults

When the designers of the three sites informed experimenters about the age span of the subjects in the target groups the experimenters became somewhat reluctant as two out of the three sites had children aged from 7 to 15 years as the main target group. Testing children is known to be different from testing adults. The experimenters were not clear about what exactly would be different. However, a decision was made to adhere to the target group for the most part, but also to conduct a split test on the Mosquito web site, including both adults and children. This mixed approach was used in order to get a feeling for the differences in the results between adults and children. Instead of being a problem, the challenge turned out to be two important conditions to test.

Written vs. oral answers to questions concerning entertainment

In the majority of the literature on usability evaluation interviews are used to test the subjective satisfaction of the subjects (c.f. Nielsen, 1993). Whether this is the most successful approach when investigating entertainment and experiences was on of the study questions, as was whether subjects should be asked about this in writing in the oral interview? Both approaches were explored in this study. For the ESC, interviews were mostly oral. In the evaluation of the Mosquito web site, the same questions were given to the subjects, but this time in writing. With Total Defence most of the interviews were oral, mainly because all the subjects were children.

Procedure

Most of the tests consisted of three parts; a pre-test questionnaire, with background information about subjects; a think-aloud session, with more or less structured user tasks, and finally a post-test questionnaire or interview, with questions about the site and the test.

Pre-test questionnaire

The pre-test questionnaire included ten (10) questions regarding age, gender, whether the subject had visited the web page before and the level of expertise of the subjects regarding computer work in general and web surfing. Finally, the subjects were asked if they had participated in any kind of experiments earlier[2]. The information received in this part of the study was used as an aid in the analysis phase.

Think-aloud session

The think-aloud session has been discussed earlier, but generally, the time for the sessions was 20-30 minutes. If the subjects had not finished after 30 minutes, the test leader stopped the session and moved on to the post-test questionnaire and/or the interview. In some cases a short scenario was used[3] mainly to contextualize the use of the web sites.

Post-test questionnaire or interview

Various types of post-test questionnaires were used[4] as some included questions about aspects of fun. All questionnaires included questions on both the evaluation *itself* as well as on the web site. In those cases where the session included pairs of users, this aspect was also covered in the questionnaire and the interview. In the cases where interviews were used, the approach was exploratory, i.e. experimenters explored various types of questions in different sessions in order to gain knowledge about how the investigation of entertainment in the context of web usability could be explored in interviews.

Results

In general, the tests produced numerous results relevant to both the evaluation and design of entertainment web sites. Below the focus is on results concerning evaluation, beginning with some overall findings and continuing with more specific findings, sorted according to the different parameters elaborated. It is also important to highlight the importance of results regarding the design of the web sites, since these findings exemplify important methodological results.

Before exploring results connected with the conditions focused on in the study, some overall findings are identified. The first main finding is concerned with the relation between fun of use and ease of use. Testing the *fun of use*, as in the case of evaluations of EWSs compared to the *ease of use*, as in evaluations of more traditional applications and web sites, may differ a great deal. However, the ease of use must not be forgotten in evaluations of entertainment in EWSs. For instance, overall results from the study show that subjects do not find it amusing when there is a chance they are being fooled. On one of the sites one of the buttons in the navigation bar was marked 'Fun Stuff'. Subjects interpreted it as indicating an area where games, e-postcards and other fun stuff could be found on this type of site.

> "*Fun Stuff, that is where the fun things are found…)what is this..oh no, this is not…fun"*
> (Example of a comment in an evaluation of ESC)

Instead of what they expected, subjects got a commercial film from the sponsors with a high level of sound. Almost all the subjects were surprised and/or annoyed by this part.

> *"..what is this..oh no, this is not…fun"*
> (The continuation of the above comment in an evaluation of ESC)

None of the subjects showed any interest in the content of the commercials after being tricked.

> *"The thing – that 'Fun Stuff', that was annoying – I did not like it. It felt like I was fooled into going there and after the click I was trapped into watching. It was probably the sponsors of the web site.."*
> (Example of comment by subject in interview after use session – ESC)

Another example of the clear link between ease of use and fun of use shown in the study was when problems occurred in relation to navigation. Overall throughout all the sites, navigation was not something subjects found amusing when it was not useful. Instead, it seemed to have a negative influence on the overall experience of the EWS, i.e. problems regarding ease of use influenced the concept of fun of use.

> *"..but it was sometimes difficult to find [the way] in the site..to answering the questions was hard because of this…I felt…I got it wrong at first…they [the designers] could have done it better"*
> (Example of a subject, frustrated with the poor navigation in the ESC web site)

Even if people wanted to explore, experience and be entertained on all these sites, in an exploratory and time-consuming way, this did not include the navigation. The subjects became frustrated as soon as anything went wrong in the navigation.
These examples from the study highlight the importance of considering the relation between ease of use and fun of use.

Another finding from the study concerned the importance of regarding intervention from experimenters when evaluating fun in relation to web usability. The level of intervention on the part of the experimenters was intentionally varied

in sessions in the study. Using a scale from 1 to 5, where 1 is no intervention, the tests on average rated around 3. The varying of the level of intervention followed two patterns; (1) Situated level, and (2) level set in advance. Here, the first situated approach was much more reliable in obtaining useful data. Setting the level in advance only disturbed the test in an unnatural way, no matter what level was set. Having fun on one's own can be embarrassing when two silent people are watching. Instead, if the experimenters smile with the subject and ask follow-up questions the whole situation is regarded as more natural by the subjects, as the results from the interviews show. This was clear at least for the subjects tested singly. Pairs required less intervention and the experimenters could follow the more traditional Usability Engineering type guidelines for the level of intervention in these cases. One example of this is to give firm guidance about difficulties not being tested, such as hardware problems.

Overall, being successful in evaluating fun in the context of web usability requires experience on the part of evaluators. It is easy to spoil results in this type of evaluation by making mistakes about intervention. This might be an argument for using a larger number of subjects when evaluating fun than when evaluating traditional web usability. If a larger number of subjects is used, mistakes regarding intervention in some sessions, with resulting possible breakdowns, could to some extent be rectified.

The results regarding the conditions explored in the study are discussed below.

Pairs vs. singles

Entertainment is well suited to being tested in pairs, as entertainment and exploration lend themselves to group activities. Some of the single sessions were rather slow, due to the lack of conversation, and it was obvious that the user was influenced by being in an evaluation session as opposed to an authentic use situation. The test situations conducted with pairs of users went very smoothly, and seemed natural to the subjects. In the interviews after the use sessions, the subjects were asked about the fact that they had been working in pairs and all commented that this was positive. Some also commented that it was more natural to discuss interaction with the web site with a collaborator, than 'thinking out loud' as in traditional evaluation sessions using the technique, where single users are evaluated.

> *"I think it was fun to sit together…it was the opposite..to sit together and talk is more natural than thinking aloud [by myself]..this is more natural"*
> (Subject 9 & 10, collaborating at the web site of ESC)

This aspect is also connected with the level of intervention from the experimenter. If tested in pairs, subjects have less need for feedback on their experiences from the experimenter. It fits well with the idea that sharing an experience with someone else makes it better. However, tendencies to 'show off', compete, dominate etc. could also be observed, especially during the tests using children.[5]

Another important aspect regarding testing in pairs is the social dimension of the session. As described above, this approach to evaluating EWSs facilitated the think-aloud aspect. However, it is important to identify the trade-off in these sessions between facilitating the think-aloud aspect and what is actually evaluated. In sessions with pairs, another dimension is added to the use of EWSs and that is the social dimension, the interaction between subjects. When analyzing results it is important to consider this trade-off, since in real life the web sites are not necessarily used in pairs.

Structured vs. unstructured user tasks

As stated earlier, three approaches were used; structured user tasks, unstructured user tasks and a mixture of the two. This resulted from the fact that many researchers have pointed out the difficulties of giving subjects well-defined assignments when testing web sites in general and especially this type of site. Real-life use of EWSs is usually of an exploratory kind, i.e. users explore the web site rather than search or navigate it to find information. This might be an argument for using an unstructured approach in evaluating EWSs. However, findings in the study show that an approach with structured user tasks might not be such a bad idea after all depending on the type of entertainment included in the evaluated EWS. For instance, one approach used in the study was giving structured tasks to subjects in order to test features where subjects create and send postcards, or where they mix their own song. This approach was regarded as successful as exemplified by a quotation from one of the subjects in the study:

> *"It was good to use tasks, so I knew what to do…it is easy to get stressed [in test situations]and this was better…also, I would have tried the same [things/ features] if I had used it [the web site]on my own.."*
> (One example of an answer, given by a subject in the ESC study, to a question in the post-test interview, about the given tasks)

For some subjects, the experience and entertainment took over and they became immersed and therefore forgot the test situation[6]. In this case the test was successful without the use of structured tasks – there was no need to assign tasks to such subjects. However, in some of the tests the subjects performed the task as

quickly as possible as if there was a time constraint. In this type of test, structured user tasks did not provide any support at all in testing entertainment web sites.

These situations also occurred when the mixed approach, with both structured and unstructured user tasks, was used. When new circumstances suddenly appeared, such as moving from given tasks to being recommended to perform free surfing, the subjects became confused.

The opposite order worked better, but it became clear that for some sections of the web sites, structured user tasks constituted a poor choice of evaluation technique. Once subjects were asked to perform structured tasks, they worked hard against the clock instead of spending time as they probably would have in a non-test situation. In some circumstances, such as passages in games or riddles, subjects often wanted to do some exploration and not get it right at once. Traditional usability evaluation here would just spoil the fun. The levels of support from experimenters and the presence of built-in keys in web sites were critical. Too much or too little help would spoil the entertainment. There were numerous examples of this.

As in the case of trade-off between social dimensions and facilitating think-aloud in sessions, there is also a trade-off in using structured versus unstructured tasks in test situations in the context of EWSs. This could be described as a trade-off between providing a use setting as natural as possible and gaining valid results. When an EWS is structured and designed in an exploratory way, this also argues for an exploratory approach when it comes to evaluation, i.e. unstructured tasks would be considered the first choice. However, in any evaluation of web usability some questions are considered to be too important to answer on the basis of results from evaluations. If a study is designed to be completely based on exploration and unstructured tasks only, there is always a risk that specific questions will not be answered.

Testing children vs. adults

The study included a large number of evaluations involving children because the target groups of the web sites evaluated were children and teenagers. As mentioned above, experimenters were rather reluctant to include children in evaluations, which led to the exploration of this condition – testing children vs. adults. After the tests, the experimenters involved in testing children were very enthusiastic. Testing entertainment in the context of web usability with children thus proved to be a successful approach. Compared to adults, children exhibit a different pattern when testing usability and some aspects of this were of specific significance in dealing with entertainment. The children seemed to 'play around' with the site, and explore interfaces more frequently than the adults included in the

study. This points to one advantage in that as children need fewer incitements to explore less intervention is needed. It was more common for children to become immersed in the current activity than adults, who were less 'carried away'. In some ways, children can be more spontaneous, as when a pop-up message states they are wrong - without giving any correction. This happened in one situation in the evaluation of the edutainment web site Total Defence, where a pop-up message stated that the child could not explore information to complete a specific task. The child responded out loud:

> *'strange… how am I to learn then'.*
> (Subject, 9 years old, testing the Total Defence web site)

Such spontaneity was more rare among adults, who generally appeared to be 'better behaved' culturally. However, adults were able to verbalize their actions more easily than some children who became completely silent during the test. When adults were tested in single sessions, the tests were generally more successful in that they produced understandable results for the experimenters. This was not always the case with single children, who often fell silent when faced with difficult situations. Some of the children also became afraid during the tests. One child felt so afraid that the test had to be interrupted while the teacher was asked to come and sit in on the test session. However, in this specific case the session could be completed successfully after the arrival of the teacher.

Results regarding design emerged, especially when testing the children, i.e. the true target group of the sites. For example there were problems with the level of language, the complexity of some tasks and some basic assumptions regarding experience in free text search meant that a number of tasks and scenarios made it almost impossible for children to cope with the site. In this regard it is vital to emphasize the importance of designers having a general knowledge about the experience of the target audience, especially when it comes to children. Developments in knowledge, skills and experience in general vary a great deal between children aged 7 and those aged 14, and it is essential to bear this in mind when designing for this group.

A final comment concerning testing children concerns the context of the usability tests. The experimenters soon realized that the children answered their questions with the idea at the back of their minds that if they responded negatively to the questions there would be no more exciting computer experiments in the school, ever again. Some children also made comments like:

'Oh, I missed again – do I not get a ticket to the cinema now…'.
(Subject, 9 years old, testing Total Defence web site)

Thus, social aspects must also be taken into consideration when testing children.

Written vs. oral answers on questions regarding entertainment.
As mentioned, the tests contained different types of questions to which the subjects answered in writing or orally. The pre-test questionnaires and some of the post-test questionnaires were answered in writing. In the think-aloud test and the interview, which in some cases was the post-task questionnaire, the answers were given orally. The subjects currently had no problems answering questions about background information or about errors, mistakes, effectiveness and other more common questions related to traditional Usability engineering. The results were similar to those found in other studies, and guidelines for these questions can be found in the literature (c.f. Nielsen, 1993). However, the situation was more complex when questions were asked regarding experience, fun, entertainment and etc. One question was used specifically to get the subjects, in a way, to free their minds. The question was:

'If you were to describe the site as a person (or as a car) and you had to describe this person to a friend, how would you describe this person?'[Later in the evaluations 'a car' was also used]

This question functioned very well orally and a very rich picture of how the web sites were regarded by users was obtained. The subjects needed a little help with follow-up questions from the experimenter before the answers came, but when they did, the result was outstanding. In the interview below the subject is asked the question but with reference to a car.

S: "I would say it is a BMW! A clean BMW."
E: "A BMW – why is that?
S: "It is luxurious"
E: "Is that significant in how the web site appears to you?
S "Yes, and it is red"
E: "Why?"
S: "It is cool too…"
(Discussion between evaluator (E) and a subject (S) in the ESC study)

Using the oral approach definitely led to a revelation of the character of the site. In contrast, in the questionnaires this question, and others concerning entertainment, fun, experience etc., were answered very quickly and briefly and most of the answers were quite difficult to relate to or fully understand. Some examples of written answers to the same question as above are presented below:

"Fun, chaos person, shallow, no deeper character."
(Subject, 23 years old, testing Mosquito)

"High on drugs, dizzy, believes everything is fun…"
(Subject, 23 years old, testing Mosquito)

"A clown"
(Subject, 24 years old, testing Mosquito)

"Like a studio 54 disco person"
(Subject, 23 years old, testing Mosquito)

These quotations show that it is difficult to come to an understanding and interpretation of how the subjects characterize the web site from written answers. The level of interpretation needed is so high that the results might be irrelevant. The experimenters simply did not understand what the subjects meant. One guideline when testing entertainment is that questions and answers should be administered orally if possible, in order facilitate any necessary follow-up questions.

Discussion

The evaluations of EWSs using empirical usability evaluation methods in the study presented above produced a rich source of information for how to conduct such evaluations. Overall the tests showed that even if, in experiments, traditional usability evaluation methods may be too structured and, in a way, too interventionist for the subjects to have as much fun as they would in a totally authentic situation, they still have a great deal of importance in the design of EWS. To some extent, the results came as a surprise to the experimenters. Initial assumptions about the applicability of the traditional methods included in this study, were that these methods would only give feedback on traditional usability problems. This was not the case. Initially, the plans within the research group were to develop and try out totally new approaches to empirical evaluations of aspects of fun in relation to web usability. The main reason for using traditional methods instead was that

no empirical evidence could be found in related research at that point in study, to use as basis in favour of developing new approaches should be developed. This early assumption proved wrong after this part of the study and instead the focus was placed on how to revise and refine the traditional methods. For instance, *how* to conduct a Think Aloud Protocol, *how* to write and ask interview questions concerning entertainment and fun and *what* important aspects to include in pre-test questionnaires. The subjects in the first part of the study taught us a lesson, and made us reconsider the design of the rest of the study, i.e. we decided to use traditional methods instead of inventing totally new ones for evaluating fun.

The next chapter deals with the first part of the evaluation using traditional inspection methods. The experts used were all experienced in HCI-related issues, as well as in usability evaluation. This part of the study is to be compared with that in the following chapter, Chapter 7, where the same study design was used, but the evaluations were conducted by so-called novice users, i.e. students taking an undergraduate course in HCI.

Footnotes

[1] High-resolution screenshots from the web sites are presented in Appendix III. Live versions of the web sites are also provided on the CD-ROM, attached to the thesis.
[2] For a complete description of the questionnaire – see Appendix I
[3] For a complete description of the scenarios – see Appendix I
[4] For a complete description of the post-test questionnaires – see Appendix I
[5] No quote is used, to back this up as such things are very delicate. Instead this aspect was an interpretation by the evaluators of situations which occurred repeatedly in evaluations of the Total defence and Mosquito web sites using children.
[6] No quote is used, as in the situations when this aspect occurred, the audiotape was empty. However, these situations were analyzed and it emerged that the subjects became so immersed that they forgot to talk out loud.

Chapter 6

Using traditional inspection methods – experts

Background

The next step in the approach to investigate how to find ways to improve traditional methods to better suit evaluations of fun and entertainment in the context of entertainment web sites was to conduct evaluations using traditional inspection methods (IM). The general aim was to explore the applicability of this type of method in the evaluation of fun and entertainment in relation to web usability.

Heuristic Evaluation and Design Walkthrough are the methods included in this phase of the study. These methods were chosen because they can be considered good examples of two different genres of inspection methods. Heuristic Evaluation uses a list of guidelines, or heuristics, in order to structure and guide evaluations, and can thus be seen as a somewhat structured inspection method. Design Walkthrough on the other hand is known to be one of the freest and unstructured inspection methods in existence in the area of HCI.

Experts applied these two inspection methods to evaluation of web sites. In this part of the study two web sites were evaluated; one information retrieval web site was chosen as a control site and one entertainment web site. The selection of the entertainment web site was based on the list of features discussed earlier. The chosen web site 'scored high' on these features, i.e. a large number of the features could be found in the web site. The target group for the web site was people aged between twenty (20) and fifty (50) years. All the experts fell within this target group.

The design of this phase of the study was then repeated once more with only minor changes, using other types of experts, so-called 'novices'. The novice expert phase is described further in Chapter 7.

Method

Below, the experts, material, design of this section of the study and the procedure used in the study is described in more detail.

Experts

Experts were recruited, as already mentioned, from among colleagues in the Department of Informatics at Umeå University, Sweden. Since the profile of the experts has an impact on judging and analysing the results from the expert evaluations and interviews, a thorough inquiry into the experts was conducted before any evaluations were made. Information obtained regarding the experts included gender, age, type and level of contact with the research discipline of HCI and with Usability Engineering. The experts were also asked to judge their level of experience in evaluation in general, use of Heuristic evaluation and Walkthrough evaluation, evaluation of web sites in general and finally in their use and evaluation of entertainment web sites in particular[1].

Materials – web sites

Two types of web sites were evaluated in this part of the study, one so-called, information retrieval web site (IRWS) and one entertainment web site (EWS). The reason of this choice was that the former could be used as a control site. A control web site was used in this phase of the study, because of its within-subject design, as described earlier. If no control site had been used, it would have been difficult, if not impossible, to determine the extent to which reported problems in evaluation methodology on the part of a specific expert applied to evaluations of EWSs only, or whether these problems occurred in the evaluations – by the same expert – of all types of web sites.

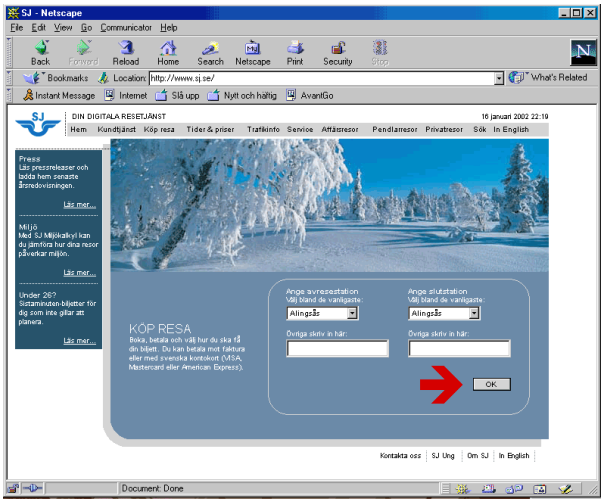

Figure 6.1 A screenshot from the homepage of the SJ (Swedish Railways) web site

Swedish Railways (SJ)

The information retrieval web site chosen for this study was the 'SJ' (Swedish Railways) site. The 'SJ' web site contains information about various areas of business within Swedish Railways, e.g. timetables, prices, routes etc. The web site was chosen as a good example of a traditional IRWS. It is reasonable to suppose that the main activity of visitors to such a site is first and foremost information retrieval. A screenshot of the home page of the web site is shown below:

Skyscraper

The entertainment web site (EWS) chosen for evaluation in this phase was 'Skyscraper'. It was a part of the Paregos corporate web page and they considered it a design project where new ideas and concepts were developed and tested. It aroused a lot of media interest when released and won prizes in 2001 such as 'Utmärkt svensk form'[2] ('Excellent Swedish Form') and first prize in "Webbspelen' (Web games). In the latter case the jury's motivation was:

> *"This site stretches the limit of what is feasible. By serving a mix of creativity, design and technique in a playroom of highest international standards the visitor is offered a top rate experience."* (http://www.Paregos.com)

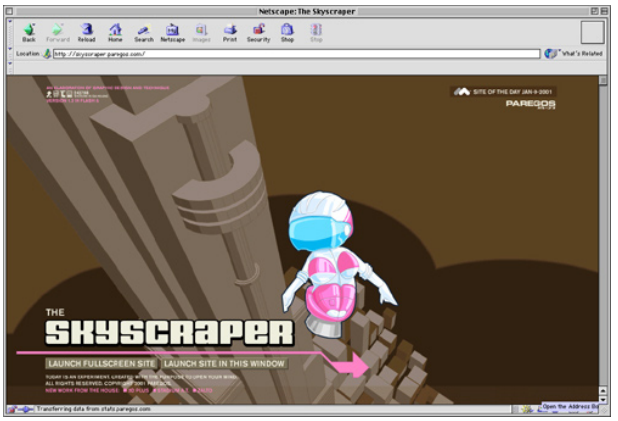

Figure 6.2 A screenshot from the homepage of the web site 'Skyscraper'

The site was used as an object of study in this phase because it contains a wide range of entertaining themes and features, based on the list of typical features included in EWSs, described in detail in Chapter 3. It was thus regarded as a good example of an EWS. A screenshot from the web site is shown below:

Design of the study

A group of ten (10) HCI experts were given the task of applying various kinds of inspection methods traditionally used when evaluating entertainment web sites. Experts reported their results both in the form of usability problems found and comments on the methodology itself. The objects of study in the evaluations were two web sites, one traditional, so called, information retrieval web site and one entertainment web site. Individual interviews with the experts were subsequently conducted to discuss the results and arrive at two main categories; (1) comments on application of the traditional inspection methods on EWSs in relation to IRWS, and (2) suggestions for changes to the traditional approaches and for new solutions.

Each expert received extensive written documentation including steps to work through[3]. In the documentation, the order in which the web sites were dealt with was switched in 50% of the cases to balance the so-called sequence effect (c.f. Solso, 1999). In other words, to eliminate problems related to the order in which the web sites were evaluated, as the same evaluation process was used in both cases. A general overview of the process is presented below:

- *Introduction* – background information about the complete procedure and brief description of the web sites to be evaluated
- *Questionnaire* – including questions about background information and earlier experience in evaluation with inspection methods on the part of the experts.
- *Evaluation of web site 1*
    - Evaluation by *Design Walkthrough*
    - Evaluation by *Heuristic Evaluation* (Heuristics by Jacob Nielsen)
    - Comparison of both approaches and suggestions for new heuristics
- *Evaluation of web site 2*
    - Evaluation by *Design Walkthrough*
    - Evaluation by *Heuristic Evaluation* (Heuristics by Jacob Nielsen)
    - Comparison of the both approaches and suggestions for new heuristics

The documentation served two purposes: (1) to supply information and instructions concerning the entire process of evaluation, and (2) to work as a form for the experts to complete by making comments on all the included steps. The documentation was tightly structured, with steps to follow and assignments to complete in order to provide conditions as similar as possible for all the experts and facilitate analysis of data obtained. This documentation was later used as a basis for interviews conducted with the experts, where both the results reported from the evaluations of web sites and comments about methodology were discussed.

Procedure

As mentioned earlier, 50 % of the experts evaluated first the IRWS and then the EWS and 50 % vice verse to counteract the risk of sequence effect. In all cases the order within the methods was the same, i.e. first the Design Walkthrough was used, followed by Heuristic Evaluation. The main reason for this approach was to counterbalance the influence of the heuristics included in Heuristic Evaluation for the Design Walkthrough evaluation. It was regarded as important that the walkthrough evaluation should be conducted without input from any specific list of general guidelines, but as an overall free exploration of the web site.

In the first evaluation using the Design Walkthrough evaluation method, the experts received a 'checklist' of key features in the web site to cover. This was done to ensure that the experts conducted similar evaluation processes, as they had differing levels of experience in using these evaluation techniques. The experts were also encouraged to write down problems and other comments in the documentation. In the documentation about Design Walkthrough, it was

emphasized that the method is an evaluation method with no specific heuristics. Instead, it is a free and individual 'walk' through the web site, where the expert was welcome to comment in writing both positively and negatively. The time limit for this part of the evaluation was estimated at 30-40 minutes.

In the second web site evaluation a brief description of the rationale of the Heuristic Evaluation method was supplied and the experts received Jakob Nielsen's list of heuristics. They were encouraged to note down any new problems and comments, in addition to the problems and comments already described in the walkthrough evaluation. The experts were also asked to link, if possible, each comment with a related heuristic from Nielsen's list. If a problem or comment was noted which had no corresponding existing heuristic, it was also to be included in the documentation, and designated 'Other'. The documentation emphasized that more general, and perhaps also positive, comments were welcome in the annotation, not just problems. The estimated time taken for this part of the study was 30-40 minutes.

After the evaluations using both methods, the experts were asked to make a comparison of the results between the two. The comments made in the walkthrough were to be marked with the number of any corresponding heuristic, if possible. This produced a situation where the experts had a number of comments in the documentation from the two evaluations marked 'Other'. They were then requested to suggest possible new heuristics, based on these comments. The above process was repeated in the both evaluations of the two web sites.

As mentioned, the documentation was used as a basis for interviews conducted with the experts after completion of the evaluation. The main themes covered in the interviews were reported problems and comments about the evaluated web sites, the evaluation process in general, the suitability of the two approaches in each case and the heuristics proposed by the expert. The length of interviews lasted about one hour.

Results

There were two types of results obtained from this phase; (1) the experts' suggestions for new heuristics to be used for the evaluation of entertainment web sites, and (2) the experts' comments on other, more general, types of problems they found when evaluating the web sites. In the latter case, the experts were also encouraged to propose suggestions for how to conduct evaluations of this kind, more generally, on entertainment web sites, based on their experiences from the above evaluations.

Below, the experts' suggestions for suitable heuristics for evaluating EWSs are given and each heuristic was marked with the number of the expert who suggested it. This list of heuristics is followed by the general comments on methodology for evaluating EWSs, as described by the experts.

Suggested heuristics for evaluation of entertainment web sites

The list below comprises suggested heuristics from the evaluation forms and includes every suggestion given by the group of experts in this phase sorted according to the number assigned to each expert (1-10[4]). In addition, the author of this thesis has commented on her reporting of the suggested heuristics. Some are 'directly reproduced', others have been 'revised to understand context' or 'revised because of length'. Each suggested heuristic is marked with the relevant comment. The table below contains a complete list of the suggested heuristics in this phase.

Expert	Suggested heuristic Description
1	Contribution of the metaphor Decide if the contribution of the underlying metaphor should be structural or content-related. (Directly reproduced) Visual impression vs. expectations Do not let the visual impression create expectations the interaction cannot meet. (Directly reproduced)
2	No new heuristics proposed
3	No new heuristics proposed
4	Product versions For this type of web site it is important that the choice of browsers and plug-ins is thought through. (Partly revised to understand context)
5	No new heuristics proposed
6	No new heuristics proposed
7	Exploratory design The design should invite the user to explore the web site. (Directly reproduced)

	Playability - gameplay It is important to clearly visualize the gameplay. Otherwise there is a risk that the user will feel cheated in that he/she had expected to be able to do other things than what is actually the case. (Partly revised to understand context) Durability Is there enough content for a longer 'visit' or 'stay'. (Directly reproduced) Clarity of genre The web-site target group should be clear, regarding e.g. age, special interest(s), event etc. in order to avoid misunderstandings about quality. (Partly revised to understand context)
8	Animations Animations which show directions (for instance flying packages, boxes etc.) should correspond to functionality. (Directly reproduced) Personal (re)connection It is important to keep the user inside the web site, independent of activity. Avoidance of a situation where the user is 'thrown out'. Possibility also to assign user an identity which can be used if user gets thrown out or voluntarily chooses to re-enter the web site. (Partly revised to understand context) Fast feedback As interaction is often fast on this type of web site, it is important to have clear text and quick feedback on information. (Partly revised to understand context) Support of social navigation Support of 'social navigation' should be provided, i.e. visualizing of other users on the web site. It enhances the interest as the user seeing that there are others either simultaneously present or who have been there before. (Partly revised to understand context)
9	Affordances / Visability of objects All elements should clearly show their status in relation to the environment, or what they do. (Directly reproduced) Icon clarity Icons should unmistakably indicate what they stand for. (Directly reproduced) Support of models The designer should facilitate the creation and understanding of mental models for how functions are connected, or what they are doing. (Directly reproduced)
10	No new heuristics proposed

Table 6.1 Overview of suggested heuristics by experienced experts.

More general comments about methodology for evaluating entertainment

Comments and implications for improvements in inspection methods for evaluating EWSs which emerged in the interviews, are given below. This section is structured as follows: The comment or implication as given by experts is marked 'implication'. The implication is then contextualized by empirical evidence from the interviews. This evidence is presented as 'background' and finally, the methodological changes made before further investigation in the following phase of the study are presented and labelled 'solution'. Generally, the comments the experts made in the interviews constitute input into the overall evaluation methodology investigated in the study, concerning e.g. design of documentation, steps included in the whole evaluation process, etc.

1. Possibility of creative feedback when evaluating with heuristics

Implication: Opportunity should be given to provide positive as well as negative feedback when evaluating EWSs using the Heuristic evaluation method. The documentation should also provide opportunities for further motivating comments about the interface for each heuristic.

Background: In the first part of the study, the documentation handed out to the experts was designed in such a way that it was understood to ask only for problems, and no positive comments. This was not completely true, as some parts in the documentation encouraged experts to give both positive and negative feedback. However, this was obviously not stated clearly enough, and some of the experts felt 'trapped' in a process where only problems were allowed.

> *"I was somewhat confused by the labelling 'problems' – I was evaluating fun not only function. It was OK, but perhaps it could be clearer."*
> (Expert 8 – Interview after the evaluations in this part)

Solution: To solve this, the evaluation documentation to be used in the third phase of the study, presented in Chapters 11 and 12 in the thesis, was re-designed to highlight the fact that both positive and negative feedback was allowed and were equally important.

2. Importance of a high level of freedom when evaluating fun

Implication: In the context of evaluating 'fun' it is important that (at least) one part of the evaluation could be regarded as 'free', i.e. where no or only minor obligations are to be met. This is very important in the context of fun because of its exploratory nature.

Background: The evaluation process overall, with the experts required to follow 'to-do' lists, as in the Design Walkthrough evaluation, and lists of heuristics, as in Heuristic evaluation, was experienced by many experts as strictly *evaluation* with very few or no similarities to an authentic use situation. Even if Design Walkthrough is considered a free approach, the fact that suggestions were given concerning activities, e.g. features of the web site to be explored, resulted in even this part of the evaluation being seen as strictly evaluation as against real use. Experts expressed the feeling that they were in a state in which they were 'coded to search for problems' when describing their evaluations. This may imply that there is a need to tone down the focus on 'evaluation as process' in the overall evaluation methodology.

> *"When I evaluate, I am coded to find problems and I cannot escape the fact that it is an evaluation."*
> (Expert 9 – Interview after the evaluations in this part)

> *"Evaluation for me is focusing on problems…as soon as I am involved in an evaluation I search for problems"*
> (Expert 10 – Interview after the evaluations in this part)

Solution: To tone down the sense of 'pure evaluation' for the experts, the Design Walkthrough was replaced with another type of approach, called in the context of this study 'free surf' approach. This change in the evaluation methodology also limited the time experts spent evaluating, since time spent in a 'free surf' approach can more easily be restricted compared to the case when a 'to-do' list is delivered to experts in evaluations, based on Design Walkthrough method. In order to suggest a level for what might be considered as 'acceptable performance' by the expert, a time limit of 20-30 minutes was set. After this, the expert was asked to answer three brief questions about their thoughts on to what extent they as a person could be regarded as fitting the target group, about the approximate time they spent on the web site in the 'free surf' and finally, their estimation of how much of the web site they explored.

3. Reviewing rather than evaluation

Implication: Evaluation of entertainment should be seen rather as 'reviewing' than making an 'evaluation'. The reason for this is twofold: (1) The idea of 'evalution' *in itself* has a negative implication, at least for some people. Evaluators think of themselves as problem finders. In the context of fun this is awkward as many comments might be positive but are not further expressed because the evaluator

feels they are of less importance in 'evaluating' a web site. (2) Some parts will get lost when certain aspects are abstracted from a larger context and narrowed down to a detailed level, as happens when usability problems are reported in evaluations. This is true for all types of evaluations, usability or others. However, in the context of entertainment this becomes even more problematic, as an overall impression might differ a great deal from the stated comments. Therefore, it is important to allow the expression of an overall judgement especially in the context of entertainment.

Background: In one of the interviews, one expert expressed these ideas but as of s/he was the last to be interviewed in this phase, they could not be discussed further with other experts. However, the expert expressed ideas about this and gave two thoughtful examples: (1) In academic journals and conferences, the proposed papers are subjected to a review process, where reviewers are given some guidelines (or 'heuristics') as a basis for the review. However, on the basis of these guidelines, the reviewer makes an overall judgment, a review, of the paper as a whole. Even if the author of the paper does not meet one or two of the guidelines, an overall impression of the paper might still be positive. (2) When making judgments about a film, it might be true that some parts, such as the sound or the lighting in some sequences might be considered bad. If a strict list of problem-based heuristics were to be followed in an evaluation of the film, it would be judged problematic, or even 'bad'. However, the overall impression of an audience might still be positive. It might be thought 'a wonderful movie', 'a romantic story in stunning natural environments' or anything else. Here, an 'evaluation' is useless, but a 'review' is more fruitful as a judgment model. The same expert stated:

> *"I have thought about this – evaluations. Reviewing might be considered instead, as this differs from evaluations in that evaluations focus on problems – reviews reveal both positive as well as negative aspects. The fun factor is judged [reviewed] – it is not evaluated. Compare with a book or film review – Even if the language may be poor, this may be mentioned in one sentence in a review. In a heuristic evaluation, every detail of the poor language would be mentioned "*
> (Expert 10 – Interview after the evaluations in this part)

Solution: No specific solution to this problem was proposed in the subsequent phase of the study. The reason for this was a need to restrict changed conditions between different phases of the study, in order to be able to link changes in results to the right condition. In this specific case, the evaluators wanted to find out if the main problem in this phase was instead the heuristics used that caused the

identified problems and not the contradiction between reviewing and evaluation. Further, evaluators wanted to somewhat restrict the number of changes between phases to be able to follow-up the results. If a new set of heuristics would not solve the problems as expressed by this expert, a more explicit solution regarding evaluation vs. reviewing would be proposed for the very last phase of the study.

Discussion

This first phase of the study of Inspection Method evaluations of EWSs and IRWSs produced numerous indications that this type of evaluation is also applicable when investigating entertainment. Results clearly showed difficulties arising, for instance, from the heuristics used. Results from this phase provided clear evidence that the results of evaluations of EWSs are highly dependent on what set of heuristics is used. It emerged that the list of heuristics proposed by Jakob Nielsen was of little or no use in the context of entertainment. This might come as no surprise. However, what is presented here are empirical findings that actually show what was wrong combined with suggestions for alternative heuristics. Furthermore, the results from this part of the study produced a methodological input regarding *how* to consider heuristics. For instance, the heuristics might be seen more as a basis for comments rather than laws or rules allowing for only complaints and the stating of problems.

The suggested heuristics, as described above, were the subject of further investigation and analysis, in order to estimate their validity for the evaluation of entertainment. This was done as follows: the list of suggested heuristics, as presented above, was combined with the outcomes from a study with similar design as the one presented above, but conducted by a group of experts called 'novices'. That study is described in detail in the next chapter, Chapter 7. The process of the combination of the suggestions into a new list of heuristics, i.e. the analysis process, is further developed in Chapter 8. The outcome of this analysis, i.e. the list of proposed heuristics for evaluation of EWSs, is also given in Chapter 8.

Footnotes

[1] The complete questionnaire and the results of the inquiry concerning the experts are presented in Appendix II.

[2] The Swedish Society of Crafts and Design deliver the award of 'Excellent Swedish Form' every year.

[3] This documentation is further described in Appendix I.

[4] The experts are assigned the same number in all of the following phases of the study as well, i.e. 'Expert 1' always refers to the same person throughout the study.

Chapter 7

Using traditional inspection methods – novices

Background

As in the phase where experienced users applied inspection methods in the evaluation of EWSs, this phase also describes the application of such methods but this time by a group of so-called 'novice users'. The aim in this phase of the study was also to investigate the applicability of traditional inspection methods to usability evaluation, on the basis of their applicability in evaluations of entertainment and fun in EWSs. The design and procedure of the study were basically the same. The novice experts were asked to use exactly the same inspection methods in evaluations of the same web sites as the experienced users and were given the same documentation. There were, however, two differences between this and the preceding part: (1) the difference in the level of expertise in usability evaluation in general between the experienced and the novice expert groups, and (2) the fact that this part of the study included additional assignments for the experts.

The experts used in this part of the study could be called *novices* in that they were undergraduate students, taking a course in HCI in the final part of their degree course. The evaluations of the web sites were included as a compulsory assignment within the course. The novice experts worked in pairs. This type of experts was chosen in order to see what would happen if inspection methods were used by those with no previous experience. The students were given additional assignments out of a curiosity to test other inspection methods and their applicability to entertainment. Each group chose an additional inspection method and used it to evaluate an entertainment web site of their choice, i.e. in all ten additional EWSs were evaluated. The results from these additional assignments, given to the group of novice experts alone, were seen only as complementary in this phase of the study. This will be further developed below.

Method

As mentioned, the study in this phase of the study was carried out in a similar way to that in the second part of the study. There is no further description of the similar procedures, as these are described in detail in Chapter 6. Those aspects and procedures specific to this part of the study are explained below.

Expert groups

The expert group comprised twenty (20) students, divided into ten (10) groups with two students in each group. The students were free to choose their own group. The novice expert teams were asked to complete a background questionnaire, in which they made judgments concerning their background and experience in relation to the assignment. They made judgments *as pairs*, i.e. the data received from the questionnaires about the expert groups described the profile of the pairs and not the individuals[1].

Furthermore, the groups were also asked to judge their overall level of experience in evaluation, use of Heuristic evaluation, Walkthrough evaluation, evaluation of web sites in general and finally in the use and evaluation of entertainment web sites in particular[2]. Comparisons between the groups of experts and novices are neither possible nor fruitful in the context of this study, since it was obvious that major differences existed. The only interesting comparison to be made here is between earlier experiences *within* the two groups of experts. This comparison was made and the results are presented in Appendix II. Overall, the novices were similar as regards background, age, and experience both of usability evaluation in general as of using EWSs.

Materials – the web sites

The two web sites used for evaluations with Heuristic Evaluation and Design Walkthrough in this phase were, as mentioned 'SJ' (Swedish Railways) and 'Skyscraper'. For further description, see Chapter 6. In addition, each expert team chose a different entertainment web site. In other words, no two teams could choose the same extra EWS. The entertainment web sites chosen with corresponding WWW links and a brief description of each site are shown in Appendix II.

The results from evaluations of the additional web sites in this phase contributed to the process of defining new heuristics in the same way as for the two set web sites. However, the evaluations of the additional web sites are not further explored or discussed in this thesis, since they were considered to be somewhat less important. For this reason there is no further exploration or

description of these additional ten (10) web sites in the context of the thesis. The additional methods and web sites included in the assignment for the students were added as belonging to the content in the course they were taking, rather than as a contribution in this thesis.

Design of the study

The assignment for the students was more or less the same as that given to the experienced experts. There were also additional assignments in this part of the study, where the expert teams should: (1) go through exactly the same procedure but for an additional entertainment web site, (2) carry out empirical usability evaluations on all three web sites with a minimum of 3 subjects for each web site, (3) choose one additional evaluation method - empirical or inspection - to apply to the three web sites, (4) write an extensive report with descriptions of the procedure used, their findings and suggestions.

In that part of the study where the teams conducted the evaluations on the assigned web sites with Heuristic Evaluation and Design Walkthrough, extensive support was provided by their lecturers. Workshops were given, where the methods were presented and the students could ask questions. Throughout the evaluations, lecturers provided feedback to all groups whenever problems arose related to the conduct of usability evaluations with these inspection methods. However, the teams performed all the evaluations and presented the results without interference or involvement on the part of the lecturers. The author of this thesis was the senior lecturer on this course and had the main responsibility for providing guidance for this specific assignment.

In the second part of the assignment, where additional methods and an individual EWS were chosen, the students were given no or only limited help, mainly because this part of the assignment was considered of secondary importance in relation to results of this study. However, the second part provided valuable experience for the students, since it was less structured than the first part. In other words, the second part served to educate students in finding interesting and important research literature about usability evaluation. The students were also given experience in the design of studies and delivering results in reports. The methods chosen by the expert groups were questionnaires, Cognitive Walkthrough, scenario-based evaluation, focus group evaluation, feature inspection, heuristic evaluation based on design guidelines by Shneiderman, pluralistic walkthrough and inspection on the basis of Norman's seven steps[3].

The results from using the additional techniques listed above, i.e. their applicability when evaluating entertainment web sites, are briefly discussed in this thesis. It is worth mentioning that not all groups were successful in their additional

assignments, as some of them did not understand how to apply these additional methods in practice. In these cases there is no further discussion of that specific method in the thesis. Furthermore, it should be clearly stated that the results concerning the additional techniques are of minor relevance in the context of this thesis, in light of the overall aim of the study. Initially, a number of inspection methods were chosen, as described in Chapter 4. These chosen methods were thoroughly explored throughout the complete study, by experienced as well as novice experts, often in more than one phase. In the case of the evaluation methods chosen by the students, only one novice student team used each of these methods on a maximum of three web sites. Thus the quality of the data produced must be considered poor. Furthermore, it was obvious that in many cases the students had not succeeded in carrying out this assignment, because of time constraints or lack of experience in using the techniques. In some cases, however, interesting findings did emerge. Nevertheless, throughout this phase of the study, in which teams evaluated web sites using inspection methods both assigned as well as new chosen suggestions for Heuristic Evaluation are still the main focus in the results reported below.

Procedure

The core assignment was the same as that in the study with experts, as mentioned. In the Design Walkthrough evaluation, the experts received a 'checklist' of key features within the web site to cover. They were encouraged to note down problems and other comments in the documentation. The Heuristic Evaluation process was briefly described for the experts, they received the list of heuristics devised by Jakob Nielsen, and they were also told to write down any *new* problems and comments, and mark the comment with the relevant heuristic. It was emphasized that more general, and perhaps also positive, comments and not just problems were welcome in the documentation, and should be marked 'other'. After this, a comparison between the results of the two evaluations was conducted, and comments made in the walkthrough were marked with the number of any corresponding heuristic. Finally, the experts were asked to suggest new heuristics, based on statements, which they could not connect to existing heuristics.

From this point onwards, the process in the two expert evaluations diverges. The novice teams, like the experts, documented all their findings and suggestions in the handouts. However, the novice teams were also asked to write an extensive report about their findings. This report provided the basis for an oral presentation of their work, in front of the rest of the student groups and the lecturers. As mentioned, the author of this thesis was one of the lecturers. These presentations focused on the extra entertainment web site and general findings regarding

usability problems and other comments and finally on the methodological part of the assignment, i.e. how the additional method worked and what new heuristics the team proposed.

All written material was collected and the presentations were recorded for further analysis.

Results

The main results for this part of the study are given below. The types of results are (1) suggestions of new heuristics for evaluating entertainment web sites, and (2) implications of the applicability of the other evaluation methods used in this part of the study, apart from Design Walkthrough and Heuristic Evaluation.

Suggestions of new heuristics from novice teams

Team	Suggested heuristic Description
1	Feeling of movement Measure the level of ability to enter into the interface connected to control of movement. User should experience that he/she 'is' the one portrayed in the interface. The movement in the spatial dimension should be experienced as synchronous with reality. (Directly reproduced) Color mediation of feelings Does the choice of colors reflect the mood mediated, as intended by the designer? Awareness of colors effects a user's opinions about the system, for instance for exploration and remaining at the site. (Directly reproduced)
2	No new heuristics proposed
3	No new heuristics proposed
4	General impression This is not only about graphic design, but also what the web page has to offer and how this is shown and presented. Does the web site catch the users attention and curiosity and generally make a strong impression. (Partly revised due to length) Suggestions instead of problems Instead of just looking for problems, the user/evaluator often notices things that could be improved. The aim is to elicit creative thinking from the user/evaluator.(Partly revised due to length)

	Experience In some way, the level of experience should be considered in the evaluation of entertainment. Regardless of the scale used, the user should be able to value his/her subjective experience. (Partly revised due to length)
5	No new heuristics proposed
6	Create and fulfill expectations On entertainment web sites it is important to create an expectation, for instance by providing a 'fancy' or attractive design and suitable music. However, it is also important that the expectations created are fulfilled. If this does not happen the whole impression of the web site is spoiled, and the user might feel disappointed or cheated. (Partly revised due to length) Balance between information and entertainment If entertainment is the goal of the web site, it should strive to achieve a balance between entertainment and information. We mean that [entertainment] web sites are often so-called, information containers for banners and entertainment material, e.g. multimedia content etc. However, there is a limited total amount of content for a web site, and in order to achieve a good result. There must be balance. It is also important that the information is relevant and 'fresh'. Otherwise the feature has no reason to be on the web site. (Partly revised due to length)
7	Relevant content and lifetime The content must be relevant to the purpose of the web site. For instance, in order to entertain for more than just 'the moment', the entertainment parts of the web site should have a longer life. (Directly reproduced) Originality and uniqueness The web site must provide something new and unique. Otherwise there is no reason for a user to revisit this particular web site. (Directly reproduced) Design for the target group The web page must be designed, i.e. configured and filled with content, relevant to the intended target group. (This also means that it is important to find out what target group really is the receiver.) (Directly reproduced)
8	No new heuristics proposed
9	No new heuristics proposed

10	Encouragement to explore The users should be motivated to stay on the web site and explore the possibilities it can provide. A boring and difficult design might frighten users away faster and reduce their motivation to stay. (Directly reproduced) Avoid distracting information The primary and essential information must not fade into background or be disturbed by secondary information. (Directly reproduced) Simplicity and completeness The web page should give a [good] overall impression. Secondary features should be removed from the design. (Directly reproduced) Information should be reliable and not misleading All promised information should exist, if this is stated. Otherwise, it should be stated that the information, is temporarily missing. (Directly reproduced)

Table 7.1 Overview of suggested heuristics by novices.

Results form evaluations with the additional techniques

The methods given high levels of quality of performance and data, graded as 4 or 5 in quality of practice and data, as given in the table of web sites in Appendix II, are further discussed below, from the perspective of their applicability in evaluation of entertainment web sites.

Two teams used the inspection method Cognitive Walkthrough. Their findings showed that the method, as it is designed in the standard version in the research literature, is not applicable to entertainment web sites, in the sense that it does not cover any 'fun' aspects. The baseline in the method is highly structured, i.e. there are standard tasks to be performed by users and this creates difficulties in the context of EWSs, which are often unstructured and based on exploration rather than task driven. As the user group of entertainment web sites is heterogeneous it is difficult to define a proper 'general user', which is one cornerstone in the method, in order to find the most common tasks. Overall, the method is somewhat 'digital' in the reporting of findings, i.e. 'on or off', the task *could* or *could not* be completed in the system. It is difficult to grasp the whole picture in entertainment by using this type of quality measurement. It might be difficult to say that the web site is either 'fun' or 'not fun'. A richer scale, whatever is chosen, is needed to give answers regarding entertainment and fun. The general finding regarding Cognitive Walkthrough was that it is not a fruitful method to use for evaluating entertainment web sites.

The team that had tried Focus Group Evaluation decided that this inspection method could very well be used to say something about entertainment web sites.

The students pointed to such advantages as that it was a group evaluation with a heterogeneous group, it gave rise to interesting discussions about the web site. Compared with the Design Walkthrough or Heuristic Evaluation, where in their case only two evaluators were present, the Focus Group Evaluation produced faster results as the group found problems more quickly and could find things to comment on. It also gave results concerning design implications where the group discussed alternatives for features they somehow found peculiar or less fun. The students gave the method high grades for its applicability in evaluating entertainment web sites.

Pluralistic Walkthrough is somewhat similar in procedures to Design Walkthrough. In this case the two methods differed in that the team involved in the evaluation was larger and also more heterogeneous. The practice of this method closely resembled the Focus Group evaluation, as did the results. Overall, these students found Pluralistic Walkthrough useful in the sense that it gave rise to a lot of comments about the entertainment web site, which would seem to be valuable for designers to know if they were re-designing the web site or designing other, similar, entertainment web sites.

Discussion

In this part of the study undergraduate students were used as experts. Even if this is somewhat open to question it was worth doing because the students did an excellent job and worked really hard on their assignment. When they met problems, lecturers worked as mentors to help the students over their difficulties so they could finish their assignments. The focus of the mentoring was on the main assignment, i.e. to conduct evaluations using Design Walkthrough and Heuristic Evaluation and to propose new heuristics. Overall, the students produced astonishing results, especially in their proposals for new heuristics and their commitment and creativity was admirable. Had the students proposed heuristics based on guesswork with no basis in empirical findings, the results could not have been used at all, or would have been rated as less important than the suggestions made by the experienced experts. However, in the event the students produced numerous examples from their studies and could present strong arguments to support their suggested heuristics. Thus these are regarded as having the same status as those suggested in the study with experienced experts. The results from the secondary assignment, i.e. the use of supplementary methods, however, are not developed further in this thesis for two reasons: (1) there are not enough data to be able to say that the findings are completely valid; (2) it could be argued that the experts were novices who were given no guidance or mentoring, and therefore the results are of lower

quality. In the additional assignment the students worked completely on their own without guidance from lecturers, as opposed to the first part of the assignment, where extensive documentation and supervision were provided.

In the next part, a general analysis of the findings is conducted, and a new and revised set of methods is presented.

Footnotes

[1] A complete description of the questionnaire delivered to the expert pairs as well as the results is given in Appendix II.

[2] The levels of experience were described by choosing one of a number of statements included in the questionnaire. Note here, that some of the scales do not correlate with those used in the evaluations conducted with experienced experts as the students were assumed to have other types and levels of experiences, requiring the use of other values. The complete questionnaire and the results are presented in Appendix II.

[3] A table of these methods including a judgment of their applicability, as stated by groups of novice experts, is presented in Appendix II.

Part III

Re-design of traditional evaluation methods for entertainment web sites

Part 3 includes a general overview of the process of refinement and re-design of the methods used in this study. The methods were refined on the basis of findings from the first phase of the study, described in Part 2. Part 3 includes two chapters:

- Chapter 8 – Revision and re-design of *Empirical Usability Evaluation Methods.*
- Chapter 9 – Revision and re-design of *Inspection methods*

The two chapters go into more details, and in the first chapter conditions in focus and empirical evidence for re-design are discussed in order to give a better understanding of the process of revision. In the second chapter, Chapter 9, the process of designing new heuristics based on the suggestions made by the two groups of experts is described. Suggestions for overall methodological changes are also given. Finally, the chapters end with a discussion of the overall findings from the initial study.

Chapter 8

Revision and re-design of empirical evaluation methods

Traditional usability evaluation methods were selected and applied to EWSs in the first part of study. The design of the study and the methodological findings from these evaluations are presented in Chapter 5. In this chapter, these findings are summarized and developed further. On the basis of this, a revised set of methodological implications for empirical usability evaluation of fun in the context of EWSs is presented. This process is described in this chapter.

The conditions used in the initial empirical study were: (1) evaluating in pairs vs. individuals, (2) using structured vs. unstructured user tasks, (3) testing children vs. adults, and finally, (4) collecting written vs. oral answers to questions concerning entertainment. Below, each of these conditions is linked to the corresponding methods, and an overview is given of the implications of the first phase of the study.

Results from the first phase of empirical evaluation

Think-aloud protocol

Testing pairs vs. individuals

Entertainment is well suited to being tested in pairs, as entertainment and exploration are activities suitable for groups. This approach worked well, both when testing adults and children. It is natural to experience enjoyment and fun together and using a pair design for the study also facilitated the think-aloud process. However it is important to be aware of the precise activity and interaction that is being displayed. To add a social level, as in the test design for EWSs which uses a team, to a context traditionally focused on individual users may be problematic when it comes to data analysis. It is important to try to identify and ignore those findings most likely to be connected with social interaction and to focus on the findings connected only to interaction with the EWS, in this case. This needs further investigation.

Structured vs. unstructured user tasks

Results clearly showed that the level of structure in the tasks given to subjects is dependent on the type of entertainment included in the EWS that is being evaluated. If the features to be evaluated are highly immersive, as described in the experience realms (Pine II & Gilmore, 2000) and exploratory in nature, a structured approach is insufficient. In other words, no structured tasks are needed when subjects proceed to explore these features. Examples of such cases are games, music mixing, etc. In some cases a game is the main object of study, as in the case of the Total Defense web site where the only task give to the subject was 'play the game'. Arising from this, one conclusion from the study is, that if the EWS includes mainly one feature, for instance a game and nothing else, no structured tasks are required. On the other hand, in other cases where different types of entertainment features are included in the EWS, an approach that employs structured user tasks was quite effective, even if some of the features included immersive or exploratory elements. In some cases, where a completely unstructured task approach was used, subjects reacted reluctantly and showed symptoms of insecurity. Generally, it is important to be aware that subjects may feel insecure in test sessions, and this problem must be addressed. In situations where subjects may possibly feel uncomfortable, structured tasks could be the most suitable alternative.

Testing children vs. adults

Testing the target group intended as prime receivers of the web site is a crucial aspect when it comes to evaluating entertainment web sites, even if this means using children. In some cases, tests involving children may be of little value, as this type of subject in general may find it more difficult to think aloud. However, if the level of intervention is carefully considered by the evaluators, the best practice is to test using children, if they are the main or only target group for the EWS.

Interviews and questionnaires

Written vs. oral answers to questions concerning entertainment.

In general written data is easier and less time-consuming to collect and interpret, as it is shorter and more concise. However, when the questions are about entertainment values, this type of data becomes difficult to interpret, due to the subjective nature of the object of study, i.e. EWSs and entertainment. Answers become difficult to interpret and analyze, and it is often hard to understand what the subjects mean by their statements. In many cases in the first phase of the study, follow-up questions were needed, and these are only possible in oral situations, such as interviews.

Children vs. adults

To conduct inquiries based on questionnaires and interviews where children are included as subjects is rather difficult. The questions must be kept easy to understand, and easy to answer. This has important implications for output. However, the data material in this study shows that overall interviews worked better than written questionnaires.

Sources of empirical evidence for revisions of empirical evaluation methods

In the empirical evaluation method part of the first phase of the study, the main source of empirical evidences for the revision of methods derives from observations, made by evaluators in the test team. Other sources of empirical evidence are interpretations of the data material from interviews and questionnaires. Below, the changes in the traditional methods, for further investigation in Phase 3 in the study, are presented.

Re-designing and revising methods

In the re-designing and revising of methods for empirical usability evaluation in this study, a number of alternatives were possible. Such as; (1) evaluating entertainment web sites using the same methods as in the initial part of the study, refined on the basis of the findings from the first phase in the study – aimed at checking whether the refinements made seem correct; or (2) using other types of methods for the empirical evaluation of entertainment web sites, in order to arrive at implications also for these other types of methods. In this study an iterative approach is used, i.e. the same methods are used in evaluations throughout the study, and the methods in focus are refined with every iteration. This iterative approach is used both in empirical evaluation methods as well as in inspection methods. The main reason for this was to obtain continuous verification of whether methodological changes made were reasonable, based on comparisons between earlier findings and findings of evaluations where the revised methods were used. In the case of empirical usability evaluations there are two iterations and in the case of inspection methods three. The conditions in the re-designed methods are as follows:

Think Aloud protocol

Pairs vs. Individuals

If possible, the subjects are tested in pairs, in order to continue inquiry into how to separate data connected with the social dimension of interaction from that connected with authentic interaction with the EWS.

Structured vs. unstructured user tasks

For this condition, the situation has to determine the correct approach – if games are to be evaluated, no tasks are needed. If large entertainment web sites are to be tested, and especially if the subjects may consider the evaluation situation rather stressful, structured tasks are used. The first phase indicates that this is a proper division but this needed further investigation.

Testing children vs. adults

For this condition, testing the right target group is the most suitable focus – even if this means testing children. It emerged that entertainment seems particularly sensitive regarding the sense of fit to target group, which overrides the possible problems of testing children. However, the process of testing with children and young adults needs further investigation.

Interviews and questionnaires

Written vs. oral answers to questions regarding entertainment.
Based on the findings from the first phase, oral interviews are the most successful alternative when investigating aspects of fun in EWSs, and should be used if there are no obvious reasons for using questionnaires. However, more knowledge about designing interview manuscripts for this kind of interview is needed, requiring further investigation.

Children vs. adults
In the first phase the only questionnaires children answer are the pre-task questionnaires. In all of the studies with children included in the first phase, interviews are the only method used for collection of empirical evidence after use sessions, since this seemed the only reasonable study design, as it avoided putting too much pressure on the children by asking them to complete a questionnaire. This approach is successful and will be investigated further.

In the case of adults tested in environments primarily directed to a younger audience, either questionnaires or interviews appear possible, on the basis of the findings from the first phase. However, as the object of study, i.e. EWSs, is generally very subjective in its nature, it is difficult to design questionnaires which will produce understandable results. The answers are subjective and often need clarification, more easily accomplished orally. A combination with questionnaires being initially completed and then used as a basis for oral follow-up questioning, i.e. interviews, may be a possible alternative.

Discussion

Considering that the study included sixty empirical usability evaluation sessions, the findings in this part appeared to the evaluators at first glance to be a matter of common sense. As presented above, the findings in relation to the specific conditions explored, seem superficially to correlate more or less with what any informed HCI researcher might expect, without having to conduct an extensive research process. However, when all the findings from the study are considered on a deeper level, this is seen to be not really true. Many things that occurred, not necessarily connected to the above conditions, provided worthwhile information about how to conduct evaluations of usability on entertainment web sites. For instance, the question of whether it is possible to obtain *any* input of value for designers from this type of evaluation is regarded as very important and without testing these approaches the question would be impossible to answer. Here, the

answer is unconditionally in the affirmative.

Workshops were held with the design team which had designed the entertainment web sites, and they found the information provided on the basis of the findings valuable. Returning to the heuristics for, or questions regarding, evaluating methods suggested by Khan & Prail (1994)[1], many of these heuristics would be fulfilled, on the basis of the designers' reactions. They were interested in whether any function-related problems appeared, as well as whether the users were amused by the web sites. Both of these questions could be answered from the evaluations.

Another useful insight gained from these tests is the importance of using evaluators skilled enough to react intuitively in each situation. Breakdowns frequently occurred, due to the fact that it was an evaluation that was being conducted and not an authentic use of the entertainment web site. When this happened, the situation had to be stabilized by the evaluator, whose skills increased as the study proceeded. Thus it is important to remember when conducting this type of research that it might take a number of sessions before the evaluators are experienced enough to conduct useful tests and report valuable results.

On the basis of these overall findings, the part of the study that included empirical evaluation methods may still be considered to be successful.

In the next chapter, the process of revision and re-design of inspection methods is described and discussed.

Footnotes

[1] These heuristics by Khan & Prail (1994) were described in detail in Chapter 3.

Chapter 9

Revision and re-design of inspection methods

Inspection methods are the object of study in this phase of the study. In phase 1 of the study, a number of methods were used, by both a group of experts and a group of novices. The results from the evaluations by these two groups were analyzed to determine whether inspection methods would be useful in the context of evaluating entertainment web sites. As in the case of empirical valuation methods, results showed that inspection methods were also useful in this context as becomes clear both in the reporting of the evaluation results from the web sites, and in interviews conducted with experts after the evaluations in the first phase. However, in order to be more appropriate for evaluating entertainment, the methods needed revision, and it seemed important to include additional approaches. All these requirements were the results of the inspection method evaluations in the first phase of the study.

In this chapter the process of re-designing and refining the overall methodology is presented, including an overview of the main findings from the interviews, a detailed description of the proposed new heuristics for evaluating entertainment, and a description of the analysis and interpretation process with respect to these suggestions. The purpose is to make the research process as transparent as possible.

The process of designing new heuristics for evaluating entertainment

In the assignments in the first part of the study, the experts were asked to use traditional functional heuristics, as designed by Nielsen (1993). The heuristics were used in the evaluation of two types of web sites, one information retrieval web site and one entertainment web site. After this, they were asked to suggest new heuristics, on the basis of their experience in these evaluations. Generally, both groups of experts succeeded in proposing new heuristics, and the complete list of suggestions for new heuristics is displayed below. The suggestions were subdivided into two sets from the experienced and from novice experts respectively, as given below:

New heuristics from experts

Expert	Suggested heuristic Description
1	Contribution of the metaphor Decide if the contribution of the underlying metaphor should be structural or content-related. (Directly reproduced) Visual impression vs. expectations Do not let the visual impression create expectations the interaction cannot meet. (Directly reproduced)
2	No new heuristics proposed
3	No new heuristics proposed
4	Product versions For this type of web site it is important that the choice of browsers and plug-ins is thought through. (Partly revised to understand context)
5	No new heuristics proposed
6	No new heuristics proposed
7	Exploratory design The design should invite the user to explore the web site. (Directly reproduced) Playability - gameplay It is important to clearly visualize the gameplay. Otherwise there is a risk that the user will feel cheated in that he/she had expected to be able to do other things than were actually possible. (Partly revised to understand context)

	Durability Is there enough content for a longer 'visit' or 'stay'. (Directly reproduced) Clarity of genre The web-site target group should be clear, regarding e.g. age, special interest(s), event etc. in order to avoid misunderstandings about quality. (Partly revised to understand context)
8	Animations Animations which show directions (for instance flying packages, boxes etc.) should correspond to functionality. (Directly reproduced) Personal (re)connection It is important to keep the user inside the web site, regardless of activity. Avoidance of a situation where the user is 'thrown out'. Possibility also to assign user an identity which can be used if user gets thrown out or voluntarily chooses to re-enter the web site. (Partly revised to understand context) Fast feedback As interaction is often fast on this type of web site, it is important to have clear text and quick feedback on information. (Partly revised to understand context) Support of social navigation Support of 'social navigation' should be provided, i.e. visualizing of other users on the web site. This enhances the interest as the user see that there are others either simultaneously present or who have been there before. (Partly revised to understand context)
9	Affordances / Visability of objects All elements should clearly show their status in relation to the environment, or what they do. (Directly reproduced) Icon clarity Icons should indicate unmistakably what they stand for. (Directly reproduced) Support of models The designer should facilitate the creation and understanding of mental models for how functions are connected, or what they are doing. (Directly reproduced)
10	No new heuristics proposed

Table 9.1 Heuristics suggested by experts.

Suggestions of new heuristics from novices

Expert group	Suggested heuristic Description
1	Feeling of movement Measure the level of ability to enter into the interface connected to control of movement. User should experience that he/she 'is' the one portrayed in the interface. The movement in the spatial dimension should be experienced as synchronous with reality. (Directly reproduced) Color mediation of feelings Does the choice of colors reflect the mood mediated, as intended by the designer? Awareness of colors effect on user's opinions about the system, for instance for exploration and remaining at the site. (Directly reproduced)
2	No new heuristics proposed
3	No new heuristics proposed
4	General impression This is not only about graphic design, but also what the web page has to offer and how this is shown and presented. Does the web site catch the users attention and curiosity and generally make a strong impression. (Partly revised due to length) Suggestions instead of problems Instead of just looking for problems, the user/evaluator often notices things that could be improved. The aim is to elicit creative thinking on the part of the user/evaluator.(Partly revised due to length) Experience In some way, the level of experience should be considered in the evaluation of entertainment. Regardless of the scale used, the user should be able to value his/her subjective experience. (Partly revised due to length)
5	No new heuristics proposed
6	Create and fulfill expectations On entertainment web sites it is important to create an expectation, for instance by providing a 'fancy' or attractive design and suitable music. However, it is also important that the expectations created are fulfilled. If this does not happen the whole impression of the web site is spoiled, and the user might feel disappointed or cheated. (Partly revised due to length)

	Balance between information and entertainment If entertainment is the goal of the web site, it should strive to achieve a balance between entertainment and information. We think that [entertainment] web sites are often so-called, information containers for banners and entertainment material, e.g. multimedia content etc. However, there is a limited total amount of content for a web site, and in order to achieve a good result. There must be balance. It is also important that the information is relevant and 'fresh'. Otherwise the feature has no reason to be on the web site. (Partly revised due to length)
7	Relevant content and lifetime The content must be relevant to the purpose of the web site. For instance, in order to entertain for more than just 'the moment', the entertainment parts of the web site should have a longer life. (Directly reproduced) Originality and uniqueness The web site must provide something new and unique. Otherwise there is no reason for a user to revisit this particular web site. (Directly reproduced) Design for the target group The web page must be designed, i.e. configured and filled with content, relevant to the intended target group. (This also means that it is important to find out what target group really is the receiver.) (Directly reproduced)
8	No new heuristics proposed
9	No new heuristics proposed
10	Encouragement to explore The users should be motivated to stay on the web site and explore the possibilities it can provide. A boring and difficult design might frighten users away faster and reduce their motivation to stay. (Directly reproduced) Avoid distracting information The primary and essential information must not fade into background or be disturbed by secondary information. (Directly reproduced) Simplicity and completeness The web page should give a [good] overall impression. Secondary features should be removed from the design. (Directly reproduced) Information should be reliable and not misleading All promised information should exist, if this is stated. Otherwise, it should be stated that the information is temporarily missing. (Directly reproduced)

Table 9.2 Heuristics suggested by novices.

Development of a revised list of heuristics

The lists of heuristics, as presented above, were contrasted and an interpretation and combination of these lists was produced. In this process, each heuristic was judged, in order to find other similar suggestions for heuristics from another source in the lists. Similar suggestions were combined into one heuristic, which in some cases was slightly reformulated in order to cover all the suggested heuristics of from which it originated. The total number of heuristics proposed by the experts in phase 2 is fourteen (14). Some of them were judged to be of less importance and these are further described below. First, an overview of the list of abstracted heuristics and corresponding sources for the two groups of experts is shown:

Table 9.3 Overview of the analysis of proposed heuristics.

Number new heuristic	Short description[1]	Originator(s) experts	Originator(s) novices
1	Expectation vs. visual	Expert 1	Group 6
2	Explorative design	Expert 7	Group 10
3	Playability	Expert 7	
4	Durability	Expert 7 Expert 8	Group 7
5	Correspondence design – mediated feeling		Group 1 Group 4 Group 10
6	Clarity of genre		Group 4 Group 7
7	Balance of information and entertainment		Group 6 Group 10
8	Originality		Group 6

Some of the suggestions from experts and expert groups are not included in the revised list of heuristics. There are eight (8) such suggestions from the experienced experts. The reasons they were not included are lack of clarity (1), discussion of technical constraints (1), regarded as being more related to traditional heuristics (understanding etc.) (5) and finally considered too specific to this web site and therefore not generalizable (1). From the novice expert groups the number of unconsidered suggestions is two (2). The first discusses rather unclearly the experience as a whole, which all the heuristics take up. The second discusses the fact that we should not only look for problems but also try instead to find creative solutions when evaluating. This suggestion is included in the overall methodology instead of as an heuristic.

Another comment, in relation to the interpretation and analysis process is that all the suggestions for new heuristics were translated into English before the combining process began. This might be a source of mistakes on the part of the author, which must be considered. However, as the complete study was conducted in Swedish, this could apply to all the quotations and translations made concerning interpretation of any data from elsewhere.

The new list of heuristics is given below. This list is used in the following phase of the study, containing *Inspection Methods*. This phase of the study is further described in Part 4 of the thesis.

New and revised methods

Heuristic number	Name of heuristic Description
1	Visual impression vs. expectations On entertainment web sites it is important to create an expectation, for instance by providing a 'fancy' and attractive design and suitable music. However, it is also important that the expectations created are fulfilled. If this does not happen, the whole impression of the web site is spoiled, and the user might feel disappointed or cheated. The visual impression must not create expectations that the interaction cannot meet.
2	Exploratory design The design should entice the user to explore the web site. The users should be motivated to stay on the web site and explore the possibilities it can provide.
3	Playability – gameplay It is important to visualize the gameplay clearly. Otherwise there is a risk that the user will feel cheated in that he/she had expected to be able to do other things than were actually possible.
4	Durability and lifetime – amount of content Will the content sustain a longer 'visit' or 'stay'. This type of web site is often expected to support or entertain a user, often for a longer time. It should not feel as if it has been 'emptied' after just a short period of use.
5	Coherence between chosen design and desired mediated feeling Does the choice of colors reflect the mood mediated, as intended by the designer? Is the design appealing? Are the colors attractive? How should the menus be designed and where should they be placed? Does the web site catch the user's attention and curiosity and make a generally strong impression.

6	Clarity of genre – design for right target group The web page must be designed, i.e. given form and filled with content, relevant for the intended target group. This also means that it is important to find out what target group really is the receiver.
7	Balance between information and entertainment If entertainment is the goal of a web site, there should be an effort made to achieve a balance between entertainment and information. Entertainment web sites are often so-called information containers for banners and entertainment material, for instance multimedia content etc. However, there is a limited total amount of content on a web site, and there must be a balance if the result is to be good. The content must relate to the purpose of the web site.
8	Originality and freshness (uniqueness) It is important that the information is relevant and 'fresh'. Otherwise it has no reason to be on the web site.

Table 9.4 The list of new heuristics from the first phase of the study.

General methodological revisions

The implications of the general methodology were further discussed earlier in Chapter 6 and 7. Below, the implications from these two chapters are combined and summarized:

Opportunity to give both negative and positive feedback as well as the opportunity to give more motivation for comments about the interface according to the different heuristics.

'Free surf' part of evaluation. In the context of 'fun' it is important, even if evaluation is the purpose, that one part of the evaluation can be regarded as 'free', with few or no obligations to meet. The reason this is so important in this context is its exploratory nature.

Overall review rather than pure evaluation: Evaluation of entertainment may be seen as 'reviewing' rather than 'evaluating' for two reasons: (1) the notion of 'evaluation' has *in itself* a negative tone; (2) when a part is detached from a greater whole, reduced to the levels of specific 'aspects' of some kind something will be lost. Therefore, the need to express an overall judgment is important, especially in the context of entertainment.

The first two implications are included in the new overall methodology but not the third, concerning reviewing. The reason lies in the changes to the heuristics in this part of the study. Only a certain number of conditions could be changed if the results of the changes were to be fully understood. If too many changes were made, changes in reactions could become too confused for experts to be able to comment on the results of the changes. However, this implication was left to be tested in further parts of the study.

Discussion

This chapter described the process of re-designing and revising methods in that part of the study which focuses on inspection methods. The part of the study that included inspection methods revealed the advantage of having an external source of empirical evidence in the process of revision and re-design, i.e. the experts. In the part where inspection methods were explored, the experts were the main source of all the changes. The evaluators being rather administrative resources in the process. The level of interpretation by evaluators in this part was much lower than in the empirical part. The experts provided most of the solutions. This difference in the study is further discussed in Chapter 14.

This concludes the description of the process of revising and redesigning the methods and methodologies. In the next part – Part 4 – the evaluations based on the revised methodology are presented. The first study reported concerns empirical usability evaluations conducted on two entertainment web sites.

Footnotes

[1] The new heuristics are described together with their complete names later in this chapter.

Part IV

Evaluation of refined usability evaluation methods for entertainment web sites

Chapter 10

Applying revised empirical usability evaluation methods to entertainment web sites

Background

The methodological considerations used in this phase are derived from the earlier phases of the study involving evaluations using empirical usability evaluation methods. From Phase 1 in the study, implications for usability evaluation of entertainment web sites are abstracted, concerning conditions such as testing in pairs vs. individuals, use of structured vs. unstructured user tasks, whether to test children or adults, and finally the question of whether written or oral techniques should be used when asking subjects about entertainment. On the basis of the results from the empirical usability evaluations conducted in Phase 1, even general implications not unconnected with the conditions mentioned above were also presented. For instance, one of the results was that the experience and knowledge of the evaluators or experimenters was crucial when evaluating EWSs, because of the subjective nature of the object of study, i.e. the EWSs themselves. Results show that the level of intervention on the part of the evaluator was crucial in evaluating entertainment, and to ascertain this level in each single case would be difficult for an inexperienced evaluator. Another result, clearly shown in Phase I, was the interrelation between ease of use and fun in the context of EWSs. If subjects have problems regarding ease of use, i.e. in understanding what to do and how to do it, it usually influences the extent to which the site is fun (to use). These implications are taken up individually and described and discussed in depth. The main intention in this phase of the study is to further investigate and develop the implications of empirical evaluation of entertainment. The chosen web sites were all entertainment web sites and parts of ongoing or finished design projects at Paregos Mediadesign.

Method

Below, the subjects included, the material used, the design and the procedure within this part of the study are described in more detail.

Subjects

The subjects in the evaluation were high-school students aged 14-16 years, chosen by an external contact, also a student at a local high school. This person chaired a student committee, which included students from all classes at the school, and the subjects were all members of this committee. The total number of subjects in this part of the study was is (10).

The subjects received a cinema ticket, sponsored by Paregos Mediadesign, in return for their participation.

Materials – web sites

The materials in this part of the study included two EWSs – one small web-based game and one community web site for young people. The two web sites chosen for this part of the study were being redesigned at the time of the study and implications from the evaluations influenced the changes made. The game site was, '*Jernkontoret – Captain Steel*', and the community site was *Stadium - Activity Town*.

Jernkontoret – 'Captain steel'

'Captain Steel' is a small game on the 'Jernkontoret' web site, the main purpose of which is to attract a new target group to the web site. The main target group is young people aged 12-18 years. A screenshot from the game of 'Captain Steel' is shown below:

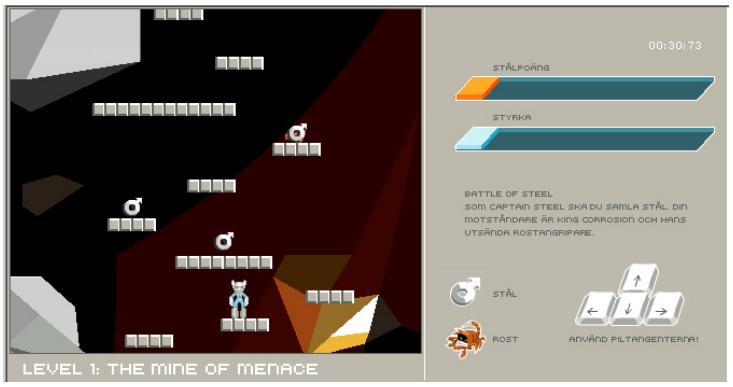

Figure 10.1 A screenshot from the game 'Captain Steel' at the Jernkontoret site (http://www.jernkontoret.se)

Stadium – Activity Town

This web site is a community site for the sportswear store Stadium, in Sweden. The community consists of thousands of registered users and the target group is mainly 15-20-year-olds.

Stadium Activity Town is a concept aimed at encouraging an active life among young people aged 12-18 years. The main idea is to do this by providing an on-line community. Through connecting to big sports events around Sweden, the web site works as an extension of the events on the web, where participants and others interested can meet before and after the happening. This community

includs themes such as local and national sports events, workout, personality and other interests. By involving the members as co-designers, and putting the web site focus on the on members' self-designed web pages and exposing them to other members, the designers hoped to create an inviting and evolving environment for people interested in sports in Sweden. Members are also informed about products in such a way that they are involved in the design of the site, which probably makes them feel special. Activity Town includes such features as; "Sports Academy", "Playground" – a platform for small games, "Library" – where members can find articles etc., "Showroom" – an area where sportswear is displayed, "SportsCamp"- where interesting events are announced, and most importantly for inter-member, communication "Comunity center". A screenshot of 'Stadium – Activity Town' is shown below:

Figure 10.2 A screenshot from 'Activity Town' at the Stadium site.

Design of study

In order to provide further input in evaluating entertainment using empirical evaluation, the study was designed as follows:

Conditions explored

Oral or written answers about entertainment: This specific group of subjects involved in evaluations in this phase, was said to be highly articulate in general, by the person helping the evaluators to recruit the students. The subjects were all members of a committee, at a particular school, which helped to structure questions together with the management at the school. The evaluators, therefore, expected that the subjects would have no difficulty expressing opinions about the web sites.

The pre-task questionnaires required written answers. The students filled in their names and marked on a scale their responses to various questions relating, for example, to their earlier experiences of using the web, making evaluations, etc. The questionnaire comprised only statements and checkboxes.

In the post-task part of the session, the intention was to use two approaches, i.e. written questionnaires and oral interviews, in order to investigate further the relation between written vs. oral answers to questions about entertainment. The subjects were to answer open-ended questions about EWSs in general, the evaluated EWS in particular and the test situation. However, as discussed below, the design of the study was changed during the first session when it became

clear that this group had problems even with questionnaires that included only alternatives to choose among. Thus, only oral interviews were conducted after each user session instead.

Structured vs. unstructured tasks: Earlier results showed that the type of task that was most suitable varied with the type of EWS. This phase includes two completely different types of EWSs, regarding this condition. Based on earlier findings it seemed natural to use unstructured tasks in the case of the game web site, i.e. Jernkontoret – Captain Steel. The earlier results gave no real guidance, however, regarding the community web site, i.e. Stadium Activity Town, which was a highly exploratory environment. On the one hand its exploratory nature suggested the use of unstructured tasks with subjects being encouraged to 'explore the environment'. On the other hand, breakdowns had occurred in some earlier cases, where subjects had shown symptoms of insecurity, probably due to the evaluation situation. As earlier results gave no clear guidance, a flexible approach was employed in evaluations of this web site, where situated decisions would be made concerning the approach to be used in relation to this particular condition.

Pairs vs. individuals: As already mentioned, this group of subjects was considered to be talkative and straightforward as individuals compared to other students of their age – as stated by the source at the school who helped with the signing-up process. This argues for an approach where single-user as well as pair-user tests would be an appropriate study design, in order to reveal any indication of differences between the two approaches. The single user sessions were regarded by the evaluators in advance as being a more authentic use of EWSs. However, earlier results suggest that this type of sessions may suffer from problems in the think-aloud process. The pair sessions had worked better earlier in getting the subjects to think out loud, but there are other drawbacks. First, the sessions are regarded as less authentic, since most web use occurs when the user is alone. Second, there is the problem of what is really being evaluated in the session – the social interaction between the subjects or the interaction with the EWS. Both of these aspects are discussed earlier in Phases 1 and 2 of the study. Despite the problems with the two approaches, it was considered important to investigate further the implications for evaluation of using pairs or individuals.

Testing children vs. adults: Earlier results in the study clearly indicate the importance of testing the intended target audience of the web site. This condition was not, therefore, further explored in this phase. All the tested subjects came within the

intended target group for the web sites, i.e. teenagers. However, as some results in this study produced further input in relation to this condition, it is still mentioned as a condition included in the test.

Procedure

In order to further explore the impact of the above conditions for evaluation of entertainment in EWSs, the procedure of the study was as follows: The subjects worked both in pairs and alone. An overview over the evaluation session is shown below:

	Number
Singles	4
Pairs	3 (6 subjects)
Total tests	7
Total subjects	10

Table 10.1 An overview of the subjects in the tests

Pre-task questionnaire – Here the subjects were asked to answer questions about age and previous experience with computers and the web and finally whether they had participated in similar evaluations or experiments before.

Free-surf approach with think-aloud protocol – The subjects were informed about the process in this part and encouraged to surf through the two web sites as they would have done if it had been an authentic use situation.
The order of the web sites in the 'free surf' sessions was:

1. Captain Steel
2. Activity Town

The order of the web sites in the evaluations could have been counter-balanced to avoid serial order effects, but in this case it seemed unnecessary for two reasons: (1) The subjects were teenagers and thought to be somewhat insecure in the situation, as the evaluations were carried out in an unfamiliar physical setting, and they were considered to be inexperienced in experimental situations. For this reason the evaluators wanted to begin with the game since it would be considered an easy start. (2) Because of the different natures of the two evaluated EWSs, i.e. one game and one community web site, the risk of serial order effect seemed minimal. Thus the presentation of the two web sites was not counter-balanced.

Post-task interview – After the uset session, the subjects were asked to answer questions about the web sites used and also to compare them to other, similar, small web-based games and community web sites they had used or visited. The

subjects were also asked about their expectations about the sites and whether they were fulfilled and what they thought of the so-called gameplay in the games, i.e. whether they thought any game play existed in the games and if so to describe it. Finally, the subjects were interviewed about the evaluation session – where they had worked in pairs, they were asked if they believed this to be an authentic use situation and if any problems, related to the fact that they had worked in pairs, had occurred.

Results

Oral or written answers about entertainment The evaluators thought that the first part of the evaluation session, i.e. the pre-test questionnaire would be easy for this specific group of subjects to complete. However, the students proved to be a little reluctant and frustrated, and some had real problems completing the questionnaire. The input from this was that written answers to open-ended questions about entertainment aspects were out of the question for the rest of this part of the study. Thus, the initial design of the study was changed and interviews were the sole method used in the post-task questions. Even when asked orally about these aspects, the answers were given quite slowly and with hesitation. It was obvious from this study that oral answers have advantages when certain types of users, for instance younger people, are involved in evaluations. It is particularly important to be aware of this in situations where longer and more thoughtful answers are required.

Testing children vs. adults: Both of the web sites clearly exemplified the need to find the proper target group. The evaluators (and the designers) had miscalculated the proper target group for the game – it was considered too easy by the subjects. So here two solutions were possible: (1) If it was important that the target group should be 13-15-year-olds the game had to be re-designed, or (2) the target group for the game had to be changed.

The need to test the proper target group also became clear to the research team with reference to the community web site. Evaluators were surprised at how the subjects explored the web site, and quickly realized that they were quite alone on-line. The social dimension, i.e. communication with others, was the obvious 'killer-app' in this community web site. Subjects also explored the community site and then commented on problems and lack of content in relation to community web sites usually visited. This group turned out to be frequent visitors to virtual community web sites and this web site did not satisfy their needs. Even if this specific fact is of no interest methodologically, it is a strong indicator of the need

to evaluate any entertainment web site using the targeted users. It had not even occurred to evaluators before the sessions that the absence of others on the web site would be as critical as it turned out to be, especially since the web site included many other sections apart from the one offering socializing with others. However, the evaluators were not a part of the target audience. It became clear in interviews that the crucial added value in a community web site – regardless of how well designed it is graphically or how loaded with other fun material it was – is communication with others. It is difficult to say how other subjects outside the target audience, adults for instance, would have reacted to the fact that no one else was present in the chat part of the Stadium Activity Town web site, since no such subjects were used in this phase of the study. However, it is reasonable to believe that adults, not necessarily used to visiting chat rooms, would investigate all parts, or at least other parts as well, of the EWS and not only go to the communication center. So, to conclude, this phase once again indicates the importance of evaluating EWSs using the intended target group.

Structured vs. unstructured tasks: In the game 'Captain Steel' it was natural to give only one task – to complete the game. Thus in this example there was no need to consider whether to use structured or unstructured tasks – the game required only one obvious task, especially because of its limited size.

In the example of the community web site, i.e. Stadium Activity Town, the web site is of a very exploratory nature which could be an argument for using a completely unstructured approach. As mentioned above, however, some of the students were a little shy and reluctant, and in these sessions evaluators intervened and provided tasks to keep the subjects going. So, even if using an unstructured approach is the initial intention in a study like this one, it might also be useful to have some minor structured tasks available to encourage the subject to keep exploring the web site, if breakdowns of the kind described above occur. It should be noted, however, that intervening with too many structured tasks to get subjects to continue the interaction with the EWS, implies a trade-off between completing a test session and observing an authentic use situation by the subject of the EWS evaluated. These sessions, however, can hardly be seen as authentic for these users, as they would have given up if no external assignments had existed. This possibility must be considered in each single case and if there is any suspicion that this is the case, it might be useful to have a follow-up interview and confront the subject with a question about this, i.e. Would the subject consider the previous session as an authentic use situation, in other words would the subject normally use the EWS in the way he or she just did?

Pairs vs. individuals: As in the case of some of the other conditions where subjects had shown symptoms of insecurity and frustration, this was also obvious in relation to the condition pairs vs. individuals. In the sessions where single subjects were tested, it was difficult to achieve a situation in which the subject would think out loud. Single subjects explored the community web site and played the game making no or only minor comments. In the pair sessions, the subjects continuously discussed with each other what was happening on screen, in an apparently natural way. Overall, this phase of the study shows clearly that evaluation of EWSs is facilitated by testing in pairs. Working in pairs on evaluations does not only facilitate the think-aloud process, it also became clear that this type of user often surfs the web collaboratively. In this sense the evaluation also became an authentic use.

Discussion

This phase of the study has further implications about how usability evaluation in relation to entertainment web sites should be considered. In the case of empirical usability evaluation in general, and of entertainment web sites specifically, it is obvious that the procedure is complicated, and requires extensive knowledge and experience on the part of the evaluators. Many things can go wrong and it only becomes obvious once everything is in place and using the system is being evaluated, what things should have been considered before the session. With regard to the interviews it is also obvious that there are so many things that are not covered. This is extremely frustrating for an evaluator – and it is true for all types of empirical usability evaluation.

Another strong implication is the need to be situated as an evaluator. This may well be important in all empirical user testing, but it is even more important when evaluating entertainment. No two sessions are the same, and it is easy to intervene in a negative way. Interventions in evaluations of fun differ from those in evaluation of function because in the former an evaluator can be *too* passive. If an evaluator does not 'play along', or at least seem be somewhat amused by the situation or interaction with the system evaluated, breakdowns will occur. It is an awkward and non-authentic situation for people to enjoy themselves with someone else watching silently and taking notes. This was one of the most difficult parts of the empirical evaluations of entertainment web sites.

In the next chapter, that part of the study where revised inspection methods are used is presented.

Chapter 11

Applying revised inspection methods to entertainment web sites

Background

The results from the earlier phase of the study using inspection methods were revised and re-designed, as described in Part 3. This new methodology was then used by the same group of experts as in earlier inspection method evaluations, in order to explore further the applicability of the method to evaluation of entertainment web sites. This was done in two steps, or iterations, in Phase 3 of the study. In the first iteration, described in this chapter, inspection method evaluations were carried out on EWSs. After the evaluations, interviews were conducted with the experts involved. On the basis of findings from these interviews, the methods were further revised and once more redesigned and used in evaluations of EWSs, as described in the following chapter – Chapter 12.

Method

The experts included, the evaluated entertainment web sites, the design and the procedure of this part of the study are described in more detail below.

Experts

The experts included in this part of the study are exactly the same group as in the first phase, so-called 'experienced experts'. The experts were chosen on the grounds that they could be considered not only usability evaluation experts but also to some extent experts in evaluating EWSs. After participating in the first phase of this study, these experts started to think about, as well as discuss, the

evaluation of entertainment on the web. As the general aim of this study was to build on the methodological framework for evaluating EWSs, created on the basis of findings from earlier evaluations using traditional inspection methods, this group of experts was regarded as having an advantage over other HCI experts.

Materials – web sites

The materials selected to be included in this phase of the study were (1) the 'Skyscraper' web site, and (2) the 'Mosquito' web site. The web sites had been used earlier in the study and will not be further described here[1].

The 'Skyscraper' web site was familiar to all of the experts, since the web site had been included in their first inspection method evaluation. In this third phase of the study it should be seen mainly as a control site, in the same way as control groups are used in empirical usability evaluation or control objects in other types of experiments using the within-subject test design. The rationale for using this design with a control web site, is that the suggestions had been made on the basis of earlier evaluations of this web site. Whether the suggested heuristics could be considered applicable in evaluations of this web site would show in the validity of the results – the applicability of the proposed heuristics would be high, if they were also approved by other experts. We wanted to use an additional entertainment web site, to see if the suggested heuristics could be regarded as generalizable to other web sites apart from 'Skyscraper'. We were aware of the problem that the heuristics not could be regarded as completely generalizable, from evaluating just two entertainment web sites, but as the assignment already entailed a lot of work for the experts, even with only two web sites, the number of evaluated web sites had to be limited. Furthermore, other steps were added during the evaluation process in the study, which further prolonged the total number of hours needed for the experts to finish the assignment. Thus although only two web sites were included in this part of the study, the experts spent many hours on them, for which we are extremely grateful.

Design of this part of the study

The experts were asked to conduct evaluations of the web sites described above, using three types of approaches – 'free surf', Heuristic Evaluation and a meta-evaluation. The design in this phase of the study was based on results obtained by experienced and novice experts in Phase 1 using traditional inspection methods. The experts were aware that it was the process of evaluation rather than then the results of evaluations that were the focus. Similar documentation as in earlier parts of the study was given to the experts for them to follow and complete. No additional information was given apart from that contained in the documentation.

The order in which the web sites were evaluated was switched in 50% of the cases. This phase of the study also ended with a one-hour interview with each expert. An overview of the evaluation of the EWSs is presented below:

- *Introduction* – background information about the complete evaluation and brief overview of the web sites to be evaluated
- *Evaluation of Web site 1*
    - Description of web site 1 and specific instructions for the evaluation
    - Part 1: Exploration and entertainment – 'free surf'
    - Part 2: Evaluation using Heuristic Evaluation
    - Part 3: Meta-evaluation (evaluation of the evaluation in itself)
- *Evaluation of web site 2*
    - Description of web site 2 and specific instructions for the evaluation
    - Part 1: Exploration and entertainment
    - Part 2: Evaluation using Heuristic Evaluation
    - Part 3: Meta-evaluation (evaluation of the evaluation itself)

Procedure

In the 'free surf' approach the experts were encouraged to freely explore the environment in the web site for a limited time of 20-30 minutes. They were then asked to answer three questions:

- How well would you say that you fit to the web site target group (on a scale from 1-5)?
- How much of the web site did you explore in your 'free surf' session (0-100%)?
- For how long did you explore the web site?

In the Heuristic Evaluation the experts were informed about the new heuristics and given some more specific details in the documentation. For instance, it was highlighted that comments could be both positive and negative. The estimated time to be taken for this part was 30-40 minutes.

In the 'meta-evaluation', the experts were encouraged to conduct an evaluation, a grading, of the suitability of each heuristic in relation to the evaluated web site. The heuristics were to be graded from 1 to 5 for applicability. There was also space left for arguments or comments about the grading of the heuristic.

All these steps were repeated for both web sites.

Interviews were conducted after the evaluation sessions to discuss the

advantages and disadvantages of the approaches used. New ideas were also discussed, for instance other possible approaches to be included, heuristics, etc. The completed documentation from each expert was used as a basis for this interview.

Results

In this phase, the results from the evaluations provided information not only about the EWSs, but significantly about the experts' judgments of the suitability of this new approach for the evaluation of these two entertainment web sites.

The responses to the questions asked after the 'free surf' session, which included the experts' fit to the web-site target group, time spent and level of viewed content on the web site[2] indicated how well-informed the judgments of the specific EWSs were. The results from the interviews had an impact both regarding changes in the heuristics and indications for general evaluation methodology.

Methodology for expert reviewing

The interviews revealed various kinds of problems regarding each heuristic, as well as suggestions for new heuristics. In addition, comments and suggestions about more general methodological implications could be abstracted for further discussion. The suggestions below came from one or more of the experts involved. Some have been slightly rephrased so as to be more general and comprehensible outside the context of the discussion. The ideas from experts were also reformulated if they were a combination of two or more suggestions. Below, the methodological implications are described in more detail.

1. Required background information about sites

Implication: Experts need information about the intended target group as well as the specific purpose of site in order to give better feedback on the heuristic 'design for right target group' and 'coherence between chosen design and desirable mediated feeling or mood'.

Background: More than one expert commented on this problem pointing out that it is not reasonable, or likely to produce useful results, if they have to base their comments on guesswork about these aspects. As the experts are asked to make judgments related to the intended target group and the purpose of the web site, they obviously require some information about these aspects.

> *"I believe that the heuristic [of designer's intentions in relation to mediated feeling] is proper and important. However, this assumes that I know the designer's intentions and I do not know that.. I need to be briefed about it, in order to judge this.. it is difficult to discover a breakdown here if I am not informed. "*
> (Expert 1)

> *"I get stuck in the first part of the heuristic [of fit to target group] – what is the target group? There is nothing wrong with the heuristic [per se] but when the target group is not identified I start speculating about what it might be…this is not good."*
> (Expert 4)

Solution: For this reason, the next evaluation provided the experts with background information about the web sites, to give them a proper base for their judgments. There were three changes altogether, all including references to background material delivered to the experts. The three heuristics considered were evaluation based on designer's intentions, the purpose of the web site and its target group. The revised heuristics were heuristics 5, 6 and 7.

2. Need for functional heuristics

Implication: It is still important to consider functional aspects when evaluating an EWS. This implies that some function-related heuristics be added to the list of heuristics. A possible source for this is the heuristics originating from Nielsen.

Background: In many of the interviews, experts indicated that they missed some, or many of the Nielsen's function-related heuristics. Even if the new heuristics covered aspects relating to 'fun', the experts felt a need to comment on more functionally related aspects of the design. A more complete list of heuristics was required. Many experts expressed the opinion that as 'fun' was closely interrelated with, and dependent on functional aspects, functional heuristics could not be excluded from the heuristic evaluation. The experts were asked what specific heuristics they generally felt a need for, but no specific traditional heuristics could be regarded as being of more importance than others. However, there was general agreement that there was no need to add all ten functional-related heuristics. The main essence in many of them could be subsumed into fewer heuristics to avoid the problem of having almost the same number of functional-related heuristics as fun related. There would be two sides to this problem:(1) First, the total number of heuristics must not be too large as this would lead to Heuristic Evaluation being thought cumbersome. (2) Second, if the number of function-related heuristics were similar to the number of fun-related heuristics, methodologically the two aspects should be regarded as equally important in entertainment. However, the overall opinion of the experts was that even if function-related aspects must be considered important in entertainment web sites, they are of less importance than fun aspects, and this should be reflected in the number of each type of heuristic included in the method.

> *"I rather missed Nielsen. I would like to keep some of Nielsen's heuristics … maybe a combination 60/40, with Nielsen's recommendations being the smaller number...or at least an abstract of his heuristics into some function-related rules for use. Function is still important, even in entertainment web sites."*
> (Expert 5)

Solution: For this reason, function-related heuristics, based on Nielsen's heuristics, and input from the suggestions made by the experts, were subsumed into two additional heuristics (Heuristics 9 and 10).

Revised heuristics

The heuristics were changed on the basis of the input summarized above. The new list of heuristics, for further exploration in the last evaluation of this phase, is presented below:

Heuristic number	Name of heuristic Description
1	Visual impression vs. expectations On entertainment web sites it is important to create an expectation, for instance by providing a 'fancy' and attractive design and suitable music. However, it is also important that the expectations created are fulfilled. If they are not, the whole impression of the web site is spoiled, and the user might feel disappointed or cheated. Do not let the visual impression create expectations the interaction cannot meet.
2	Exploratory design The design should entice users to explore the web site. The users should be motivated to stay on the web site and explore the possibilities it has to offer.
3	Playability – gameplay It is important to clearly visualize the gameplay, otherwise there is a risk that the user will feel cheated in that he/she had expected to be able to do other things than are actually possible.
4	Durability and lifetime – amount of content Is there enough content for a longer 'visit' or 'stay'? This is important as this type of web site is often expected to support or entertain a user, frequently for a longer period. It should not feel as if the web site has been 'emptied' after just a short period of use.

5	Coherence between chosen design and desired mediated feeling Does the choice of colors reflect the mood mediated, as intended by the designer? Is the design appealing? Are the colors attractive? How should the menus be designed and where should they be placed? Does the web site catch the user's attention and curiosity and make a generally strong impression. (For intentions about the mediated mood and purpose of the web site – as stated by the designers – please check the description of the web site in the documentation)
6	Clarity of genre – design for the right target group The web page must be designed, i.e. be given form and filled with content, relevant for the intended target group. This also means that it is important to discover what target group really is the receiver. (For information about the intended target group of the web site please see the description of the web site in the documentation)
7	Balance between information and entertainment If entertainment is one goal to be accomplished on a web site, there should be an attempt to achieve a balance between entertainment and information. Entertainment web sites are often so-called information containers for banners and entertainment material, for instance multimedia content etc. However, the total amount of content on a web site is limited, and there must be a balance in order to achieve a good result. The content must be relevant to the purpose of the web site. (For information about the purpose of the web site please see the description of the web site in the documentation)
8	Originality and freshness (uniqueness) It is important that the information is relevant and 'fresh'. Otherwise the feature has no reason to be on the web site.
9	Consistent navigation Does the navigation work in a consistent way, externally and internally? Externally, based on general standards and guidelines for (this type of) web site, and internally, i.e. is the navigation structure of the web site consistent throughout the whole page structure?
10	General function-related aspects Does the web site have an overall functionality which is comprehensible? Does the system give feedback about what is going on? Can a user find out by him/herself if mistakes have been made and remedy them? Are there any so-called emergency exits for users if they have made mistakes? What about help and documentation?

Table 11.2 Revised heuristics from this part of the study.

Apart from comments and suggestions about the heuristics, the experts also had numerous comments and suggestions concerning the overall methodology for evaluating fun. Below, some of them are summarized:

3. Applicability of the 'free-surf' approach

Implication: The 'free surf' approach was originally seen as being more authentic, like a real use situation, than an evaluation session. It became clear that this was important, especially with regard to entertainment. For this reason, the 'free surf' approach remains in the new methodology.

Background: In the interviews as part of the evaluations using the traditional inspection methods, many of the experts discussed the fact that they could only view the sessions where they used the web sites as 'evaluation'. There was nothing authentic about the situation, as they felt like 'judges' and not like 'visitors'. Once the 'free surf' was added to the methodology in this part of the study, many of them commented positively on it. The experts expressed relief that the 'must visit' list, used in evaluations employing the traditional inspection methods, was omitted, and that a time limit of 20-30 minutes was set for this assignment. The session was considered to be more of an authentic use situation, rather than just an evaluation.

> *"Many of these [entertainment web] sites are about being immersed, and if I don't become immersed, I might not experience the web site as it is supposed to be experienced… I believe that the 'free surf' approach works better for entertainment…. [this approach] made me step into the the role[of an authentic user] and then I felt that I may be more fair[in the judgment of the web site] when I become a part of the intended target group"*
> (Expert 4)

Solution: The changes in combination with having three simple questions, as the only steps in the 'free surf' part of evaluation seem to have made this a generally positive experience for many experts.

4. The role of Meta-evaluations

Implication: Some of the heuristics fit more or less well, depending on the type of site. One solution to this might be the ranking from the meta-evaluation of the test, so this is added. The importance/applicability of the specific heuristic to the evaluated web sites is ranked from 1-5.

Background: The main reason for a demand for a meta-evaluation in the context of fun and entertainment lies in the nature of fun and entertainment which is impossible to standardize in a general set of heuristics, unlike function-related contexts such as information retrieval. Thus it is impossible to give a global set of required fun aspects that must be present and completed for a 'fun' system to

be achieved. To show the relevance of each aspect, which is defined in Heuristic Evaluation by the heuristics used, a grading for relevance of each aspect is useful, not only when developing or judging applicability of the method in each case, which was the main intention in this iteration, but also as a part of the overall strategy for evaluating EWSs.

> *"I believe this is an interesting way to think about the heuristics, because how I judged, or ranked, the heuristics influenced my evaluation [of the web site based on these heuristics].*
> (Expert 2)

Solution: In order to overcome the problem of the non-standardizable nature of fun and entertainment in EWSs, the heuristic evaluation method was combined with the meta-evaluation approach – even when the only intention was to evaluate EWSs, and not to judge and further develop the method *per se*.

5. Overall judgment of the web site

Implication: Chance to give an overall judgment or review of the web site in one's own words.

Background: As mentioned in evaluations using traditional inspection methods in earlier parts of the study, there were some demands to be allowed to make an overall judgment of the entertainment web sites. This was not met in this part of the study, as those conducting the study wanted to determine whether it was a matter of the choice of heuristics. However, it emerged that more experts had come to realize the desirability of making such a judgment, and in the interviews conducted in this phase, they indicated that changing the heuristics did not completely solve the this problem.

> *"Sometimes I considered the heuristics and started to wonder if they really covered everything – if they allowed me to express all my opinions about the [web] site. Maybe a freer part would be appropriate, to add comments, if anything is missing [in the heuristics for this specific web site].*
> (Expert 8)

Solution: In the final evaluations in the study of EWSs using inspection methods, a specific section was added in the documentation where the experts could give an overall judgment of the entertainment web site.

Discussion

The main finding from this part of the study is the obvious need to give an overall opinion, a review, of the web site as a whole. It became increasingly clear that what we were dealing with here was not something that could simply be broken down into parts, as is traditionally done in usability evaluation. The experts started to think about the evaluation of entertainment on a deeper level in this part of the study, and interviews indicated that they were doing this not only in the context of evaluations of EWSs. Discussions about evaluating movies, games and other things, came up in the interviews. The experts expressed an interest to continuing their help in another iteration of evaluations.

Additional results from this phase are the suggested changes in existing fun-related heuristics, as well as adding functional heuristics. In addition, the role of the meta-evaluation was somewhat changed – from being a tool for evaluating the method as the object of study, to becoming a tool included in the composition of methods for evaluating the EWS as the object of study. Finally, the 'free-surf' approach replaced the traditional design walkthrough, as the 'free-surf' session gave the evaluators the opportunity to be 'users' in a more authentic way than in the design walkthrough approach.

In the final part of the study including inspection methods, these findings were implemented as part of the general methodology, and the changes were tested and judged by the experts.

The next chapter presents that part of the study which included these changes in methodology. The design of the study is described in detail and the findings are summarized. These findings are also the final findings of the refining inspection methods for evaluating the fun 'track', for the complete study.

Footnotes

[1] For descriptions, see Chapter 5 and 6.

[2] The results are summarized in Appendix II.

Chapter 12

Applying revised inspection methods to entertainment web sites – second iteration

Background

The input for the new methodologies used in the study design in this phase was derived from the earlier phases. Conclusions concerning evaluation of entertainment web sites could be drawn from these earlier phases. The implications of these are discussed and described in depth. The entertainment web sites selected for this part of the study had not been tested earlier in the study.

Methodology for expert reviewing

The main findings from the earlier phases of the study which constituted methodological input to the study design in this part were:

1. **Required background information about sites** - Information was given to the experts about intended target group as well as the purpose of site in order to facilitate better feedback on heuristic 'design for right target group' and 'coherence between chosen design and desired mediated feeling or mood'.
2. **Need for functional heuristics –** There were some changes in heuristics to be used in this phase: the language used was changed to make them more easily understandable and function-related heuristics were also added, loosely based on Nielsen's heuristics.

3. **Applicability of the 'free-surf' approach** - The design walkthrough approach was replaced by a 'free surf' approach, which was seen as being more authentic, i.e. more like a real use situation than an evaluation session. It became clear that this was important, especially in the case of entertainment, since evaluation and ordinary use of this type of web site differ widely. For this reason, the 'free surf' approach is retained in this new methodology.
4. **The role of Meta-evaluations** - Some of the heuristics fit more or less well, depending on the type of site. Here the ranking from the meta-evaluation of the test might serve as one solution thus it was also added as a tool in method for evaluating EWSs, and not just as a tool for evaluating methods. The importance and applicability of the specific heuristic when evaluating each web site is ranked from 1-5.
5. **Overall judgment of the web site –** Due to the need for a method to highlight more holistic perspectives of EWSs in evaluations, a reviewing stage was added in this last phase of the study. This was done to provide experts with the opportunity to make an 'overall judgment or review' of the web site in their own words. Commenting on details, which are understood as problematic does not necessarily correlate with an negative overall impression of the whole EWS.

Method

Below, the experts conducting the evaluations, the material used and the design and procedure in this part of the study are described in more detail.

Experts

By this stage, the experts included in the study had developed an experience in evaluating EWSs and in judging methods for doing this. Thus, this group was the first choice of experts to be included in this last phase. It was important here not only to use experts in testing interfaces in general, but experts in testing entertainment web sites in particular. If we did not use people who were experts there was a great risk that the results would correspond to those from the start of the study where inspection methods were used. Another aspect, which also strongly influenced the choice of included experts in this last phase of the study, was the advantage of being able to compare methodological approaches among the phases in the study. In this study design, the experts were able to draw on their experience from other evaluations of entertainment in order to provide even more informed feedback.

The same ten experts as had been used earlier were asked to participate in this part of the study, to conduct evaluations using inspection methods. Nine of them agreed, but one, unfortunately, had left the department, thus another person, also a member of the Department of Informatics, was asked to join the team. The background profile of this new expert was similar to the one who had left, including experience of evaluation in research and lecturing. However, the new expert differed from the others, in that s/he had no previous experience in evaluating entertainment web sites. This was taken into account in the analysis of the data.

Materials – web sites

In this part of the study the web sites evaluated were 'Vodafone – How are you?' and 'Stadium Activity Town – 'Bad Guy Monkeys'. Vodafone was up and running, but was a subject for future re-design and input from the evaluations regarding usability was to inform this re-design process. The game 'Bad Guy Monkeys' on the Stadium web site, was still under construction. Input from the evaluations would have a direct impact on the design of the game.

Vodafone - 'How are you?'

This part of the Vodafone web site was a campaign site, designed by Paregos Mediadesign AB. The target group of users of *'How are you?'* was 18-30-years-olds. The instructions to the designers of the web sites were:

> *"Use 10 different moods in eight languages and one great community to connect with people from around the world. For example "I'm in love" "I'm Gorgeous" "I'm Stuck" Purchasing different local mobile service providers Vodafone wanted to unify them by changing its name to Vodafone throughout Europe. Our mission has been to communicate this on the web supporting other media channels. Making people experience that they are part of a worldwide community, unifying the Vodafone brand."*
> The web site of Paregos: http://www.paregos.com/)

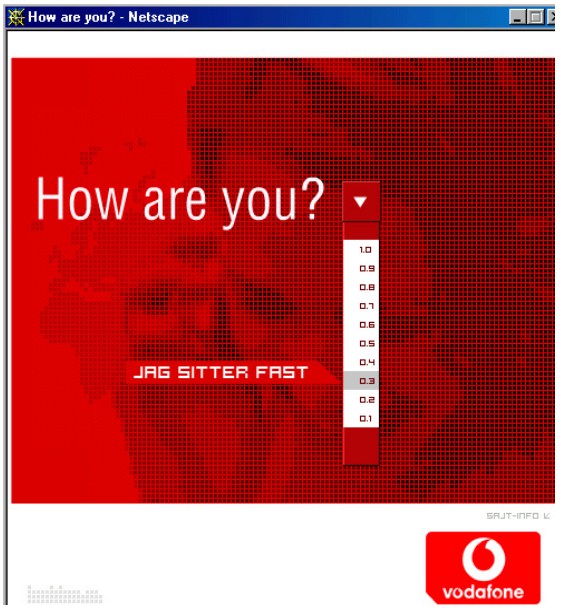

Figure 12.1 'How are you?' campaign site from Vodafone. (http://howareyou.vodafone.com/)

Figure 12.2 The game 'Bad guy monkeys' in 'Activity Town' at the Stadium site

Stadium – Activity Town – 'Bad Guy Monkeys'

The 'Bad Guy Monkeys' was included in the earlier evaluated community web site 'Activity Town' presented by Stadium – a sportswear store in Sweden. As the game had not been completed when the empirical evaluations of the web site were conducted earlier in the study, it had not been included. In this phase the focus was only the game and not on the whole 'Activity Town' site. 'Bad guy monkeys' is an adventure game influenced by soccer, intended to involve its players and arouse curiosity. It is not to be seen as a competitor to traditional soccer games. The intended target group was those familiar with computer games, aged 12-18 years.

Study design

In this part of the study experts were asked to conduct evaluations of the entertainment web sites described above, by using four types of approaches – 'free surf', heuristic evaluation, meta-evaluation, and finally an overall review of the complete web site. The design of this part of the study was based on the results from earlier phases in the study. In this phase, the experts were also aware of the fact that it was the process of evaluation itself rather than the results of the evaluations that were in focus. Similar documentation as in earlier phases was provided, for the experts to follow but they were given no additional information apart from that in the documentation. The order in which the web sites were evaluated was reversed in 50% of the cases to balance the sequence effect. This part of the study ended with an e-mail-based questionnaire including open-ended questions, addressed to the experts about the methodology used. A text-based questionnaire was used out of respect for the experts, as they had already spent numerous hours in evaluations and interviews in earlier phases of the study. A questionnaire would require less time to complete than a one-hour interview session, and was less dependent on pre-scheduled meeting arrangements for interviews, i.e. experts could complete the questionnaire at any convenient time. An overview of the design of the study is presented below:

- *Introduction* – background information about the whole process and a brief overview of the web sites to be evaluated
- *Evaluation of Web site 1*
    - Description of web site 1 and specific instructions for the evaluation
    - Part 1: Exploration and entertainment – 'free surf'
    - Part 2: Evaluation using Heuristic Evaluation
    - Part 3: Meta-evaluation (evaluation of the evaluation in itself)
    - Part 4: Overall review of the complete site.
- *Evaluation of web site 2*
    - Description of web site 2 and specific instructions for the evaluation
    - Part 1: Exploration and entertainment
    - Part 2: Evaluation using Heuristic Evaluation
    - Part 3: Meta-evaluation (evaluation of the evaluation itself)
    - Part 4: Overall review of the complete site.

Procedure

As mentioned, the first three approaches included in this evaluation were almost identical to the procedure in the earlier inspection method evaluations in the study. There were only minor differences in the heuristics used– they were supplemented with function-related heuristics, and some minor changes had been made in the language. Finally, the experts were supplied with a more detailed description of the purpose of the web sites and intended target group. For reasons of clarity these three approaches are also described here:

In the 'free surf' the experts were encouraged to freely explore the environment in the web site. The exploration was limited in time to 20-30 minutes after which they were asked to answer three questions:

- Considering the target group of the web site – how well would you say that you fit into this group (on a scale from 1-5)?
- Considering the web site as a whole – how much of it did you explore in your 'free surf' session (0-100%)?
- For how long did you explore the web site?

Heuristic evaluation

In the heuristic evaluation the experts were informed about the new heuristics and given some more specific details about the documentation. The estimated time taken for this part was 30-40 minutes.

In earlier phases, the experts had commented that they needed other heuristics and evaluation approaches to be able also to consider the positive things about the EWSs, as this was difficult using only the traditional heuristic evaluation. In this phase it was therefore interesting to see how the fun-related heuristics were used in relation to the functional aspects in heuristic evaluation. All the fun-related heuristics (1-8) were included in one group and the functional heuristics (9-10) in another. The total numbers of positive and negative comments in relation to each heuristic were counted and the two groups – fun vs. function – were compared.

First, there is some clarification of how the number of comments made by the experts about the EWSs is counted, and an explanation is given as to how they were positive or negative. Here, two choices made in the interpretation are worth mentioning. (1) If the expert made a comment and stated that it related to two heuristics, the comment was counted as two comments. (2) In those cases where the type of comment, i.e. whether it was to be considered positive or negative, was not stated by the expert, and the comment could be seen to be both positive and negative, the comment was counted as two – one positive and one negative. Finally, a comment may be important to make, regarding the comparison between the groups of comments, the fun-related including eight heuristics versus function-related including only two. It might seem strange to compare these two groups some might argue. However, as the function-related heuristics cover more or less all of the Nielsen heuristics it seems reasonable to understand the two groups as equal enough to make the comparison.

Meta-evaluation

In the meta-evaluation, the experts were encouraged to make an evaluation, a grading, of the suitability of each heuristic in relation to the particular web site. The heuristics were graded from 1 to 5 and space was given for comments about both the grading and the heuristic. These results were to be used somewhat differently in this phase of the study.

Overall review

The new evaluation approach asked the experts to express more freely their overall impression of the web site evaluated. The space allowed in the documentation for writing this review was consciously extensive, to encourage the experts to write in detail.

The overall judgment of the web site was compared to the results from the heuristic evaluation. The main interest was to see to what extent the relation between positive and negative comments about the web site correlated with the overall review of the web site. The experts expressed their judgments in the overall review, and the author interpreted these overall judgments and placed them in a scale from 1 to 5, where 1 was considered completely bad and 5 completely good.

All these steps in methodology were repeated for both web sites. As in all the other parts of the study, the order in which the web sites were evaluated was reversed in 50% of the cases to balance the sequence effect.

Results

The results from this part of the study are reported in separate parts. First, an overview of the comments on the web sites in the heuristic evaluation is discussed. Here, the emphasis lies on the relation between positive and negative comments. An overview of the meta-evaluation is also provided and discussed – it terms of whether any general patterns are present. Finally, the methodological changes are discussed, on the basis of the questionnaire addressed to the experts – asking what they thought of the changes in the methodology in this part of the study. The experts also provided further comments and input, unconnected with the listed changes, which are also discussed.

Results from the heuristic evaluation of the entertainment web sites

We wanted to examine the extent to which the comments were given in the list of Heuristics 1-8, i.e. the entertainment-related heuristics, and in Heuristics 9-10, i.e. the function-related heuristics[1]. An investigation was made into the number of positive and negative comments in the two groups of fun- and function-related heuristics.

There seems to be a pattern, as shown in both studies, in that the majority of positive comments were made in the entertainment-related heuristics. In those cases where a high percentage of positive comments were made in the functional heuristic list, the total number of comments – including both fun and functional heuristics - was low, i.e. 1-2 in total.

The negative comments were scrutinized in the same way. No such pattern as appeared in the positive comments could be found in the negative comments.

One may conclude on the basis of this, therefore, that positive comments are more frequently given in relation to fun than in relation to function in evaluations. Negative comments are given in both cases, i.e. regarding both fun and function.

Results from the overall review of the entertainment web sites

One interesting aspect to investigate further was whether there was any correlation between comments given in the heuristic evaluation and those in the overall review. Some experts had asked for the inclusion of an overall review, since this could still be positive even if there were many comments regarding specific problems. This had been pointed out in earlier investigations of web sites on the basis of other methods. In order to make this comparison the overall review was interpreted by the author[2] and positioned on a scale from 1-5. In the discussion below, the possibility of the occurrence of errors due to interpretation mistakes made by the author must be taken into consideration.

In the case of the game web site – 'Activity Town Bad Guy Monkeys'- the results of the comparison show that in eight cases out of ten, the overall judgment correlated with the balance between positive and negative comments. For instance, Expert 1 gave an overall judgment of 4[3], and the relation between positive and negative comments is 70/30. This would indicate that the overall impression correlates with the number of negative and positive comments made. Similar results were obtained also for Experts 2-8. However, this was not true for Experts 9 and 10 as they had a high number of negative comments in connection with the heuristic evaluation – 78% and 71% respectively of these experts' comments were negative. At the same time they gave an overall judgment which was approximately graded as 3. The interpretation of this would be that even if a large number of negative aspects or problems are found, the overall impression may still be positive. If the overall review approach had been excluded in this study, the results from these two experts would not have correctly indicated their attitude towards the web sites.

The same situation as described above occurred in the study of the 'Vodafone – How are you?' web site. Those cases where the overall review did not at all match the relation of positive to negative comments, are highlighted. Where the differences between review and comments on the heuristics were minor, the results are not highlighted, as the overall judgment is an interpretation by the author, and the difference cannot be considered to be completely significant in these cases. Only the obvious cases are included in this investigation.

Results from meta-evaluation

The purpose of presenting the results of the rating of the heuristics in the meta-evaluation is to give an indication of how applicable the experts considered the heuristics to be in the specific cases of the two web sites, and to investigate whether these results are significant. An overview of the results from the meta-evaluations

is presented in Appendix II. Furthermore, the mean value of the ratings of each heuristic is calculated and included in the tables presented in Appendix II. However, these mean values should only be seen as an overall *indication* on how experts in general graded the applicability of the heuristics, and not as data that could be used for detailed comparison of specific heuristics. The reason for this is that the rating scale in this case is an example of an *Ordinal scale* comprised of non-parametric values. In means that it is not possible to meaningfully compare intervals between the values, i.e. the ratings of the heuristics – the difference between grades '1' and'2' cannot be considered to be exactly the same as the difference between values '3' and '4'. (c.f. Anderson et. al., 2002).

Based on the *Mann-Whitney* and *Kruskal-Wallis* test models[4], non-parametric tests were performed on the expert ratings of the heuristics, and these tests formally show that there exists a significant difference between the ratings of the heuristics. Further, comparison of results from the non-parametric tests and the calculated mean values of ratings of each heuristic show that similar differences occur in both cases. Therefore, it can be concluded that the mean values could be used as an indication of the overall judgment of the applicability of each heuristic in this case (Anderson et. al., 2002; Solso, 1998).

Conclusions in relation to these results are that (1) with only ten experts involved in the evaluation significant values for ratings can be obtained, and further, (2) the differences between heuristics, according to the ratings, in their applicability in both of these two evaluations of EWSs are significant. Therefore, meta-evaluations can provide important evidence and, in our view, are to be included in the design of future studies of this kind.

Results about the overall methodology – from the questionnaire completed by the experts

Extended information about web sites given to the experts

All the experts agreed that this was an important change. No specific quotation is shown, as this was a general opinion. This change in the methodology can even be considered as more of a correction driven by a mistake rather than an actual implication.

Changes in the heuristics – language and functional related heuristics

Overall, the reactions from the experts to these changes were positive. Many had felt a need for this type of support for the reporting of problems found, related to functional aspects in the last part of the study. Below examples are shown of comments from the experts, given in the post-evaluation interviews.

> *"It felt safe that these two new heuristics were included. They helped to identify the problems which are function related."*
> (Expert 3)

> *"[The function-related heuristics are]..an important factor, which should always be included [in evaluations]. Regardless of what is evaluated, it is important to be able to interact [well] in a satisfying way. It was good that they were present."*
> (Expert 9)

> *"Needed indeed. Even though one might think functionality has little to do with entertainment it is still very important."*
> (Expert 8)

> *"The heuristics are definitely better this time, even if I cannot pinpoint the actual differences."*
> (Expert 3)

The 'free surf' approach remains in the methodology

In this part of the study, the 'free surf' approach was used once again, as positive feedback had emerged from earlier interviews. This gave the experts one more chance to use and judge the approach in relation to the evaluation of entertainment web sites.

> *"[The 'free surf' approach is retained in the methodology – what do you think about that?] It sounds reasonable, I agree."*
> (Expert 6)

> *"It gave me time to think, and I was also free to explore the web site on the basis of my own preferences in entertainment."*
> (Expert 5)

> *"I am happy that this approach was retained [in the overall methodology]. I agree fully that it feels more authentic."*
> (Expert 4)

Ranking of suitability of heuristics in meta-evaluation

The experts were asked about including the ranking of heuristics according to suitability. Overall, the experts were positive about this strategy, however, some concerns was raised about the extent to which the site was evaluated on the basis of heuristics or vice versa.

> *"One problem which could be considered is the extent to which the web site is evaluated on the basis of the heuristics, or whether the heuristics are evaluated on the basis of the web site."*
> (Expert 8)

> *"Worked well. Made me reflect, as an evaluator, on the heuristics I had used.."*
> (Expert 4)

> *"Good, but not that essential. I gave most of the heuristics the highest grade."*
> (Expert 2)

The comment about the evaluation of the web site vs. the heuristics is an important one to consider, especially if more inexperienced experts are used in this type of evaluation. However this problem can be dealt with reasonably easily by supplying by supplying detailed descriptions of the specific purpose of this step in evaluations of EWSs.

Chance to give an overall judgment or review

In the two earlier phases of the study, which include inspection method evaluations, experts had expressed a need for some kind of overall judgment in this type of evaluation. It was used for the first time in this phase. Some comments on its existence in the overall methodology are given below:

> *"Very good!..it felt as if my judgment of the web site was more general than just giving specifically positive or negative criticism. This part of the evaluation gave me the freedom to provide more of a summary and a final answer in the evaluation."*
> (Expert 3)

> *"Good as a summary. However, there might be a risk here, and that is that the subjective opinions of some experts will dominate. Opinions may be good, but it is not necessarily the case that these specific opinions about taste, preferences and such are relevant."*
> (Expert 10)

> *"Good. You get a final chance to say what you appreciated on the site and what was entertaining about it."*
> (Expert 9)

> *"Valuable if a large number of negative comments have been given, but the overall impression is still quite good."*
> (Expert 2)

Generally, the experts seem to have reached a consensus about the level of efficiency and applicability of the overall methodology, either because they were too exhausted after all the hours of evaluations and interviews and simply resisted giving more feedback, or the methodology can be seen as fairly complete, given the above conditions for evaluating at least these types of entertainment web sites.

The experts all answered positively to the question of whether they would consider participating in further studies, which may be an indication that the 'exhaustion' explanation is less likely to be true. Whatever the reason was for the generally positive response to the overall methodology for evaluating entertainment, there seemed to be no emergent need to proceed with the iterations by this stage of the study. The results should be regarded as initial work, to be used for further studies.

Discussion

Overall it can be concluded that inspection methods, as they were designed and combined as in the last phase of the study, can be valuable in evaluating fun and entertainment. It is important to raise some concerns when evaluating fun compared to evaluating functional aspects using inspection methods, however. In particular it is important of providing steps in the evaluation, which are as similar as possible to authentic use of the web sites. The reason for this is that in structured evaluation processes, for instance, where there are lists of heuristic steps to follow in evaluations, experts have difficulties judging to what extent the web site are fun or entertaining. The reason for this is that evaluation as activity is understood by experts, something very different from authentic use of entertainment web sites. If the aspects of fun and entertainment are to be found in expert evaluation, more structured approaches have to be combined with more free approaches. Otherwise experts are focused on evaluation rather than exploratory use and amusement.

The findings in this last study, using inspection methods, present the final proposal for a revised and re-designed methodology for evaluating fun. Earlier in Part 4 of the thesis, the end findings and conclusions regarding empirical usability evaluation methods are also shown. These findings are presented in summary in the next chapter together with some conclusions. The general implications for usability evaluation in the context of entertainment are then discussed in the final chapter.

Footnotes

[1] The tables with the complete information of the relation between the positive and negative comments given by experts in the heuristic evaluation in this phase are presented in Appendix II.

[2] The grading of the overall review from 1-5 could easily have been done directly in the evaluations, as a part of the assignment for the experts. However, this part was not included in the methodology at the time of the expert evaluations.

[3] As interpreted by the author.

[4] For a more detailed description of the statistical test models of Mann-Whitney and Kruskal-Wallis, see, for instance, Anderson et. al.(2002)

Part V
Conclusions

Part 5 is the concluding part of the thesis. It presents a summary of findings of the study described in the previous parts, reflections on the findings and the study in general, and an overall discussion. The part is organized into two chapters, Chapter 13 and Chapter 14. This division reflects a dual status of methods in this thesis, namely, methods as usability evaluation tools and methods as objects of study. The chapters comprising Part 5 deal, respectively, with (a) evaluation of fun and entertainment and (b) analysis of methods for evaluation of fun and entertainment.

Chapter 13 focuses on the implications of the study for usability evaluation of fun and entertainment. The aspects, or "dimensions", of the design of evaluation procedures suitable for evaluation of fun and entertainment are discussed in the chapter. Chapter 14 deals with a more general set of issues. It identifies a number of fundamental conceptual distinctions related to analysis of usability evaluation methods and discusses these distinctions on the basis of the findings of the study reported in this thesis. The chapter concludes with reflections on generalizability of the study, other possible approaches to evaluate experience, and prospects for future work.

Chapter 13

Summary of empirical findings

The aim of this chapter is to summarize empirical findings of the study reported in this thesis. The findings are divided into two groups representing the two main foci of the study, that is empirical usability evaluation methods and inspection methods. These empirical findings are then used as a basis for more general discussion of methods to be used for evaluating entertainment and fun in the context of web usability in general. These conclusions are discussed further in Chapter 14.

Empirical usability evaluation of entertainment web sites

The empirical evaluations conducted in this study produced a number of findings about the aspects of evaluation procedures that should be taken into account when evaluating fun and entertainment in case of entertainment web sites. Below, the findings are summarized, first with regard to the specific conditions examined in the study followed by a general methodological discussion.

Conducting an empirical usability evaluation involves making numerous decisions about the concrete design of the evaluation procedure, which would be optimal for the purposes and general context of the evaluation. A number of aspects, or "dimensions" of the design of evaluation procedures were identified in usability evaluation research as important and potentially problematic. Our study was designed to provide empirical evidence about the importance and relative advantages of these dimensions by comparing various controlled conditions employed in, the study. This evidence can be summarized as follows:

- Pairs vs. individuals – Testing entertainment web sites in pair settings works well, particularly when children and teenagers are tested. In some cases a pair session design must be regarded as an unauthentic situation, that is where the web sites are mainly intended for individual users.. However, results from the study show that authentic use of web sites designed for single use often occurs in pairs. This was clearly shown in the part of the study that involved teenagers, who use EWSs in collaboration with others, for instance, at school. In these cases evaluating in pairs is more authentic than single user evaluation. When testing pairs of subjects, such things as domination, 'showing-off' and competition within the pairs must be taken into account. Furthermore, when testing pairs, it is important to always be aware of what is being evaluated, i.e. the interaction between the *subjects* or the interaction between the *subjects* and the *web site*.
- Structured vs. unstructured activities – Traditionally, the use of structured tasks is a common approach in the context of usability evaluation. In this study, the subjects were asked to make evaluations that included both structured and unstructured tasks. As many EWSs are exploratory in nature, providing subjects with unstructured tasks appears to be a reasonable approach. However, depending on the type of entertainment web site evaluated, a structured approach with specified assignments for subjects to complete is not such a bad approach. The main reason for this, according to the results from the study, is that some subjects are frustrated when the assignment is too unstructured or free. Breakdowns occurred in some sessions for this very reason. However, in highly exploratory web environments or where only one task is concerned, for instance in web sites which are comprised only of a game of some kind, the use of unstructured tasks is a more applicable approach.
- Testing children vs. adults – Children as subjects are more spontaneous and more willing to explore. In successful evaluations, where no breakdowns related to the evaluation *per se* occur, it is possible to obtain data of high quality from them. If children are the target group of the evaluated EWS, some aspects might be impossible to test on any other group. However, it might still be worthwhile to also include adult users in these evaluations, since adults are often better at thinking in abstract terms and verbalize more easily.
- Written vs. oral answers to questions regarding entertainment – Oral answers are to be preferred when asking questions regarding entertainment, because of the subjectivity of the answers and the possibility of asking follow-up questions.

Finally, our tests showed the importance of being situated and intuitive as an experimenter, if useful results are to be obtained when testing entertainment. As a subject, to laugh in a silent crowd is difficult.

The free-surf approach was fairly successful in the case of the Mosquito site. In most cases, the single users never asked for the tasks offered as support. When tasks were assigned to the subjects, they only needed one or two to get going. There was only one case where the subject requested all the pre-defined tasks.

The think-aloud method presented some difficulties. Some subjects showed signs of stress, uncertainty and had difficulties verbalizing even in the practice session. These subjects also continued to have problems verbalizing their thoughts throughout the rest of the test session. This greatly affected the level of intervention, as the test crew had to encourage the subjects, several times at short intervals, to start verbalizing. In general, the subjects had difficulties verbalizing continuously throughout the test. None of the single users managed to sustain continuous verbalization throughout the whole session, without encouragement from the test crew.

The subjects, who worked in pairs to do the pre-defined tasks, had few difficulties verbalizing, compared to those who worked alone free-surfing. These subjects also required less intervention, since they discussed their thoughts with each other in a way that eliminated the need for encouragement to verbalize from the evaluators. The outcome of the verbalizations differed between the single-user group and the group working in pairs. The single-users more frequently expressed subjective, emotional thoughts and provided more information on the impressions the different aspects of the site created. The pair verbalizations about the tasks, tended to express various strategies and problem-solving approaches, rather than subjective thoughts and impressions of the site. This may simply reflect the different instructions the two groups were given, or perhaps that a goal-orientated approach to thought is easier to verbalize than a more intuitive process of thinking, based on subjective impressions. A summary of conclusions regarding empirical usability evaluation

1. Test subjects might have difficulties verbalizing their thoughts when interacting with this kind of web site. This entails a relatively high level of intervention on the part of the evaluators that needs to be considered in the design of the evaluation.
2. Subjects working together reduce the level of intervention, but differ from subjects working alone in terms of the content of their verbal reports.
3. A carefully designed study with structured tasks may provide important information, even when the web site is entertainment focused.

4. The free-surf approach cannot to be relied on alone, but is effective when supplemented with additional tasks offered as support.
5. Even if web sites are entertainment focused, traditional usability methods may provide useful tools for practical evaluation, if they are chosen with regard to the functionality of the site to be evaluated.
6. The behavior of the test people interacting with this type of experience focused web sites, as opposed to, for instance, information retrieval, is often unpredictable and presents a number of factors that are difficult to control during the test. The evaluator must, therefore, be very careful and flexible in designing the test, and be prepared to change and refine methods during the evaluation in order to compensate for unforeseen behavior and effects.

Evaluation of entertainment web sites using inspection methods

The main findings from the last part of the study where inspection methods were used, refined and revised, can be summarized as:

Providing experts with general information about web sites

It is important that the information about the intentions and aims of the EWSs evaluated is as extensive as possible for valid judgments about the web site to be evaluated. All included experts agreed on the necessity of extending the information about the web sites' intended target group as well as on the originators' goals for the web site, as interpreted by the designers.

Changes in the heuristics – language and functional related heuristics

There seems to be a relation between functional aspects and fun and entertainment aspects in EWSs. The number of heuristics changed from eight to ten in the last evaluation using inspection methods. The additional heuristics were function related. Overall, the experts were positive to this change, and the results from the evaluations of the web sites also show that the heuristics were widely used, which may indicate the need for this type of heuristics – even when entertainment web sites are evaluated.

The 'free-surf' approach is retained in the methodology

When evaluating the usability of any system, it is always important to set up a use situation, which is as authentic as possible. This is also true for evaluations of EWSs. However, as some of the experts commented, it may be more difficult in

these cases. The 'free-surf' approach was highly valued by the experts in evaluating entertainment web sites. The reason for this was that the evaluation session, as designed in the first place, turned out to be far from an authentic use session of entertainment web sites. The experts could not escape from the fact that they were evaluating the web site and not entertaining themselves. This differs from evaluating pure function, where the difference between evaluation and use is less. For this reason, the 'free-surf' approach remained in the overall methodology for evaluating entertainment.

Ranking of suitability of heuristics in meta-evaluation

The meta-evaluation was introduced into the study mainly to serve as a tool for the study of methods as 'objects of study' and not to supply any information to the process of evaluating entertainment web sites as 'objects of study'. However, the experts implied that this was a valuable tool, even in the latter case. The reason lay in the nature of the entity 'entertainment web site', which must be considered highly individual. In some entertainment web sites, playability is very important and in others playability is not applicable at all. The meta-evaluation was seen as a tool to mediate this applicability in each case.

A possibility to give an overall judgment or review

Fun and entertainment are difficult to judge just by investigating their parts. It needs a more holistic approach, where the greater whole is bigger than the sum of its parts. This part of the overall methodology came from an idea developed by some of the experts early in the study. The differences between the concepts of 'evaluation' and 'reviewing' were highlighted, where evaluation often is seen as 'revealing problems' but reviewing is more about 'giving an overall judgment'. In the context of entertainment this seemed relevant. This approach was tested in the last inspection method evaluation, and two types of results indicated its importance in the methodology. (1) The overall judgment, given in the reviews, did not always correlate with the balance between positive and negative comments given in the Heuristic Evaluation, i.e. the rate of negative comments could be high, but the overall review might still be positive. (2) The second type of result indicating the importance of the overall review was the number of positive responses from the experts. In general, they were very positive about the presence of this approach in the overall methodology.

Summary of conclusions regarding Inspection Methods

- Inspection Methods are to be considered a proper choice when evaluating entertainment web sites.
- The re-designed methodology, as described above, has produced good results regarding the applicability to evaluating entertainment.
- It is crucial to mix structured approaches with more free approaches, when evaluating entertainment using inspection methods. One example of this is that in some cases the rate of negative comments was high, but the overall judgment, as stated in the review part of the evaluation, was still positive or very positive.
- The results in this study must be seen as a basis for future work and not as a general approach applicable to evaluation of all types of entertainment web sites.

Discussion

The large number of evaluations conducted on EWSs in this study produced numerous findings, relevant to both the EWSs *per se,* and how this type of evaluation can be conducted methodologically. Throughout the thesis, the findings about the EWSs *per se* are not singled out and reported, as the main research question covered in the thesis concerns methods rather than the web sites evaluated in this specific study. Nevertheless, these findings about the web sites are important for the results and conclusions in this study. They work as a basis for discussions with the experts in interviews, they highlighted what required heuristics would be necessary for the experts, and in every user session observed in the empirical evaluations in the study, they indicated how methodology could be developed for future evaluations of EWSs. Finally, the findings about the EWSs *per se* produced important implications for how to operationalize fun and entertainment in the context of web usability. One example of this was the framework of *form* and *content*, where the empirical findings from evaluations of the web sites informed the extent to which this framework was useful in this context, i.e. does division of the object of study into form and content provide any guidance in evaluations of this kind to. Furthermore, empirical findings about web sites indicated, throughout the study, the division, and boundary, *between* the two concepts. Even if the reported empirical findings of the EWSs *per se* are not specified in each case, they should be regarded as being the essence of the thesis, from which other, higher-level conclusions are drawn. The next chapter presents the general conclusions of the study.

Chapter 14

Discussion

The findings of this thesis indicate that traditional usability evaluation methods can be applied to evaluation of entertainment and fun in the context of web usability. More specifically, existing empirical evaluation and heuristic evaluation methods were found to produce relevant and potentially useful evidence when applied to evaluation of entertainment web sites. However, it was also found that existing methods have serious limitations and needed to be further developed and revised in order to become more applicable. The proposed changes to the methods, based on findings from the study in this thesis, are described and summarized in Chapter 13. It was demonstrated that after the revision there was an improvement in the quality of the methods when used as analytical tools for usability evaluation. Therefore, the main conclusion of the thesis can be formulated as follows; while the underlying concepts and principles of web usability can be employed in evaluation of entertainment and fun, specific evaluation methods need to be revised. This conclusion will be elaborated upon and made more specific in the discussion below which focuses on the concrete implications of the study for the evaluation and design of usability evaluation methods targeting fun and entertainment.

As already mentioned, considering evaluation methods as objects of study requires additional levels of work in a research process, compared to the situation where the methods are used mainly as tools to obtain knowledge about a system. In the latter case, the quality of the product being evaluated is the main measure of success – the question is to what extent the product can be said to fulfill its purposes and goals. The method used in the evaluation process is typically described and explained mainly or only to show that the results are valid and reliable. The main focus of evaluation reports in such cases is on the findings regarding the system evaluated.

When, on the other hand, evaluation methods *per se* are the object of study other aspects must also be taken into consideration. Here, the focus is on the purposes and goals of the process, i.e. the use of the method, and the measures of success are not the same as when methods are used as tools. Instead, when the focus is on the method, the purposes and goals of the method are used to identify the extent to which the method is successful, i.e. whether or not the method fulfills its purposes and goals. In this case a detailed description of the research process is necessary in order to fully investigate the use of the method, in this case an evaluation method. Thus there has to be an analysis in order for results to be delivered. To understand methods as objects of study, it seems reasonable also to use them as tools for only then can they be fully understood and their applicability judged. Finally, higher-level methods, or meta-methods, need to be described as in the method-as-tool case, in order to show that the results in this process are also valid and reliable.

In Chapter 2 and Chapter 3 a number of conceptual distinctions related to methods in general were discussed. Some of these distinctions can be used in order to generalize the findings from this study to the evaluation of fun and entertainment in the context of web usability on a more general level.

These theoretical distinctions can help in the understanding, for example, of the measures of success that could be used when judging specific methods. On the basis of discussions in Chapter 3, one possible solution is to decide whether the method is *process- or product-oriented*. On the basis of this it is then possible to identify whether it is the process or the product of the method that is in focus when choosing suitable measures of success. Another possibility in judging the applicability of evaluation methods in the context of web usability is to use existing *lists of heuristics for the evaluation of methods*. Two examples of heuristic lists were discussed in Chapter 3 and on the basis of these, more generalizable conclusions are drawn from the empirical findings from the study in this thesis.

Another issue to be taken into account when evaluating entertainment and fun in the context of web usability is which aspects of entertainment and fun should be evaluated when using the methods. As discussed above, mainly in Chapter 2, users of web sites may have a wide range of experiences that can be classified as fun, entertainment etc. Not all of them are relevant to web-site evaluation, since many of them have nothing or very little to do with the web sites *per se*. It is important to consider the distinction between relevant and irrelevant aspects of fun and entertainment. One possibility is to identify whether the intended fun and entertainment associated with a web site are related to *form* or to *content*, whether fun linked to form can be meaningfully measured, or whether fun included in the

content is beyond the scope of such evaluations. For instance, when evaluating a joke, it is possible to evaluate among other things, the context in which the joke appears, whether the receiver can understand the language of the joke, etc. The extent to which the joke is fun *per se,* however, is largely beyond the reach of the evaluation, since this is very individual, situational, and culturally dependent.

The above-mentioned distinctions, that is (1) methods as tools vs. methods as 'objects of study' (2) process- vs. product-oriented methods (3) application of heuristics of useful methods, and (4) division of entertainment web sites into form vs. content, are developed in more detail below. Furthermore, other possible approaches that could be used for evaluation of fun and entertainment in the context of web usability are discussed. In addition, potentially problematic ethical aspects regarding evaluation, entertainment and the relation between research and practice are highlighted. Finally, possible directions for future work on the basis of this study are also proposed and concluding remarks are made regarding what the results in the study might mean on a more general level, not only in the context of evaluation of entertainment web sites.

Methods as tools vs. methods as 'objects of study'

The results of the usability evaluations in this thesis where the entertainment web sites were the objects of study, both types of methods, inspection methods as well as empirical usability evaluation methods, were equal, i.e. both types of methods provided fruitful, understandable and useful results, from a designer's point of view. However, where evaluation methods were the object of study the results of using these two types of methods were more unequal.

In the inspection method part of the study, the experts were all lecturers or researchers in HCI, which was an advantage both in regard to the quality of the reflections over the use of the methods and in that they had no problems adding additional methods for evaluation of the methods. However, the methods used were all single expert methods, which provided no opportunity for the evaluators to observe the use of the methods. This information had to be collected later in interviews.

The situation differed in the case of empirical usability evaluations. The subjects were, in general, not sufficiently educated to make informed judgments about the evaluation method per se. Some of the subjects could be considered to be skilled enough to make these judgments. The input from these subjects was a valuable resource in the study. However, in general, the subjects should be regarded as novices in HCI research. Adding additional methods into these user sessions was

not an alternative because the subjects could not be expected to conduct this type of evaluation of their own participation in the evaluations. The only source of input to this evaluation of methods as objects of study was the observations made by the evaluators.

This unequal situation could be seen as a problem, but it could also be regarded as a contribution made by the study, as became clear during the performance of the study. In the initial phases there was no realization that these differences were so obvious but as the study proceeded, it became obvious that the input for further revision and re-design of the empirical usability evaluation methods clearly differed regarding the number of suggestions and creativity. In the part where inspection methods were judged and re-designed the situation was the opposite. The experts involved produced fruitful input, both regarding the use of additional methods for evaluation of the methods as well as in interviews. The experts seemed to be endless sources of insights, an important consideration when judging evaluation methods.

Process-oriented vs. product-oriented usability evaluation methods

Most of the evaluation methods included in the study should be considered to be process-oriented, as discussed in Chapter 3. Only the Heuristic Evaluation method was seen as product-oriented. The reason for this is that this method was the only one that involved any normative requirements of the object of study, i.e. the web site. As mentioned, the distinction between process- and product-oriented methods may be a valuable tool for judging the methods. If the method is to be considered as process-oriented, the judgment focuses on the process of evaluation, or evaluation, of the method. In product-oriented approaches, the judgment focuses on the properties of the product.

This may be used as an explanation for some of the results in the study. In the Design Walkthrough example, the experts made a lot of comments about how the *process* was designed, such as the fact that it was too structured. As this approach did not in any way address the properties of the web sites, the focus was only on the process. In the case of Heuristic Evaluation, the opposite was true. Here, the experts made no or only a few comments about the process. It seemed of less importance – to some extent already decided in advance. On the other hand, the heuristics were elaborated on in all the interviews. The set of heuristics was completely revised, re-designed and finally supplemented with additional heuristics, until a majority of the group of experts considered the set suitable. This is a good example of how product-oriented methods are judged and valued.

Judgments on the basis of heuristics for methods

In some examples of research *about* methods, heuristics for judging methods in order to obtain useful and applicable methods have been developed, as discussed in Chapter 3. Two examples of this kind of heuristics were developed by Khan & Prail (1994) and Muller et.al. (1993). The heuristics developed by Khan & Prail consider methods used in design in general, whilst the heuristics developed by Muller et.al. concern usability evaluation methods specifically. It is not possible or fruitful to use all the heuristics in the context of evaluation methods within this thesis for various reasons, but on the basis of the findings of the study, some of them must be considered very useful and applicable in judging the applicability of the included methods.

With regard to the heuristics developed by Khan & Prail (1994) it is important to consider the extent to which they are applicable to process- and/or product-oriented methods. Another restriction in the context of these heuristics is that they are mainly applicable in cases where evaluations are conducted within a larger context, i.e. in a design process, since the majority of the heuristics use judgments linked to designers. In the context of this study, some of these heuristics were actually used in the collaboration with the designers involved in the project.

The heuristics developed by Muller et.al. (1993) cover, for instance, the number of unique classes of usability problems found by each method, the proportion of serious problems, the relation between benefit and cost, and the likelihood of finding problems undiscovered by other methods. Here, judgments of evaluation methods are not dependent on the presence of designers, which is an advantage with these heuristics compared to the first group. On the other hand, these heuristics are mainly intended to be used in studies where two or more methods are used, as all of the heuristics are based on comparisons between different methods. As a large number of methods were used in the context of this study, it was possible to apply these heuristics. To mention just one example – the heuristic of 'uniqueness', i.e. the likelihood of finding problems in a method, undiscovered by other methods – in this case the improved level of applicability in the developed and re-designed methods overall in the study easily can be measured. This is true both for empirical usability evaluation methods and inspection methods.

Division of entertainment web sites into form vs. content

In this study the entertainment web sites were regarded as including a certain form and a certain content. The content was understood as the message of the EWS and was closely linked to the originators' purpose of the EWS, as understood by designers. The form included graphic form, overall structure and navigation and finally, what was labelled 'added value'. Dividing EWSs into form and content as was done in this study was fruitful in that it highlighted what was included amongst the aspects to be considered in evaluations and what was not. More specifically, the content was considered to be 'beyond reach' and was not included in the aspects of the web sites that were evaluated. The important thing to consider was the form. The rationale behind this was the focus mainly on aspects that could be controlled by designers. Generally in all EWSs the form was what the designers had added to the content, which had mainly been provided by the originator of the web site.

Other possible approaches to evaluate fun and entertainment.

The approach to the evaluation of fun and entertainment proposed in this thesis represents just one possible research strategy on this issue. The general problem of evaluating experience, as mentioned above, is very broad. This study addressed only a subset of potential research questions and explored a subset of potentially applicable approaches. We believe further studies are needed in this area.

In the context of Usability Engineering, there are a large number of alternative approaches available. For instance, other usability evaluation methods could have been chosen as objects of study. The methods chosen in this study are generally well-known and widely used in the HCI research community. But even with this as the criterion for choice, other alternatives would have been possible, such as, Cognitive Walkthrough, Focus Group Evaluation and Feature inspection. Other heuristics could also have been used for Heuristic Evaluation, for instance, the list of 'Eight Golden Rules of Interface Design' as presented by Schneiderman (1998, pp. 74-75) or Keith Instone's heuristics for web usability (See http://www.usableweb.org). There are also a number of quantitative methods used in Usability Engineering which could have served as alternative methods for evaluation, covering such aspects as 'time to finish task', number of errors' and 'rate of errors' (Nielsen, 1993) to mention just a few.

There are a number of other methodological alternatives in related research into evaluating fun, pleasure and experiences. A number of statistical methods are used, to analyze these aspects of design, for instance Kansei Engineering (Nagamachi, 1995), various kinds of Factor Analysis (c.f. Schenkman & Jönsson, 2000) and techniques where, for instance, the counting of smiles is applied (c.f. Höök et. al., 2000).

There are also a number of standard protocols for studies concerning the pleasure of products and systems, one example being SEQUAM – Sensorial Quality Assessment (Bonapace, 1999). These methods, or protocols, are developed and standardized as structured and standardized 'toolboxes', available to anyone to pick and use – researcher or practitioner alike.

As regards the theoretical basis for analysis of the results, two possible alternatives could be mentioned, theories related to entertainment and fun and theories traditionally used in the context of HCI research. Even if, as in the first case, the choice is limited to related theories in entertainment, pleasure, affective computing and theories of methods, a large number of alternative theories exists. This is also true of possible theories in the HCI research field, Activity Theory, Language/Action Theory, Distributed Cognition, etc, to mention just a few. For the interested reader a number of publications provide detailed descriptions of each theory. Another possible information source is to search for comparisons between the theoretical approaches in the area of HCI. (c.f. Kaptelinin et.al., 2003).

The actual choice of theoretical foundation is an issue that is widely discussed in research in general. Some argue that researchers should remain faithful to the theories they usually use, to avoid eclecticism, i.e. 'choose the method that suits the purpose in every specific case'. Others argue that there is no 'generally applicable theory', which leaves us only the alternative of trying to find the most suitable theory in every given case.

Would the results differ if other methods had been used in the evaluations conducted, if other subjects and experts had been chosen and if other theories had been used in the analysis? The answer to that question is unconditionally – yes. As in all research, the results must be considered on the basis of the methodological choices made throughout the study. The aim in this study is to make a contribution to the research communities of HCI and Informatics and hopefully, the choices made have produced interesting results.

General issues in relation to evaluating entertainment and fun

When conducting any type of research, a number of questions should be asked regarding ethical aspects. This is also the case in the context of evaluating fun, entertainment and entertainment web sites. Examples of such questions are whether we could, should and/or want to evaluate entertainment (at all). Another possible question to ask is why we should evaluate entertainment. Becoming involved in design processes, as was the case in this study, raises a number of issues about design and responsibilities for the product or system designed. For instance, one problem is where the actual responsibility for the design and its consequence lies. A high level of interactivity between research and practice was involved in this study which raises the question of whether the purpose of researchers should be to 'help' people working in practice or only to question their work.

These issues have not been thoroughly covered here, as they should be seen to fall outside the scope of this thesis. However, some comments may be made in relation to these issues. For instance, whether entertainment is, or should be, evaluated is interesting in view of the questions asked in this thesis. In general, evaluation of entertainment is a common activity in society worldwide. Movies and books are continuously reviewed in the media for instance. In culture, as in arts, this is part of the game that goes on, in and around cultural events.

Internationally, there are more examples of evaluations of, or competitions in entertainment. Perhaps the biggest event is *The Academy Awards* commonly known as *The Oscars* – the movie academy awards given by the Academy of Motion Picture Arts and Science in the United States. This is an enormous event, viewed worldwide by billions of people in 150 countries. The awards are divided into such categories as the 'best picture', 'best actor and actress in leading roles', 'best music' etc. The voters involved in this process are thousands of people working in the American film industry, actors, producers, directors etc. – comparable with experts in evaluation. Other, similar events are *The Golden Globe Awards*, for motion pictures as well as television series and films. There are numerous examples, in the music industry such as the MTV Music Awards, where the worldwide music channel MTV gives awards of various kinds to the 'best song', 'best album', 'best female singer', 'best group', best R'n B', etc. Here, the 'people's choice' is used as a basis for deciding winners – comparable with empirical evaluation in the context of usability evaluation. Another example, discussed earlier in the context of this thesis is the *Eurovision Song Contest.*

It might still be argued that it is impossible to evaluate entertainment, but the *Eurovision Song Contest, The Oscar's, Golden Globe Awards* etc., are examples where we choose to do it anyway, despite arguments that 'it is impossible'. Whether this can be transferred to the context of entertainment as a part of information technologies is another question – but one that hopefully will be answered in the near future.

Concluding remarks

Possible approaches for future work in the evaluation of aspects of fun and entertainment in the context of web usability may include, but are not restricted to, the following: (1) Evaluations of other types of entertainment IT artifacts than those covered in this study, with traditional usability evaluation methods in order to refine and develop these methods further. (2) Evaluation of entertainment web sites using other types of usability evaluation methods than those in this study, in order to find possible ways to make these better suited to their purposes in informing the design of EWSs. (3) Draw on other disciplines and areas within or outside academia, in the context of fun and entertainment, in order to find out how these areas evaluate fun and entertainment. These influences can be used as inspiration in designing and developing new methods and techniques for evaluating usability in the context of fun and entertainment. (4) Base the research on other theories than those used in this study. Examples of areas where suitable theories may be found are perhaps other theories about fun and entertainment than covered in this thesis, or theories in some way connected to HCI research. The chosen theories may also be used for different purposes, i.e. they may have roles in the research process that differ from those in the process in this study. Examples of this are letting theory inform the development of methods on the basis of a theoretical analysis of the EWSs, or to conduct evaluations on EWSs followed by a theoretical analysis of the empirical results with further development of the methods as a final step in the research process. Other alternatives would also be possible.

Overall, findings from the study in this thesis indicate that valuable findings for designers regarding aspects of fun and entertainment in entertainment web sites can be obtained if evaluations are conducted using applicable evaluation methods. Thus, it is extremely important to continue the effort to develop methods and techniques for usability evaluation – both for inspection methods and empirical usability evaluation methods. When it comes to the development of inspection methods, the challenges include finding proper heuristics to support the experts in using Heuristic Evaluation, providing conditions for experts which bridge the gap

between evaluation and authentic use, developing complementary methods for use in combination with existing methods etc. In empirical evaluation of entertainment in the context of web usability, the most crucial aspect might be to consider how to arrange a setting that is as natural and authentic as possible when evaluating fun, as this seems to be important for the results. In addition, it is crucial to consider carefully the level of intervention. The setting up of such conditions as testing in pairs and providing unstructured tasks are just one step on the road to success in evaluating fun and entertainment in the context of web usability. Further steps have to be taken and other conditions have to be explored.

The issue of operationalizing fun and entertainment in the context of usability evaluation is critically important – in the context of entertainment web sites as well as in web usability in general. In this study, a number of theoretical frameworks were consulted to find a suitable basis for operationalization. However, the result of this investigation shows that it was difficult to find a framework that was useful for this purpose. However, one possible explanation for this is that theories, applicable for operationalization of fun and entertainment in the context of web usability, do exist but were not covered in the theoretical investigation in this thesis. Another explanation is that the very nature of entertainment and fun, i.e. that fun and entertainment are exploratory, situated and subjective, makes it extremely difficult, if not impossible, to fully operationalize it to conduct controlled experiments and evaluations. On the basis of the findings of this thesis it is difficult to decide which of the possible explanations is true. However, the findings from this study do indicate that solutions may be found even if it is the case that fun and entertainment in this context are difficult or impossible to operationalize on the basis of general theories. This study was based on the characteristics and criteria of the system to be evaluated, in this case EWSs. As it turned out, operationalization based on these characteristics helped in operationalization of fun and entertainment. It might be the case that only then can the results of evaluations of fun and entertainment be sorted into those that fall within, and those that are beyond, the scope of what is being measured.

The study reported in the thesis also makes it possible to draw the following conclusion. As mentioned in the Introduction and Chapter 1, extending the scope of usability to include fun and entertainment may evoke criticisms. A potential argument against this extending is that usability and fun are two different and independent aspects of system's quality: for instance, a system can be perfectly usable even if designers fail to make it fun and entertaining. However, the experience of conducting this study indicates that considering fun and entertainment as aspects of usability has important advantages, at least in

the case of entertainment web sites. Not only does it allow employing existing methodology of usability evaluation for dealing with fun and entertainment, it also provides evidence about the relationship between functional and "experiential" aspects of a web site design. In addition, in our experience, considering fun and entertainment as aspects of web usability, makes perfect sense for practitioners, that is, web designers. Therefore, it appears a more promising way to address new issues and concerns in system design is extending the scope of traditional usability rather than creating additional separate fields of research and practice.

The strongest impression after all the use sessions conducted in this study as well as in other related projects, is how extremely important the findings are, when external sources are used as subjects in an empirical usability evaluation. No matter how many design awards or prizes the designs or designers have won, and no matter how experienced the expert conducting expert evaluations is – it will always be impossible to predict everything that happens when authentic users of a system are investigated. Even if we feel dissatisfied with our methodologies, and even if we have to struggle to meet challenges in designing and conducting these evaluations, the effort is always worthwhile, considering the interesting and important results this type of evaluations produce.

Hopefully, this study will encourage and inspire readers to continue research along the lines of this thesis. It is important to ensure that usability is regarded as a key aspect that needs to be considered and incorporated into the design of IT artifacts.

References

Aboulafia, A., L. Bannon, et al. (2001). "Shifting Perspective from Effect to Affect: Some Framing Questions." Proceedings of The International Conference on Affective Human Factors Design: 508-514.

Agarwal, R. &Karahanna, E. (2000) Time Flies When You're Having Fun: Cognitive Absorption and Beliefs About Information Technology Usage. MIS Quarterly Vol. 24 No. 4, pp. 665-694. December 2000

Anderson, D., Sweeney, D., Williams, T.A. (2002)Statistics for Business and Economics. South Western College Publishing Cincinnati.

Alvesson, M., Sköldberg, K. (1994). Tolkning och reflektion. Vetenskapsfilosofi och kvalitativ metod. Studentlitteratur. Lund.

Amant, R. S. and R. M. Young (2001). "Artificial Intelligence and Interactive Entertainment." Intelligence (Summer 2001): 17-19.

Bates, J. (1994). "The Role of Emotion in Believable Agents." Communications of ACM 37(7): 122-125.

Bevan, N. (1998). Usability Issues in Web Site Design. In Proceedings of UPA'98. Washington DC.

Bell, B. (1992). Using programming walkthroughs to design a visual language. Technical Report CU-CS-581-92. Ph.D. diss., University of Colorado, Boulder, CO (Reference in Usability Inspection Methods. Mack, R.L. & Nielsen, J. (eds.) John Wiley & Sons, Inc. New York.)

Bias, R.G. (1994). The Pluralistic Usability Walkthrough: Coordinated Empathies. In Usability Inspection Methods. Mack, R.L. & Nielsen, J. (eds.) John Wiley & Sons, Inc. New York. (63-73)

Bittanti, M. (2002). The Technoludic Film: Images of Video Games in Movies (1973-2001). Entertainment Computing: Technologies and Applications. Proceedings of IFIP First International Workshop on Entertainment Computing (IWEC 2002). R. Nakatsu and J. Hoshino. Boston, MA, Kluwer Academic Publishers.

Blythe, M. & Wright, P. (2003). From usability to enjoyment. Introduction in Funology: From usability to Enjoyment. (eds.) Blythe, M., Overbeeke, K., Monk, A.F., Wright, P. Human-Computer interaction series Vol.3. Kluwer Academic Publishers. (pp. XIII-XIX)

Bolter, J., D. and R. Grusin (2002). Remediation. Understanding New Media. Boston, MA, The MIT Press.

Bonapace, L. (1999). The ergonomics of pleasure, in W.S. Green and P.W. Jordan (eds) Human Factors in Product Design: Current Practice and Future Trends, London: Taylor & Francis, pp. 234- 248.

Bonner, J. V. H. (2002). Envisioning Future Needs: From Pragmatics to Pleasure. Pleasure with Products. W. S. Green and P. W. Jordan. London, Taylor & Francis: 151-158.

Borges, J. A., I. Morales, et al. (1996). Guidelines for Designing Usable World Wide Web Pages. In Proceedings of ACM Conference on Human Factors in Computing Systems, CHI'96.

Borges, J. A., I. Morales, et al. (1998). Page Design Guidelines Developed Through Usability Testing. Human Factors and Web Development. C. Forsythe, E. Grose and J. Ratner. Mahwah, NJ, USA, Lawrence Erlbaum Associates, Publishers.

Carroll, J.W. & Thomas, J.C. (1988). Fun. SIGCHI Bulletin. January 1988, Vol.19 No.3.

Chen, H., Wigand, R.T., Nilan, M.S. (1999). Optimal experience of Web activities. Computers in Human Behavior 15(1999) (585-608). Pergamon, Elsevier Science Ltd.

Chin, W.W. & Lee, M.K.O. (2000). A Proposed Model and Measurement Instrument for the Formation of IS Satisfaction: The Case of End-user Computing Satisfaction. In Proceedings of the 21st International Conference on Information Systems, ICIS, 553-563

Csíjszentmihályi, M. (1990). Flow. The psychology of optimal experience. Harper and Row, NY.

Creusen, M. and D. Snelders (2002). Product Appearance and Consumer Pleasure. Pleasure with Products. W. S. Green and P. W. Jordan. London, Taylor & Francis: 69-75.

Danielsson, K. & Wiberg, C. (2002). IT Basketball – A sporty Virtual Environment: An Evaluation of Usability, Presence and Interest. In proceedings of IRIS25 Informations systems Reserch seminar In Scandinavia.

De Angeli, A. (2001). The Unfriendly User. Exploring Social Reactions to Chatterbots. Proceedings of International Conference on Affective Human Factor Design, Singapore.

De Angeli, A., P. Lynch, et al. (2002). Pleasure versus Efficiency in User Interfaces: Towards an Involvement Framework. Pleasure with Products. W. S. Green and P. W. Jordan. London, Taylor & Francis: 97-111.

Desmet, P. (2003). Measuring emotion: Development and application of an instrument to measure emotional responses to products. In Funology: From usability to Enjoyment. (eds.) Blythe, M., Overbeeke, K., Monk, A.F., Wright, P. Human-Computer interaction series Vol.3. Kluwer Academic Publishers. (111-123)

Desurvire, H., J. Kondziela, et al. (1992). What is Gained and Lost when Using Usability Methods Other than Empirical Testing. Proceedings of ACM CHI'92 Conference on Human Factors in Computing Systems -- Posters and Short Talks: 125-126.

Diaper, D. (1989) Task observation for human-computer interaction. In Diaper, D. (Ed.), Task Analysis for Human-Computer Interaction. Ellis Horwood, Chichester, 210-237

Dix, A., Finlay, J., Abowd, G., Beale, R. (1998). Human-Computer Interaction. Prentice Hall, Europe.

Draper, S. W. (1999). "Analysing fun as a candidate software requirement." Personal Technology 3(Special issue: Computers and fun): 117-122.

Dyer, R. (1992). Only Entertainment. Routledge. London.

Ehn, P. & Löfgren, J. (1997) Design of Quality-in-use:Human-Computer Interaction Meets Information Systems Development. In Helander, M., Landauer, T.K, Prabhu, P.(eds.) Handbook of Human Computer Interaction Second, completely revised edition. Elsevier Science B.V.

Evans, E.A. (1993). A Modular Design for User Satisfaction Assessments. In ACM SIGUCCS XXI 1993. 325-329.

Fabricatore, C., Nussbaum, M., Rosas, R. (2002). Playability in Action Videogames: A Qualitative Design Model. In Human-Computer Interaction, 2002, Vol. 17, (311-368)

Federoff, M. (2003). Improving games with User Testing: Getting Better Data Earlier. Game Developer Magazine, June 2003.

Fjellman, E. & Sjögren, J. (2000). Interaktiv underhållning inför framtiden. Telematik 2004. KFB-rapport 2000:10 & TELDOK Rapport 133. Stockholm.

Forlizzi, J. & Ford, S. (2000). The Building Blocks of Experience: An Early Framework for Interaction Designers. In Proceedings of Designing Interactive Systems 2000 (DIS'00), Brooklyn, NY

Frank, A. and N. Lundblad (2002). The New Role of Gaming. How Games Move Outside Entertainment. Entertainment Computing: Technologies and Applications. Proceedings of IFIP First International Workshop on Entertainment Computing (IWEC 2002). R. Nakatsu and J. Hoshino. Boston, MA, Kluwer Academic Publishers.

Gaines, B. R., Shaw, M. L. G., Chen, L. L-Y. (1996). Utility, Usability and Likeability: Dimensions of the Net and Web. In Proceedings of WebNet96. San Francisco, CA, AACE (Association for the Advancement of Computing in Education.

Gray, W.D. & Salzman, C. (1998). Damaged Merchandise? A Review of Experiments That Compare Usability Evaluation Methods. Journal of Human-Computer Interaction. Special issue: Experimental comparisons of usability evaluation methods. Vol.13, No.3. 1998. 203-262

Grose, E, Forsythe, C., Ratner, J. (1998). Using Web and Traditional Style Guides to Design Web Interfaces. In In Forsythe, C., Grose, E., Ratner, J., Human Factors and Web Development. Lawrence Erlbaum Associates, Publishers. Mahwah, NJ, USA.

Harrison, A.W. & Rainer Jr, R.K. (1996). A General Measure of User computing Satisfaction. In Computers in Human Behavior, Vol.12 (1996), No. 1, 79-92

Huang, M-H. (2003). Designing web site attributes to induce experiential encounters. In computers in Human Behavior 19 (2003) 425-442

Höök, K. (2000). Steps to take before IUIs become real, Journal of Interaction with Computers, Vol. 12, no. 4, February 2000

Höök, K., P. Persson, Sjölinder, M. (2000). Evaluating Users' Experience of a Character-Enhanced Information Space. Journal of AI Communications 13(3): 195-212.

Höök, K., Svensson, M. (1999) Evaluating Adaptive Navigation Support. In A. Munro, K. Höök, and D. Benyon (eds.), Footsteps in the snow: personal and social navigation in information space, Springer Verlag.

Jefferies, R., Miller, J.R., Wharton, C., Uyeda, K. (1991) User interface evaluation in the real world: a comparison of four techniques. In proceedings of CHI'91, Conference on Human Factors and Computing Systems, New Orleans, LA. 119-124

Jegers, K. & Wiberg, C. (2001). Evaluating Experience: Implications for usability tests conducted on entertainment web sites. In proceedings of IRIS24 Informations systems Research seminar In Scandinavia.

Jegers, K. & Wiberg, C. (2003a) Satisfaction and Learnability in Edutainment: A usability study of the knowledge game 'Laser Challenge' at the Nobel e-museum. In Proceedings of the International Conference of Human Computer Interaction, Krete, Greece, June, 22-25, 2003

Jegers, K. & Wiberg, C. (2003b) FunTain: Design Implications for Edutainment Games. In Proceedings of ED-MEDIA 2003, AACE, Honolulu, Hawaii, June 22-25, 2003

Jensen, J.F. (2000) Trends in Interactive content & Services. In proceedings of WebNet2000, San Antonio Texas (pp. 281-286)

Jordan, P.W. (1999). Pleasure with Products: Human factors for body, mind and soul. In W.s Green and P.W. Jordan (eds.) Human Factors in Product Design: Current Practice and Future Trends. (pp. 179-188) Taylor & Francis. London.

Jordan, P.W. (2000). Designing Pleasurable Products. An Introduction to the New Human Factors. Taylor & Francis, London.

Kaasgaard, K. (2000). Software Design & Usability. Copenhagen, Denmark, Copenhagen Business School Press.

Khan, M.J. & Prail, A. (1994). Formal Usability Inspections. In Usability Inspection Methods. Mack, R.L. & Nielsen, J. (eds.) John Wiley & Sons, Inc. New York. (141-172)

Kaptelinin, V. & Nardi, B., Bodker, S., Carroll, J., Hollan, J., Hutchins, E., Winograd, T. (2003). Post-cognitivist HCI: Second-Wave Theories. Panel on the CHI 2003 conference, In Proceedings of CHI 2003, (pp. 692-693)

Karat, J. (1997). User-Centered Software Evaluation Methodologies. Handbook of Human-Computer Interaction. M. Helander, T. K. Landauer and P. Prabhu, Elsevier Science B.V.: 689-704.

Karat, C.-M., R. Campbell, et al. (1992). Comparison of Empirical Testing and Walkthrough Methods in User Interface Evaluation. Proceedings of ACM CHI'92 Conference on Human Factors in Computing Systems: 397-404.

Karat, J., Jeffries, R., Miller, R.M., Lund, A.M., McClelland, I., John, B.E., Monk, A.F., Oviatt, S.L., Carroll, J.M., Mackay, W.E., Newman, W.M., Olson, G.M., Moran, T.P. (1998). Commentary on "Damaged Merchandise?" Journal of Human-Computer Interaction. Special issue: Experimental comparisons of usability evaluation methods. Vol.13, No.3. 1998. 199-201

Karat, J. & Karat, C-M. (2003). That's entertainment! In Funology: From usability to Enjoyment. (eds.) Blythe, M., Overbeeke, K., Monk, A.F., Wright, P. Human-Computer interaction series Vol.3. Kluwer Academic Publishers. (125-136)

Karvonen, K.(2000). The Beauty of Simplicity. In Proceedings of CUU'00, Arlington, VA ACM

Kvale (1997). Den kvalitativa forskningsintervjun. Studentlitteratur. Lund

Langer, S.K. (1977). Feeling And Form. A theory of art developed from Philosophy in a New Key. Prentice Hall, New Jersey.

Laskowski, S., Downey, L., L. (1997) Evaluation in the Trenches: Towards Rapid Evaluation. In Proceedings of ACM Conference on Human Factors in Computing Systems, CHI'97.

Laurel, B. (1993). Computres as Theatre. Addison-Wesley. Reading, MA.

Lewis, C., P. Polson, et al. (1990). Testing a Walkthrough Methodology for Theory-Based Design of Walk-Up-and-Use Interfaces. Proceedings of ACM CHI'90 Conference on Human Factors in Computing Systems: 235-242.

Lewis, C. & Warthon, C. (1997). Cognitive Walkthroughs. In Handbook of Human-Computer Interaction, Secnd, completely revised edition M. Helander, T.K. Landauer, P.Prabhu (eds.) Elsevier Science B.V (717-732)

Lindgaard, G. & Dudek, C. (2003). What is this evasive beast we call user satisfaction? Interacting with computers 15 (2003), 429-452

Löwgren, J. (1993). Human-computer interaction. What every system developer should know. Studentlitteratur, Lund.

Mack, R.L. & Nielsen, J.(1994). Executive Summary of Usability Inspection Methods. John Wiley & Sons, Inc. New York.

McLuhan, M. (1994). Understanding Media: The Extensions of Man. MIT Press. Reprint edition.

Mahmood, M.A., Burn, J.M., Gemoets, L.A., Jacquez, C. (2000). Variables affecting information technology end-user satisfaction: a meta-analysis of the empirical literature. International Journal of Human-Computer studies (2000) 52, 751-771.

Malone, T. W. (1980). What Makes Things Fun to Learn? Heuristics for Designing Instructional computer Games. Proceedings of the joint symposium of Third SIGSMALL symposium and the first SIGPC symposium, September, 1980.

Malone, T. W. (1982). Heuristics for Designing Enjoyable User Interfaces: Lessons from Computer games. Proceedings of Conference on Human Factors and Computing Systems, CHI'82.

Marcus, A. (2002). "The Cult of Cute: The Challenge of User Experience Design." Interactions: 29-34.

Mayhew, D.J. (1998). Introduction in In Forsythe, C., Grose, E., Ratner, J., Human Factors and Web Development. Lawrence Erlbaum Associates, Publishers. Mahwah, NJ, USA.

Monk, A. (2002). Fun, communication and dependability: extending the concept of usability. Closing plenary at HCI2002, (for Human factor Advanced Module)

Monk, A.F. & Frolich, D. (1999(. Computers and Fin, Personal Technology, 3(1).

Monk, A. And Gilbert, N. (1995). Inter/disciplinary research. In Monk, A. and Gilbert, N. (eds.) Perspectives on HCI. Diverse Approaches. Academic Press. London

Monk, A., M. Hassenzahl, et al. (2002). Funology: Designing enjoyment. Conference on Human Factors and Computing Systems,CHI 2002, Minneapolis, Minnesota, USA.

Muller, M. J., T. Dayton, et al. (1993). Comparing Studies that Compare Usability Assessment Methods: An Unsuccessful Search for Stable Criteria. Proceedings of ACM INTERCHI'93 Conference on Human Factors in Computing Systems -- Adjunct Proceedings: 185-186.

Nagamachi, M. (2001). Kansei Engineering – Tutorial at International Conference on Affective Human Factor Design, Singapore.

Nelson, H. & Stolterman, E. (2003) The Design Way. Intentional Change in an Unpredictable World. Educational Technology Publications. New Jersey

Newman, W. M. and M. G. Lamming (1995). Interactive Systems Design. Cambridge, Addison-Wesley.

Nielsen, J. (1993). Usability Engineering. Academic Press.

Nielsen, J. (1994a). Usability Inspection Methods. Presentation of tutorial at CHI'94, Boston, MA.

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

Nielsen, J. (1999). User Interface Directions for the web. In Communications of the ACM. Vol.42, No.1.

Nielsen, J. (2000). Why You Only Need to Test With 5 Users. Jakob Nielsen's Alertbox, March 19, 2000 (http://www.useit.com/alertbox/20000319.html (2003-05-09))

Nielsen, J. (2003). User empowerment and the fun factor. In Funology: From usability to Enjoyment. (eds.) Blythe, M., Overbeeke, K., Monk, A.F., Wright, P. Human-Computer interaction series Vol.3. Kluwer Academic Publishers. (103-105)

Nielsen, J. & Landauer, T.K.(1993). Mathematical Model of the Finding of Usability Problems. Proceedings of INTERCHI'93. 206-213

Nielsen, J. & Philips, V.L. (1993). Estimating the Relative Usability of Two Interfaces: Heuristic, Formal and Empirical Methods Compared. In proceedings of INTERCHI'93, Amsterdam, The Netherlands. 214-221.

Nielsen, J. and R. Molich (1990). “Heuristic Evaluation of User Interfaces.” Proceedings of CHI’90: 249-256.

Norman, D. (2002). Emotion & design: attractive things work better. Interactions, Vol. 9, No. 4 (July 2002), 36-42

Norman, D. (forthcoming). Emotional Design: Why We Love (or Hate) Everyday Things. New York, Basic Books.

Novak, T.P., Hoffman, D.L., Yung, Y-F. (1998). Measuring the Flow Construct in Online Environments: A Structural Modeling Approach. Marketing Science and the Internet Mini-Conference, MIT.

Noyes, J. and R. Littledale (2002). Beyond Usability, Computer Playfulness. Pleasure with Products. W. S. Green and P. W. Jordan. London, Taylor & Francis: 49-59.

Olson, G.M. & Moran, T.P.(1998). Introduction to this Special Issue on Experimental Comparisons of Usability Evaluation Methods. Journal of Human-Computer Interaction. Special issue: Experimental comparisons of usability evaluation methods. Vol.13, No.3. 1998. 199-201

Olsson, C. (2000a) To Measure or Not to Measure: Why Web Usability Is Different From Traditional Usability. In proceedings of WebNet2000, San Antonio Texas (pp. 425-430)

Olsson, C. (2000b). The usability concept re-considered: A need for new ways of measuring real web use. In proceedings of IRIS23 Informations systems Research Seminar In Scandinavia, Doing IT Together. L. Svensson, U. Snis, C. Soerensen, H. Fägerlind, T.Linderoth, M. Magnusson, T. Östlund. Laboratorium for Interaction Technology, University of Trollhättan/Uddevalla.

Ottersten I. & Berntsson, J.(2002) Användbarhet i praktiken. Studentlitteratur, Lund.

Overbeeke, K., Djadjadiningrat, T., Hummels, C., Wensveen, S., Frens, J. (2002). Let's make things engaging. Chapter 1 in Funology: From Usability to Enjoyment. Blythe, M.A., Overbeeke, K., Monk, A. F., Wright, P.C. (eds). Kluwer Academic Publishers

Pagulayan, R.J., Steury, K.R., Fulton, B., Romero, R.L. (2003). Designing for fun: User-testing case studies. In Funology: From usability to Enjoyment. (eds.) Blythe, M., Overbeeke, K., Monk, A.F., Wright, P. Human-Computer interaction series Vol.3. Kluwer Academic Publishers.

Patton, M.Q. (2002). Qualitative Research and Evaluation Methods. Third edition. Sage Publications. Thousand oaks.

Pavlik, J.V. (2000). The Structure of the New Media Industry. In The media and entertainment industries. Greco, A.N. (ed) Allyn & Bacon, Boston, MA (214-247)

Picard, R.W. (1998). Affective Computing. MIT Press.

Pine II, B. J., Gilmore, J. H. (2000) The Experience Economy: Work is Theatre & Every Business a Stage. Harvard Picard, W. R. (1998) Affective computing. The MIT Press. Boston, Massachusetts

Pinhanez, C., C.-M. Karat, et al. (2001a). "Less Clicking, More Watching": An Option for Entertainment on the Web? Conference on Human Factors and Computing Systems,CHI 2002.

Pinhanez, C., C.-M. Karat, et al. (2001b). "Can Web Entertainment Be Passive." Proceedings of WWW01.

Pinhanez, C., C.-M. Karat, et al. (2001c). "Less Clicking, More Watching": Results of the Iterative design and Evaluation of Entertaining Web Experiences. INTERACT'01.

Polson, P., and Lewis, C. (1990). Theory/based design for easily learned interfaces. Human-Computer Interaction, 5, 2&3. (191-220).

Popovic, V. (2002). Activity and Designing Pleasurable Interaction with Everyday Artifacts. Pleasure with Products. W. S. Green and P. W. Jordan. London, Taylor & Francis: 367-376.

Preece, J. (1993). A Guide to Usability. Human Factors in Computing. Addison Wesley, London.

Preece, J. (1994). Human Computer Interaction, Addison-Wesley.

Ratner, J. (1998). Easing the Learning Curve for Novice Web Users. In Forsythe, C., Grose, E., Ratner, J., Human Factors and Web Development. Lawrence Erlbaum Associates, Publishers. Mahwah, NJ, USA.

Redmond-Pyle, D., Moore, A. (1995). GUIDE - Graphical User Interface Design and Evaluation - A Practical Process. Prentice Hall Europe.

Reinmoeller, P. (2002). Emergence of Pleasure: Communities of Interest and New Luxury Products. Pleasure with Products. W. S. Green and P. W. Jordan. London, Taylor & Francis: 125-134.

Roberts, T.L. & Moran T.P.(1982). Evaluation of Text Editors. In Moran, T.P.(ed.)Eight Short Papers in User Psychology. Xerox Palo Alto Research Centers, Palo Alto, CA

Ruecker, S. (2002). Carrying the Pleasure of Books into the Design of the Electronic Book. Pleasure with Products. W. S. Green and P. W. Jordan. London, Taylor & Francis: 135-150.

Sawyer, P., Flanders, A., Wixon, D. (1996). Making a difference: the impact of inspections. In Proceedings of CHI'96, Vancover, British Columbia, Canada. 376-382.

Scheirer, J., Fernandex, R., Klein, J., Picard, R.W. (2002). Frustrating the user on purpose: a step toward building an affective computer. Interactive with Computers 14, 2 (2002), 93-118. TR 509

Schenkman, B. N. and F. U. Jönsson (2000). “Aesthetics and preferences of web pages.” Behaviour & Information Technology 19(5): 367-377.

Schneiderman, B. (1997). Designing information-abundant web sites: issues and recommendations. International Journal of Human-Computer Studies. No. 47, (pp.5-29)

Shneiderman, B., (1998) Designing the User Interface: Strategies for effective Human-Computer Interaction, Third edition. Addison-Wesley

Solso, R.L., Johnson, H.H., Beal, M.K.(1998). Experimental Psychology. A Case Approach. Sixth edition. Longman – Addison Wesley Longman, Inc. New York.

Spool, J., Scanlon, T., Schroeder, W., Snyder, C., DeAngelo, T. (1999) Web Site Usability: A Designer’s Guide. Morgan Kauffman Publishers Inc. business school Press, Boston, Massachusetts

Spool, J.(2002). Testing web sites: Five users is nowhere near enough. Conference on Human Factors and Computing Systems,CHI 2002, Minneapolis, Minnesota, USA.

Stolterman, E. (1991). Designarbetets dolda rationalitet. En studie av metodik och praktik inom systemutveckling. (The hidden rationale of design work – A study in the methodology and practice of system development) Doctoral dissertation. Department of Information Processing, Umeå University. Umeå, Sweden. ISSN: 0282-0579

The new shorter Oxford English Dictionary, vol. 1.

Thomas, P. and R. D. Macredie (1994). “Games and the design of Human-Computer Interfaces.” Education & Training technology international journal of AETT 31(2): 134-142.

Thomas, P. and R. D. Macredie (2002). “Introduction to The New Usability.” ACM Transactions on Computer-Human Interaction, Vol. 9(No. 2): 69-73.

Tiger, L. (1992). The Pursuit of Pleasure. Transaction Publishers. New Brunswick (U.S.A.) & London (UK)

Tractinsky, N., A. S. Katz, et al. (2000). "What is beautiful is usable." Interacting with Computers 13: 127-145.

Wallén, G. (1996). Vetenskapsteori och forskningsmetodik. Studentlitteratur. Lund.

Webster's Third New International Dictionary – Unabridged (Springfield, Mass.: Merriam-Webster, 1961)

Wiberg, C. (2001a). From ease of use to fun of use: Usability evaluation guidelines for testing entertainment web sites. In Proceedings of Conference on Affective Human Factors Design, CAHD, Singapore

Wiberg, C. (2001b). Join the Joyride: An Identification of Three Important Factors for Evaluation of On-line Entertainment. In proceedings of WebNet 2001, Association for the Advancement of Computing in Education, Charlottesville, VA.

Wiberg, C. and Wiberg, M. (2001c). Configuring Social Agents. Proceedings of Conference of Universal Accessability in Human Computer Interaction, New Orleans.

Wharton, C., Rieman, J., Lewis, C., Polson, P. (1994). The Cognitive Walkthrough Method: A Practitioner's Guide. In Usability Inspection Methods. Mack, R.L. & Nielsen, J. (eds.) John Wiley & Sons, Inc. New York. (105-140)

Virzi, R. A. (1997). Usability Inspection Methods. Handbook of Human-Computer Interaction. M. Helander, T. K. Landauer and P. Prabhu, Elsevier Science B.V.: 705-715.

Wixon, D., Jones, S., Tse, L., Casaday, G. (1994). Inspections and Design Reviews: Framework, History and Reflection. In Usability Inspection Methods. Mack, R.L. & Nielsen, J. (eds.) John Wiley & Sons, Inc. New York. (77-104)

Wolf, M. J. (1999). The Entertainment Economy. London, Penguin Books.

Vora, P. (1998). Human Factors Methodology for Designing Web Sites. In Forsythe, C., Grose, E., Ratner, J., Human Factors and Web Development. Lawrence Erlbaum Associates, Publishers. Mahwah, NJ, USA.

Whittaker, S., Terveen, L., Nardi, B.A. (2001) A Reference Task Agenda for HCI. In Carroll, J.M.(ed) Human-Computer Interaction in the New Millenium. ACM Press. Addison Wesley. New York.

Appendix I

Material in the studies

In this appendix, all material given to users and experts is listed and shown, together with brief descriptions of the context in research process.

Mainly two types of material are described in this appendix.

1. Material delivered to subjects and experts
2. Interview questions and questionnaires

It should be noted that the complete material given to experts are published separately in, "Collected data material of Ph D thesis 'A Measure of Fun: Extending the scope of web usability (CDM). The listings below should be seen as showing an overview of material delivered and used in the various phases of the study. In order to clarify the description of the material, any delivered material, interview questions and questionnaires are presented.

Empirical usability evaluations

Background information questionnaires

The subjects started the evaluation session by filling in background information on a questionnaire. The questionnaire is briefly described below:

Written pre-evaluation questionnaire – background information

1. Name (optional)
2. Age
3. Have you visited this web page before?
4. How would you consider yourself as a web surfer in general (Novice – Expert)
5. Do you surf the web often (Yes, every day – No, less than once per day)?
6. Do you use computers a lot in your work or at school (Work mainly with support of computer in my work/school – I never use computers at work/school)?
7. How would you rate your interest in ESC (None whatsoever – Very big)?
8. Are you primarily a MAC or PC user (MAC or PC)
9. What browser do you mainly use (Explorer, Netscape, Other (which?)
10. Have you participated in a user test before? If so, approximately how many?

The empirical evaluation session

The empirical evaluations were partly based on a scenario. This was mainly to get the user to go back in time, as one of the web sites had a limited life. It was the *Eurovision Song Contest* web site, which was an event site for a competition that took place nearly one year before the evaluation.

The scenario for the Eurovision Song Contest

Below, the scenario used in evaluations of the *Eurovision Song Contest* is shown:

Scenario

You have been on a trip and missed the finals of Eurovision Song Contest. You come home the day after the finals and want to get updated about what happened and get a feeling of the whole competition. In front of you, you already have the web site, so you just for have to get going. You should imagine that you have never been to the web site before.

Tasks used for the Eurovision Song Contest evaluation

Tasks – Eurovision Song Contest

1. During the finals there were two hosts. Who were they and what did they do earlier?
2. Please give me the name of the winner, who came second and third, and please state their scores.
3. Please give me four earlier winners of Eurovision Song Contests.
4. Send a postcard.
5. What happened with the Italian member of the jury in 1990?
6. Who won the ESC in 1981 and from what country came the winners?
7. What is the name of the ESC expert on the web site?
8. Create your own ABBA remix and send it to a friend.
9. How many photographers have been involved in the production of this web site?
10. Who came third in the Finnish competition in 2000?
11. Please sign up for a free newsletter.
12. Find out the price of an ESC bag.
13. Luxemburg was not involved in this years' competition. Why not?
14. Ingela Pling Forsman has been involved in writing a number of Swedish songs that have competed in the ESC. How many?
15. What country has won most times? How many times has it won the ESC?
16. Please close the web site.

Post task questionnaire and interviews

As an initial question was asked in this phase about which approach to use when discussing entertainment in relation to the evaluations - oral or written, both approaches were explored. Below, the oral interview questions as well as the written questionnaire questions are shown:

Post-evaluation questions (Eurovision Song Contest)

1. Did you think anything in the test was unpleasant?
2. What do you think about the level of the questions? Easy/hard? Other comments.
3. This time we used tasks – do you think these gave a typical view of how you would use this web site?
4. Other comments regarding the test in itself?
5. Could you relate the web page to any other web page you have visited, i.e. did it remind you of anything else on the web?
6. Was it hard to navigate on the web page? Arguments?
7. What was good or pleasant on the page? Arguments?
8. What was bad or unpleasant on the page? Arguments
9. Would you revisit the page?
10. If I ask you to describe the web page as a car, how would you describe it? Brand, color, standard etc.

Additional questions for pair session

1. Do you think it is more complicated to work in pairs than if you had done the test by yourself?
2. Do you think this type of session – a pair session –gives a typical view of an authentic use situation on this web site?
3. Would either of you have done anything differently than you did in this session?

Expert tests – HCI experts and novices

Material delivered to experts and novices

The experts and novices received handouts as support for their evaluations. The complete handouts can be found in CDM. An overview of the handouts is given below:

- Introduction
- Questionnaire – background information (shown below)
- List of heuristics by Jakob Nielsen
- Evaluation of web site 1
    - Description of web site 1 and specific instructions for this evaluation
    - Report form for *walkthrough*
    - Report for for *heuristic evaluation*
    - Report form for comparison of both approaches and suggestions for new heuristics
- Evaluation of web site 2
    - Description of web site 2 and specific instructions for this evaluation
    - Report form for *walkthrough*
    - Report for for *heuristic evaluation*
    - Report form for comparison of both approaches and suggestions for new heuristics

Below, the background questionnaire for the evaluation is given:

Questionnaire – background information

1. Are you a man or a woman? (Man/Woman)
2. How are you? (20-30, 30-40, 40-50, 50-60)
3. How have you come into contact with the HCI research field?
    a. Through lecturing in HCI
    b. Through research in HCI
    c. Through both lecturing and doing research in HCI
    d. Other:
4. How much experience do you think you have of evaluation of interfaces or systems?
    a. I have never before performed any type of evaluation
    b. I have lectured in evaluation and my students have conducted

evaluations, but I have never, practically, done this by myself
   c. I have a little experience in evaluation of this kind
   d. I have quite a lot of experience of this, as I have conducted a large number of evaluations
   e. I have extensive experience in this, as I both lecture and do research in the specific subject.
   f. Other:

5. How much experience do you have of Heuristic Evaluation with Jakob Nielsen heuristics?
   a. I have never heard of Heuristic Evaluation
   b. I have heard about Heuristic Evaluation, but have never practiced it
   c. I have a good knowledge of the method, but have never used it
   d. I know it well and have used it a couple of times
   e. I know it well and have used it numerous times
   f. I have performed Heuristic Evaluation before, but with another set of heuristics (please specify which set)
6. How much experience do you have of evaluation with Design Walkthrough?
   a. I have never heard of Design Walkthrough
   b. I have heard about Design Walkthrough, but have never practiced it
   c. I have a good knowledge of the method, but have never used it
   d. I know it well and have used it a couple of times
   e. I know it well and have used it numerous times

I have performed Walkthrough before, but of another kind (please specify which kind)

7. How much experience do you have of evaluation of web sites?
   - None
   - A little
   - Average
   - A lot
   - Very extensive

8. How much experience do you have of visiting/using entertainment web sites (in a wide sense)
   - None

- A little
- Average
- A lot
- Very extensive

9. How much experience do you have of evaluating entertainment web sites?

- None
- A little
- Average
- A lot
- Very extensive

Interview questions and questionnaires (experts only)

Questions for experts after completion of the evaluation

1.You have now evaluated Swedish Railways and Skyscraper in various ways. Can you describe what you did and what your procedure was?

2. Did you have any difficulties accomplishing the task?

3. Which web site did you evaluate first? (control only)

4. If we start with this first web site; how did you think the heuristics used worked in relation to the walkthrough session?

5. When you came to the second web site, did you use (perhaps without thinking about it) the heuristics which were fresh in your memory, even in the walkthrough evaluation?

6. Did you have many 'general' comments, that were not necessarily 'problems', which in the end resulted in proposals for new heuristics?

7. Did you have many comments concerning values of entertainment, experiences etc., or are most comments based on function-related aspects?

8. Do you think it is right to evaluate usability on web sites such as Skyscraper? Please elaborate and go into detail.

9. Do you have any suggestion for another method or technique you have tested or read about, which you believe could work as well as, or even better, based on your experiences from your evaluation of these web sites.

10. If you had three wishes about how entertainment web sites should be designed, what would these be?

11. Do you have any further comments from your evaluations of these web sites?

Expert tests – revised heuristics

Material delivered to experts

The experts received handouts to support their evaluations. The complete handouts can be found in CDM. An overview of the handouts is given below:

- Introduction
- Web site 1
  - Description of web site 1 and specific instructions for the evaluation
  - Part 1: Exploration and entertainment
  - Part 2: Evaluation
    - Instructions
    - Evaluation form
  - Part 3: Meta-evaluation (evaluation of the evaluation itself)
    - Instructions
    - Evaluation form
- Web site 2
  - Description of web site 2 and specific instructions for the evaluation
  - Part 1: Exploration and entertainment
  - Part 2: Evaluation
    - Instructions
    - Evaluation form
  - Part 3: Meta-evaluation (evaluation of the evaluation itself)
    - Instructions
    - Evaluation form

Interview questions and questionnaires

The experts were involved in post-evaluation interviews. The questions are shown below:

Questions for experts after completed evaluation

Part 1: The evaluation process in general

1. Now you have evaluated Mosquito and Skyscraper, by freely surfing around, and then using of heuristics. Can you describe what you did and how you did it.
2. Did you have any difficulties completing the task?
3. What did you think of your evaluation process this time compared with your last evaluation (free surfing this time and then heuristics, compared to Design Walkthrough and Heuristic Evaluation with Nielsen's heuristics)?
4. Do you think that either of these approaches (as you used this time) would have worked as a stand-alone approach for evaluating these web sites, without using the other?
5. Do you miss any of the stages or anything else, from your earlier evaluation, which you think would to enrich your evaluation results?

Part 2: The meta-evaluation

1. In the third part of every evaluation this time you were asked, in a way, to evaluate your own evaluation, or conduct a meta-evaluation. Do you have any comments about this – did you understand what you were supposed to do?
2. If we look through the heuristics, how have you given your points, on the basis of relevance or suitability for the evaluated web sites? Can you point to any heuristic, which you believe is very good or bad?
3. You were also asked to comment and make suggestions for changes in the heuristics, or to propose new ones. If you have done this, how have you formulated the new proposals?

Part 3: Evaluation of entertainment – general discussion

1. Did you have any comments about values in relation to entertainment or related aspects?
2. Do you think it is right to evaluate entertainment in this way on web sites?

Empirical usability evaluations – revised methodology

Interview questions and questionnaires

Questions for subjects after completing of the evaluation

Part 1: The web site

1. Are you members of other similar 'communities' or chat sites?
2. If so – how does this (Stadium) work compared to this/these?
3. What did you think of the content of the web site?
4. What was the best or worst thing on the web site?
5. Does the web site fulfill the expectations you had of Stadium Activity Town?
6. Iron Man
    a. What did you think about this?
    b. Do you usually play similar web based games?
    c. How would you consider the feeling in the game, or playability?

The test in itself

7. Do you usually surf with someone or are you mostly by yourself when surfing the web?
8. Was there anything that you considered unpleasant during the test?
9. Would you consider this session an authentic or normal use situation, or was it unrealistic because it was an evaluation?

Expert evaluations – further revised heuristics and methodology

Material delivered to experts

The experts received handouts to support their evaluations. The complete handouts can be found in CDM. An overview of the handouts is shown below:

- Introduction
- Web site 1
  - Description of web site 1 and specific instructions for the evaluation
  - Part 1: Exploration and entertainment
  - Part 2: Evaluation
    - Instructions
    - Evaluation form
  - Part 3: Meta-evaluation (evaluation of the evaluation itself)
    - Instructions
    - Evaluation form
- Web site 2
  - Description of web site 2 and specific instructions for the evaluation
  - Part 1: Exploration and entertainment
  - Part 2: Evaluation
    - Instructions
    - Evaluation form
  - Part 3: Meta-evaluation (evaluation of the evaluation itself)
    - Instructions
    - Evaluation form

After Part 1, the exploration part, the experts were asked to answer three brief questions about the web site. These questions were:

Post-exploration questions

1. Regarding the target group of the web site – how well do you think you to fit into this group (from 1-5, where 1 is 'not at all' and 5 is 'completely')
2. Regarding the web site as a whole, how much of it did you explore during your session (0-100%)
3. How long did your exploration last?

Interview questions and questionnaires

Instead of conducting oral interviews, as the experts had already spent many hours on the earlier work, an e-mail-based questionnaire was posted to them. They replied to the questions in written form. The questionnaire was as follows:

Questions for experts after completion of the evaluation

1 .Did you have any difficulties in accomplishing the task?

2. What did you think about the new evaluation methodology this time, compared to your other evaluations [in the other phases] regarding the changes? Please post comments on the basis of the methodological changes (stated below):

2.1. Opportunity to give both negative and positive feedback as well as opportunity to give more motivation for comments about the interface according to the different heuristics.

2.2. Some minor changes in the language of the heuristics in order to make them more understandable.

2.3. Adding of function-related heuristics (based to some extent on Nielsen's heuristics)

2.4. Information to experts about intended target group as well as purpose of site in order to allow better feedback on heuristic 'design for right target group' and 'coherence between chosen design and desirably mediated feeling or mood'.

2.5. The 'free surf' approach was generally seen as being more authentic, like a real use situation, which is important especially when it comes to entertainment. That is the reason this approach remains in this new methodology.

2.6. Some of the heuristics fit more or less well, depending on the type of site. Here the ranking from the meta-evaluation of the test can be one solution and this is added. The importance/applicability of the specific heuristic to the evaluated web sited is thus ranked from 1-5.

2.7. Chance to give an 'overall judgment or review' of the web site in one's own words.

3. Are there any more changes that you think should be added, based on earlier evaluations? Feel free to argue for your proposal.

4. Do you have any more comments the questions above do not cover?

Appendix II
Selected data material

A large amount of data material was obtained from all the different parts of the study. For pedagogical reasons, in the reporting of the study in the thesis some parts were excluded from the main chapters. The structure of this appendix is based on the referring parts in the thesis as follows:

- Part 2
    - Empirical usability evaluation
    - Inspection method evaluation –experts
    - Inspection method evaluation – novices
- Part 4
    - Empirical usability evaluation – revised methodology
    - Inspection method evaluation – revised methodology
    - Inspection method evaluation – further developed methodology

Part 2 – empirical usability evaluation

Overview of subjects in the three studies

The table below shows data for the tests in relation to different sites. ESC corresponds to *Eurovision Song Contest*, M to *Mosquito* and T to *Total Defence*.

	ESC	M	T
Contacted	e-mail	A:Personally/ C:through teacher	35:Through teacher
Positive/total	20/80	A:10/15 C:11/35	13 /35
Group(s)	Adults	C=Children, age 7-14	Children, age 7-14
Group(s)		A=Adults; age 20-30	

Table AII.1 An overview of selection and grouping of the subjects in tests

Table AII.2 An overview of the subjects in the tests

	ESC	M	T
Singles	14	17	1
Pairs	3 (6 subjects)	2 (4 subjects)	6 (12 subjects)
Total tests	17	9	7
Total subjects	20	21	13

Written pre-evaluation questionnaire – background information

1. Name (optional)
2. Age
3. Have you visited this web page before?
4. How would you consider yourself as a web surfer in general (Novice – Expert)
5. Do you surf the web often (Yes, every day – No, less than once per day)?
6. Do you use computers a lot in your work or at school (Work mainly with support of computer in my work/school – I never use computers at work/school)?
7. How would you rate your interest in ESC (None whatsoever – Very big)?
8. Are you primarily a MAC or PC user (MAC or PC)
9. What browser do you mainly use (Explorer, Netscape, Other (which?)
1. Have you participated in a user test before? If so, approximately how many?

Subjects – Eurovision Song Contest web site

Below, the subjects are further described in more detail, sorted on each part of the study. First, the part where the website *Eurovision Song Contest* was evaluated. The first table below gives an overview of sex, age, whether the subject used mainly a PC or a MAC and finally, as some of the subjects worked in pairs, the number of collaborator in pair sessions.

Subjects	Man/Woman	Age	PC/MAC	Grouped with
1	W	30-40	PC	-
2	M	20-30	MAC	-
3	M	30-40	PC	-
4	W	20-30	PC	-
5	M	40-50	PC	-
6	M	20-30	PC	-
7	M	30-40	MAC	-
8	W	20-30	PC	-
9	W	30-40	MAC	10
10	W	20-30	PC	9
11	M	30-40	PC	-
12	W	20-30	PC	13
13	M	20-30	MAC	12
14	W	50-60	MAC	-
15	M	30-40	PC	16
16	W	40-50	PC	15
17	W	50-60	PC	-
18	M	50-60	MAC	-
19	M	50-60	PC	-
20	W	20-30	PC	-

The complete questionnaire can be found in Appendix 1

Table AII.3 An overview of information of subjects about the evaluation of Eurovision Song Contest.

Subjects	Experienced as web surfer (1-5)	Frequency of surfing (1-5)	Use of comp in work (1-5)	Interest in 'schlager' (1-5)	Earlier visits to web site (number of times)	Earlier user tests (number of times)
1	4	1	5	4	NO	NO
2	3	4	4	2	NO	NO
3	5	5	5	2	YES (2)	YES (3)
4	4	5	5	2	NO	YES (1)
5	4	5	5	2	NO	NO
6	4	5	5	2	NO	NO
7	5	4	5	1	NO	YES (3)
8	3	5	5	3	NO	NO
9	4	4	4	2	NO	YES (4)
10	3	4	5	3	NO	YES (1)
11	3	4	5	2	NO	NO
12	5	5	5	3	NO	YES (3)
13	3	5	5	4	NO	NO
14	4	4	5	4	NO	NO
15	1	1	5	5	YES (2)	YES (1)
16	4	2	5	5	NO	NO
17	1	1	3	1	NO	NO
18	5	5	5	1	NO	YES (1)
19	4	5	5	2	NO	NO
20	4	3	5	3	NO	NO

Table AII.4 An overview of subjects' earlier experience in the evaluation of the Eurovision Song Contest web site.

The complete questionnaire can be found in Appendix 1

Subjects – Mosquito web site

The general information of the subjects in the part where the web site 'Mosquito' was evaluated is as follows. These tables refer to the group of adults:

Subjects	Age	PC/MAC	Browser (E, N, other)	Grouped with
1	20-30	PC	E	-
2	30-40	MAC	N	-
3	20-30	PC	E	-
4	20-31	PC	N/E	-
5	20-32	PC	E	-
6	20-33	PC	E	-
7	20-34	PC	N	-
8	20-35	PC	N	-
9	20-36	PC	N/E	-
10	20-37	PC	E	-

The complete questionnaire can be found in Appendix 1

Table AII.5 An overview of information about subjects in the 'adult' group in the evaluation of the Mosquito web site.

The experience of this group is as follows:

Subjects	Visited web site earlier	Experienced web surfer (1-5)	Frequency of surfing (1-5)	Use of comp in work (1-5)	Earlier tests (number of tests)
1	NO	5	5	4	NO
2	NO	4	4	4	YES (2)
3	NO	4	4	4	NO
4	NO	3	5	4	YES (1)
5	NO	2	5	4	YES (2)
6	NO	5	4	4	YES (4)
7	NO	4	3	3	YES (1)
8	NO	4	5	5	YES (2)
9	NO	5	5	5	YES (2)
10	NO	5	5	4	NO

The complete questionnaire can be found in Appendix 1

Table AII.6 An overview of information about subjects in adult' group in the evaluation of the Mosquito web site.

The second group of subjects used in evaluations of Mosquito web site were all children 9-10 years old. The general information about this group looks as follows:

Subjects	Age	Earlier visits	Seen on TV	Internet use (1-5)	PC/ MAC	Earlier tests	Grouped With subject
1	9	NO	YES	1	-	NO	
2	10	NO	YES	5	PC	NO	
3	10	NO	YES	1	-	NO	
4	10	NO	-	1	-	NO	
5	10	NO	-	2	-	NO	
6	9	NO	YES	2	PC	NO	
7	10	NO	YES	5	PC	NO	
8	9	NO	YES	1	PC	NO	9
9	10	NO	YES	2	PC	NO	8
10	9	NO	YES	1	-	NO	11
11	9	NO	YES	2	-	NO	10
The complete questionnaire can be found in Appendix 1							

Table AII.7 An overview of the subjects' earlier experience in the 'children' group in the evaluation of the Mosquito web site.

The earlier experience of the same group is shown below:

Subjects	Visited web site earlier	Experienced web surfer (1-5)	Frequency of surfing (1-5)	Use of comp in work (1-5)	Earlier tests (number of tests)
1	NO	5	5	4	NO
2	NO	4	4	4	YES (2)
3	NO	4	4	4	NO
4	NO	3	5	4	YES (1)
5	NO	2	5	4	YES (2)
6	NO	5	4	4	YES (4)
7	NO	4	3	3	YES (1)
8	NO	4	5	5	YES (2)
9	NO	5	5	5	YES (2)
10	NO	5	5	4	NO
The complete questionnaire can be found in Appendix 1					

Table AII.8 An overview of subjects' earlier experience in the 'children' group in the evaluation of the Eurovision Song Contest.

Subjects – Total Defence

In the study of the 'Total Defence' web site, the subjects were all children. The complete background information for this group is given below:

Subjects	Age	Earlier visits	Internet use	PC/MAC	Earlier tests	Grouped with
1	9	NO	1	PC	NO	2
2	9	NO	4	PC	NO	1
3	9	NO	5	PC	NO	4
4	9	NO	1	PC	NO	3
5	9	NO	5	PC	NO	6
6	9	NO	5	PC	NO	5
7	9	NO	3	PC	NO	8
8	9	NO	4	PC	NO	7
9	10	NO	4	PC	NO	10
10	9	NO	3	PC	NO	9
11	9	NO	1	-	NO	-
12	10	NO	1	MAC	NO	13
13	10	NO	1	-	NO	12
The complete questionnaire can be found in Appendix 1						

Table AII.9 An overview of general information and experince of subjects in the evaluation of the Total Defence web site.

Part 2 – Inspection method evaluation – experts

Experts

The profile of the chosen experts is important to be aware of, in order to be able to interpret the results. The profiles of the experts are further described below:

Table AII.10 Profile of the experts

Experts	Man/Woman	Age	Contact research area
1	M	30-40	2
2	W	20-30	2
3	M	20-30	2
4	W	20-30	1
5	M	30-40	1
6	W	20-30	1
7	M	40-50	2
8	M	20-30	3
9	M	20-30	2
10	M	30-40	1

The complete questionnaire can be found in Appendix 1

Experience of the experts

Table AII.11 Experience of the experts

Exp.	Evaluation in general	Heuristic Evaluation	Walkthrough Evaluation	Evaluating www	Use/visit Entertain. web sites	Evaluat Entertain. Web sites
1	3	4	3	2	4	2
2	3	4	4	3	4	3
3	4	4	3	4	5	4
4	3	3	3	2	4	2
5	3	4	2	3	2	1
6	3	3	3	2	2	2
7	3	3	2	3	2	2
8	3	4	4	3	3	3
9	4	5	4	5	4	4
10	1	2	2	2	2	1

The complete questionnaire can be found in Appendix 1

Part 2 – Inspection method evaluation – novices

Team	Man/Woman	Age	Contact research area
1	W/W	30-40	Education
2	M/M	20-30	Education
3	M/M	40-50	Education/Work
4	W/W	20-30	Education
5	M/M	20-30	Education
6	M/M	20-30	Education
7	W/M	20-30	Education
8	M/M	20-30	Education
9	M/M	20-30	Education
10	W/W	20-30	Education

The complete questionnaire can be found in Appendix 1

Table AII.12 Profiles of the teams

Experience of the teams

Team	Evaluation in general (1-4)	Heuristic Evaluation (1-6)	Walkthrough Evaluation (1-6)	Evaluating www (1-5)	Use/visit Entertain. web sites (1-5)	Evaluat Entertain. Web sites (1-5)
1	2	4	4	2	3	1
2	1	2	2	2	4	2
3	2	4	4	4	3	2
4	2	4	4	3	3	2
5	2	2	4	2	4	1
6	2	4	3	3	4	2
7	2	3	6	2	2	1
8	2	4	4	2	4	2
9	2	4	4	2	4	2
10	2	4	2	2	3	2

The complete questionnaire can be found in Appendix 1

Table AII.13 Experience of the teams

Table AII.14 Overview of the additional web sites evaluated
The complete questionnaire can be found in Appendix 1

Team	Name of web site	Web link to web site	Brief description of web site
1	Chili	http://www.chili.se	A webzine for young people with articles, chat support, postcards, games etc.
2	Arthistory	http://www.arthistory.se	Web site about art history with many multimedia features
3	Silikon	http://wwww.silikon.nu	Support web site for TV show 'Silikon' and webzine mainly for young women
4	Temptation Island	http://www.tv5.se/temptationisland	Support web site for TV show 'Temptation Island'
5	The Simpsons	http://thesimpsons.com	Global site – the Simpsons cartoon
6	Plannja Basket	http://www.plannjabasket.com	Web site of basket ball team 'Plannja'
7	Global fun	http://www.globalfun.com	Entertainment portal with a community part, games etc.
8	Guggenheim	http://www.guggenheim.org	Virtual museum
9	Skrattnet	http://www.skrattnet.com	Information retrieval-based site about jokes
10	Disney	http://www.disney.com	Global site of Disney Corp.

Team	Chosen method for evaluation	Type of method (EEM)/ (IM)	Level of quality of performance and data (very poor – excellent, 1-5)
1	Questionnaires	EEM	3
2	Cognitive walkthrough	IM	4
3	Scenario based evalution	IM	1
4	Questionnaires	EEM	3
5	Focus group evaluation	IM	5
6	Cognitive Walkthrough	IM	4
7	Feature inspection	IM	4
8	Heuristic evaluation using Schneiderman's design guidelines	IM	3
9	Pluralistic walkthrough	IM	5
10	Inspection on the basis of Donald Norman's 'seven steps'	IM	1

Table AII.15 Overview of the additional methods the novice expert evaluations

Part 4 – Empirical usability evaluation – revised methodology

Session	Num. of subjects	Man/ Whoman	Age	Experience of web use	Freq. of web surfing	Freq. of use of comp. at school	Earlier experience of evaluations
1	2	M/W	15	5	5	5	No
2	1	W	14	3	1	3	No
3	1	M	14	5	5	3	No
4	2	M/M	13/ 14	5	5	3	No
5	1	W	14	3	2	2	No
6	1	M	13	3	4	2	No
7	2	M/W	15/ 15	5	4	3	No

Table AII.16 An overview of the subjects in this phase of the study and their earlier experience

Procedure

The subjects worked both in pairs and alone. An overview of the evaluation session is presented below:

	Number
Singles	4
Pairs	3 (6 subjects)
Total tests	7
Total subjects	10

Table AII.17 An overview of the subjects in the tests

Part 4 – Inspection method evaluation – revised methodology

Fit to target group

The experts were asked to judge what type of the target group the website had, and then estimate how well they fitted into this group. The results are shown below:

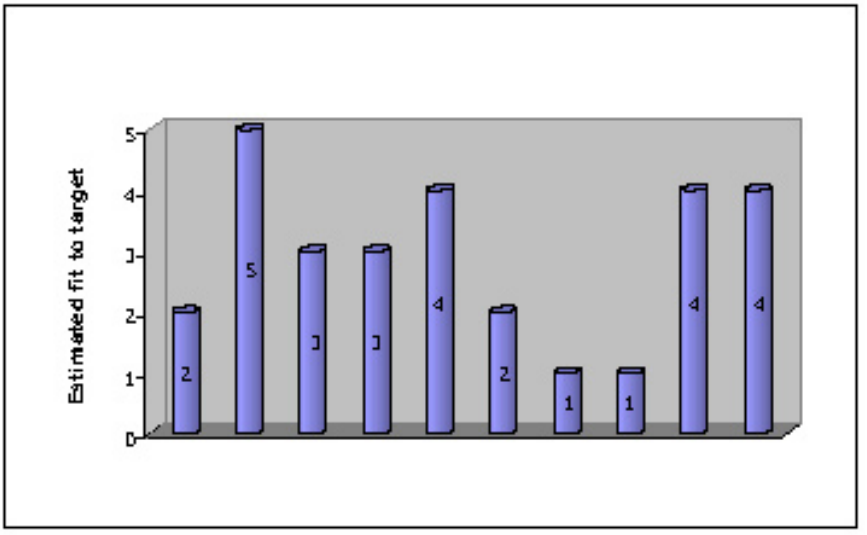

Figure AII.1 Estimated fit to target group of experts when evaluating the Mosquito web site

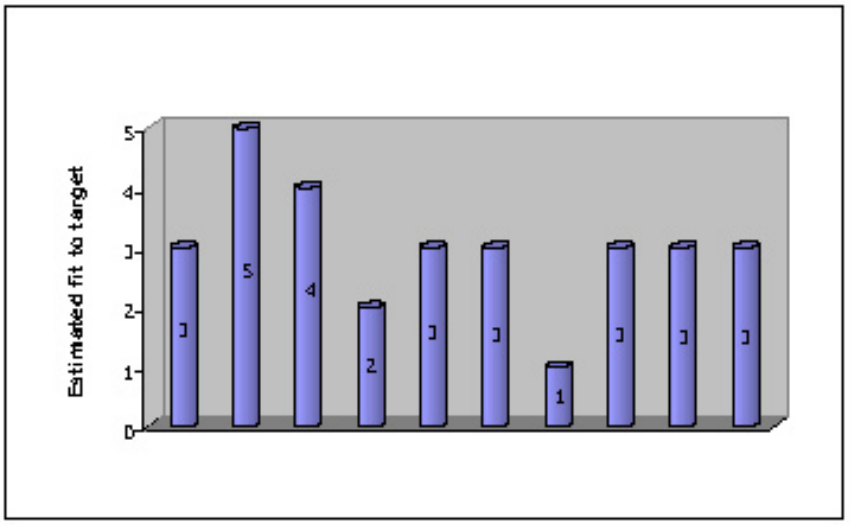

Figure AII.2 Estimated fit to target group of experts when evaluating the Skyscraper web site

Part of website viewed

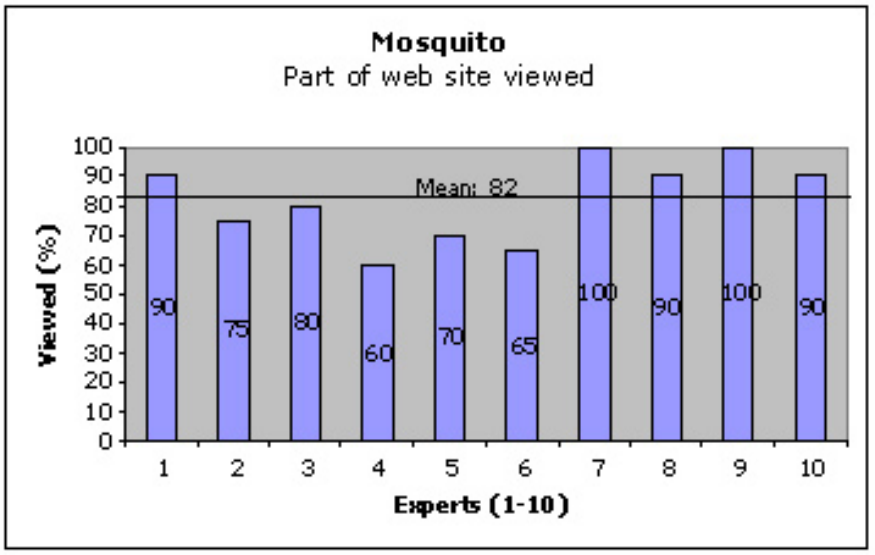

Figure AII.3 Estimated part of the web site viewed by experts when evaluating the Mosquito web site

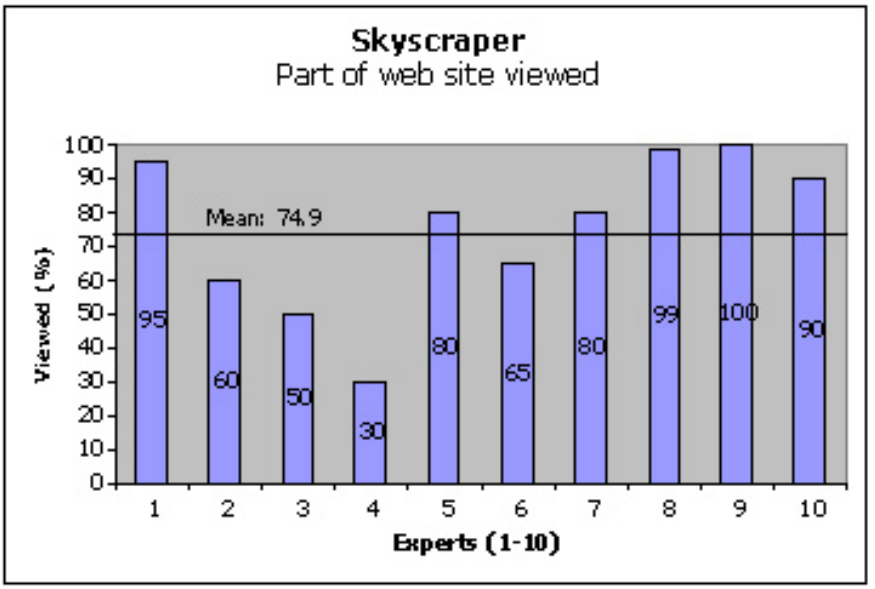

Figure AII.4 Estimated part of the web site viewed by experts when evaluating the Skyscraper web site

Time spent on website – free-surf session

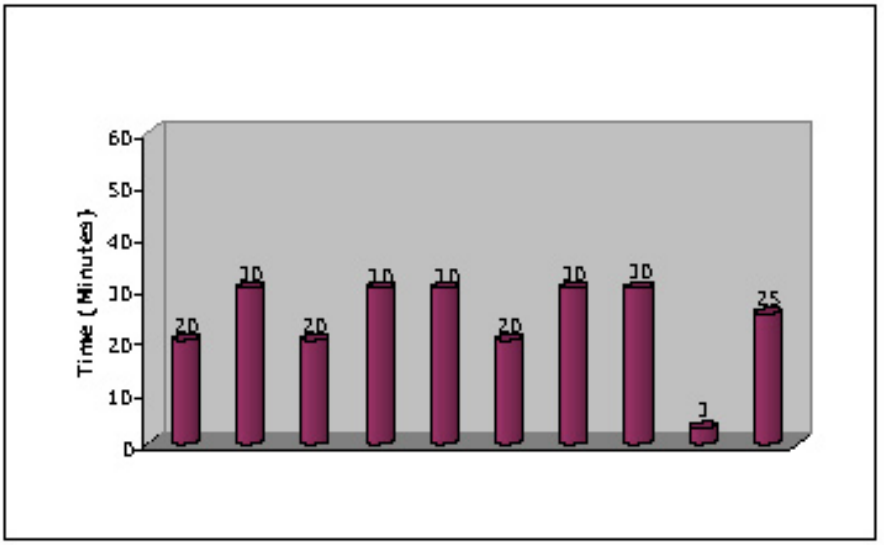

Figure AII.5 Estimated time spent on the web site viewed by experts when evaluating the Mosquito web site

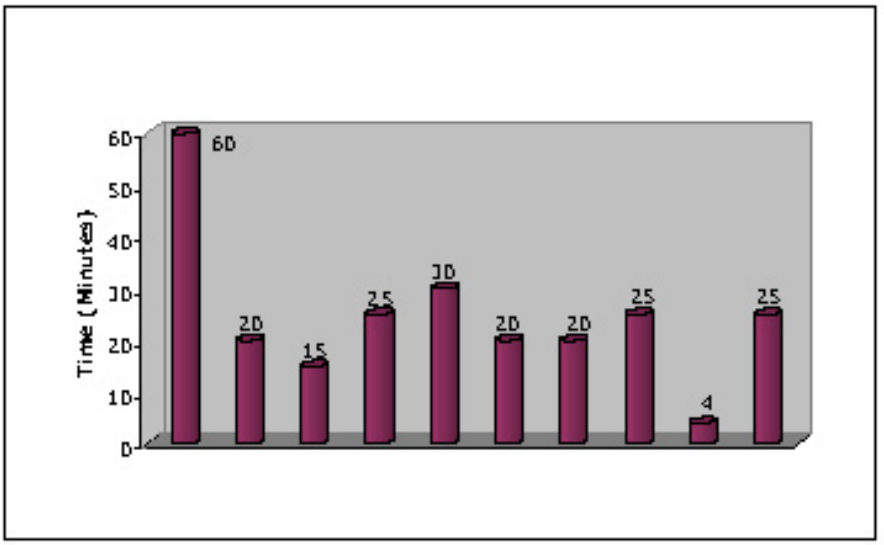

Figure AII.6 Estimated time spent on the web site viewed by experts when evaluating the Skyscraper web site

Part 4 – Inspection method evaluation – further developed methodology

First, the comments from the 'Vodafone' website are given:

Expert	Total number	Positive	Positive (% of total)	Negative	Negative (% of total)	Overall judgment
1	11	8	73	3	27	4
2	10	7	70	3	30	4
3	12	6	50	6	50	3
4	9	1	11	8	89	3
5	3	2	67	1	33	4
6	4	2	50	2	50	3
7	8	1	13	7	88	2
8	9	6	67	3	33	4
9	9	2	22	7	78	3
10	7	2	29	5	71	3

Table AII.18 Overview of comments by experts when evaluating the Vodafone web site

Below, the comments from the 'Stadium – Activity Town' are given:

Table AII.19 Overview of comments by experts when evaluating the Stadium web site

Expert	Total number	Positive	Positive (% of total)	Negative	Negative (% of total)	Overall judgment
1	-	-	-	-	-	-
2	15	6	40	9	60	3
3	19	14	74	5	26	4
4	10	5	50	5	50	3
5	14	6	43	8	57	3
6	8	1	13	7	88	2
7	11	5	45	6	55	4
8	11	1	9	10	91	4
9	6	0	0	6	100	3
10	5	0	0	5	100	4

First, the overview of positive comments for the 'Vodafone' website:

Table AII.20 Overview of positive comments for the 'Vodafone' website

Expert	Total amount positive	Heuristic 1-8	Heuristic 1-8 (% of total)	Heuristic 9-10	Heuristic 9-10 (% of total)	Overall Judgment
1	8	7	88 %	1	12 %	4
2	7	4	57 %	3	43 %	4
3	6	4	67 %	2	33 %	3
4	1	1	100 %	0	0 %	3
5	2	1	50 %	1	50 %	4
6	2	2	100 %	0	0 %	3
7	1	1	100 %	0	0 %	2
8	6	6	100 %	0	0 %	4
9	2	0	0 %	2	100 %	3
10	2	1	50 %	1	50 %	3

Below, an overview of positive comments for the 'Stadium' website:

Expert	Total amount positive	Heuristic 1-8	Heuristic 1-8 (% of total)	Heuristic 9-10	Heuristic 9-10 (% of total)	Overall Judgment
1	-	-	-	-	-	-
2	6	6	100 %	0	0 %	3
3	14	12	86 %	2	14 %	4
4	5	5	100 %	0	0 %	3
5	6	5	83 %	1	17 %	3
6	1	1	100 %	0	0 %	2
7	5	5	100 %	0	0 %	4
8	1	1	100 %	0	0 %	4
9	0	0	-	0	-	3
10	0	0	-	0	-	4

Table AII.21 Overview of positive comments for the 'Stadium' website

The same overview was made for the negative comments. Below, an overview of negative comments is given, first, for 'Vodafone':

Expert	Total amount negative	Heuristic 1-8	Heuristic 1-8 (% of total)	Heuristic 9-10	Heuristic 9-10 (% of total)	Overall Judgment
1	3	2	67 %	1	33 %	4
2	3	2	67 %	1	33 %	4
3	6	4	67 %	2	33 %	3
4	8	7	88 %	1	12 %	3
5	1	1	100 %	0	0 %	4
6	2	2	100 %	0	0 %	3
7	7	6	86 %	1	14 %	2
8	3	1	33 %	2	67 %	4
9	7	3	43 %	4	57 %	3
10	5	5	100 %	0	0 %	3

Table AII.22 Overview of negative comments for the 'Vodafone' website

Further, an overview of negative comments by the experts on the 'Stadium' website:

Expert	Total number negative	Heuristic 1-8	Heuristic 1-8 (% of total)	Heuristic 9-10	Heuristic 9-10 (% of total)	Overall Judgment
1	-	-	-	-	-	-
2	9	3	33 %	6	67 %	3
3	5	3	60 %	2	40 %	4
4	5	2	40 %	3	60 %	3
5	8	4	50 %	4	50 %	3
6	7	1	14 %	6	86 %	2
7	6	4	67 %	2	33 %	4
8	10	6	60 %	4	40 %	4
9	6	5	83 %	1	17 %	3
10	5	4	80 %	1	20 %	4

Table AII.23 Overview of negative comments for the 'Stadium'website

Results from meta-evaluation

Below, the results from the meta-evaluations are given:

STADIUM	Heur.									
Expert	1	2	3	4	5	6	7	8	9	10
1	5	5	1	1	1	5	1	4	1	2
2	5	5	5	5	4	4	4	3	4	5
3	5	5	5	5	5	5	5	4	5	5
4	5	4	4	5	5	5	4	4	5	5
5	5	3	3	4	4	5	4	5	4	4
6	5	5	5	4	4	4	4	1	5	5
7	4	4	4	4	4	3	4	3	3	3
8	5	5	1	5	3	4	3	5	5	5
9	5	5	5	5	2	3	1	5	3	5
10	5	5	4	5	5	5	3	5	5	5
Mean	4,9	4,6	3,7	4,3	3,7	4,3	3,3	3,9	4,0	4,4

Table AII.24 Overview of the results from the meta-evaluation of heuristics when evaluating the 'Stadium' web site.

VODAFONE	Heur.									
Expert	1	2	3	4	5	6	7	8	9	10
1	5	5	1	1	1	5	1	4	1	2
2	5	5	4	5	5	2	2	4	4	4
3	5	5	3	5	5	5	3	4	5	5
4	5	5	4	5	4	4	5	5	5	3
5	5	3	2	1	5	3	2	4	4	5
6	4	4	4	5	5	5	4	4	4	4
7	4	4	3	4	2	3	3	1	3	3
8	5	5	1	5	3	4	3	5	5	5
9	5	5	1	5	4	5	1	3	5	5
10	5	5	3	5	5	5	3	5	5	5
Mean	4,8	4,6	2,6	4,1	3,9	4,1	2,7	3,9	4,1	4,1

Table AII.25 Overview of the results from the meta-evaluation of heuristics when evaluating the 'Vodafone' web site.

Appendix III

The web sites evaluated in the study

The web sites used in the study are described briefly below to give the reader an overview of the numerous sites involved. The web sites are both so-called information retrieval web sites and more entertainment related web sites. The former type focus mainly on sharing information in an efficient way, while the latter have a variety of entertainment purposes. One information retrieval web sites is used mainly to have a control web site in order to show what results could be related to entertainment web sites and what could be related to evaluation of web sites in general.

The entertainment web sites are:

- Eurovision Song Contest
- Mosquito
- Totalförsvaret (Total Defence)
- Skyscraper
- 'How are you?' – Vodafone
- Activity Town – Stadium
- Jernkontoret – Captain Steel

The information retrieval web site is:

- SJ (Swedish Railways)

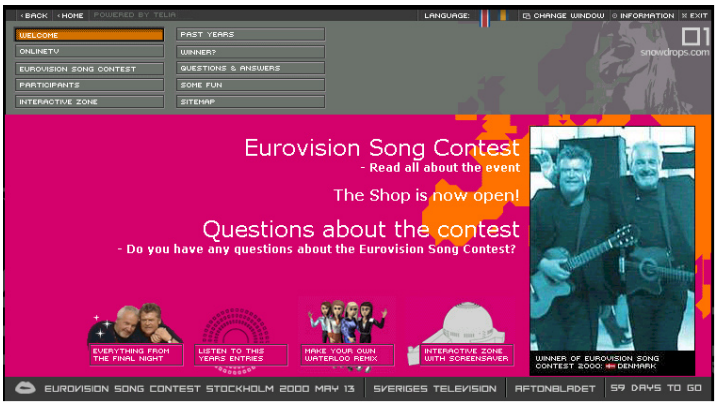

Figure AIII.1 The ESC home page

Figure AIII.2 The 'Waterloo remix' page

Figure AIII.3 The Mosquito home page.

Figure AIII.4 The Hong Kong Yoyo page.

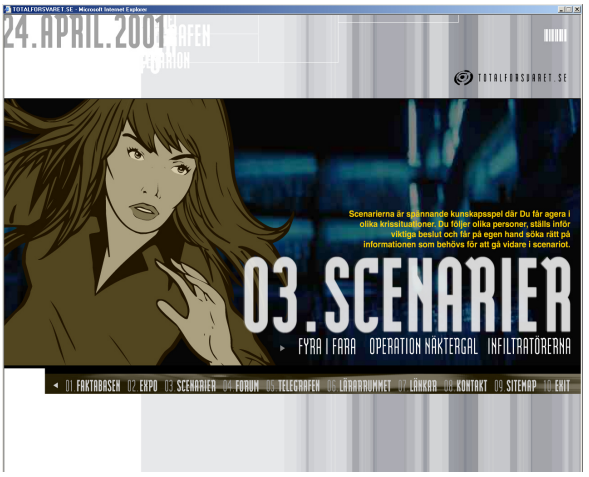

Figure AIII.5 The Totalförsvaret Web Site:

Figure AIII.6 The scenario 'Fyra i fara' (Four in The Scenario Section danger).

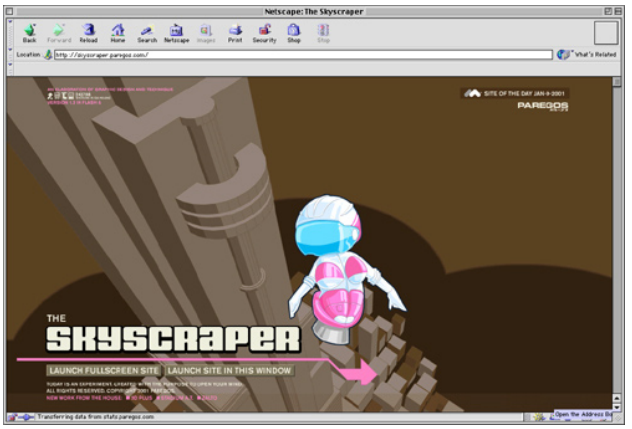

Figure AIII.7 The Skyscraper web site at the Paregos web site.

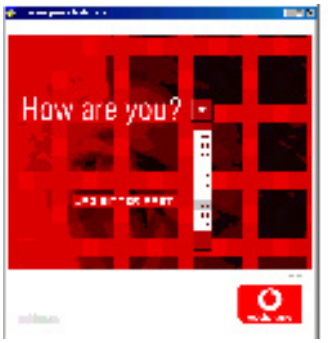

Figure AIII.8 Screenshot of the 'How are you?' campaign site from Vodafone. (http://howareyou.vodafone.com/)

Figure AIII.9 A screenshot from 'Activity Town' on the Stadium web site.

Figure AIII.10 screenshot from 'Activity Town' on the Stadium web site – the game ‚Bad Guys Monkeys

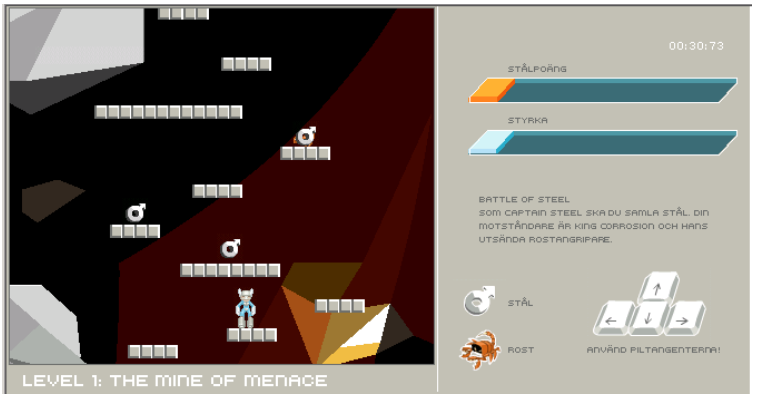

Figure AIII.11 A screenshot from the game 'Captain Steel' on the Jernkontoret site

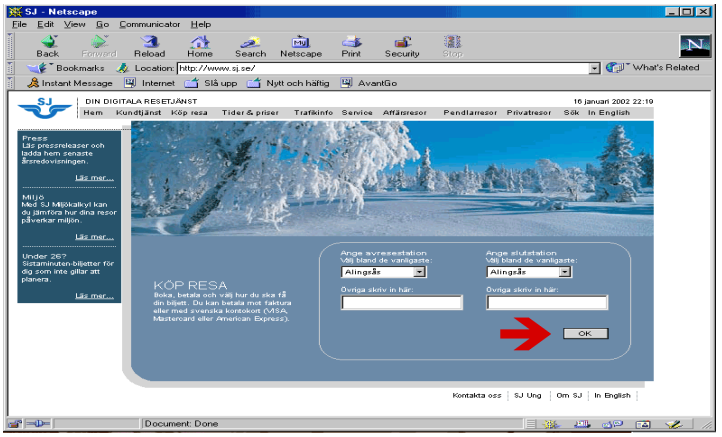

Figure AIII.12 A screenshot from the homepage of the SJ (Swedish Railways) web site .