Proceedings of The International Conference on Affective Human Factors Design
Asean Academic Press, London, 2001

From *Ease of Use* to *Fun of Use* : Usability Evaluation Guidelines for Entertainment Web sites

Charlotte Wiberg

Center for digital commerce, Department of informatics, Umeå University, 901 87 Umeå, Sweden

Abstract
Observations collected during evaluation studies conducted on three web sites are reported in this paper. The sites have different target groups and represent different types of entertainment web sites. Every web site was tested with a group of approximately 20 subjects. The intention was to test these groups in different ways according to different usability evaluation techniques. The study focused on comparison of the following conditions: (1) Subjects working individually vs. in pairs, (2) Levels of structure in sessions, traditional tasks vs. 'free surf' vs. combined task analysis and 'free surf' (3) Testing children vs. testing adults. (4) Written vs. oral answers to questions concerning entertainment. The whole set was tested on the three sites in a number of ways. This to gain knowledge in order to give implications for testing different types of sites. The result show upon how different conditions in tests could be set in order to give fruitful guidance in finding usability issues connected to entertainment.

Keywords Usability, evaluation test design, entertainment, pleasure, www

1. Introduction

Evaluation of entertainment web sites, specifically designed to be affective to the user, challenges traditional evaluation frameworks in the area of *Usability Engineering* (Olsson, 2000a, 2000b). Traditionally, *Usability Engineering* has so far focused on *ease of use* of interfaces. When it comes to the World Wide Web, and especially entertainment sites, the use situation is more about experiences, changes of emotional states, and levels of engagements than about information retrieval, error rates, and the ability of the user to (as quickly as possible) perform a task. Recent studies of Internet use or "surfing" show that when people are enjoying themselves, time flies and the focus is on the present activity (Agarwal & Karahana, 2000). How long time it takes for a particular user to find information on a page, as have been an important measure up to now, is an issue of less importance. According to this background the new measure of success is *fun of use*. This opens up the need for new usability evaluation approaches. The problem however, is that even though it is clear that *fun of use* needs to be considered in more detail, little or no attention has been paid to new methods of entertainment web sites evaluation. The question is if we can use traditional methods at all, can they be combined and elaborated in order to give us guidance, or do we need to develop totally new frameworks of methods to evaluate *fun of use*.

The rest of this paper is structured as follows; The paper starts with a brief summary of earlier related work within the field of *Usability Engineering*, *Human-Computer Interaction (HCI) in general* and *affective computing*. The method used in the experiment is described on a more general level as well as more specifically. The results of the tests are described according to the different conditions in the test and finally the paper concludes by summing up the findings into implications for entertainment web site evaluation.

2. Related work

Related research has been done in different areas. Some of this is summarized below. The related work in the mentioned areas is chosen from related research material, such as conference proceedings, books, leading journals and web sources.

Usability engineering is a field, in which practitioners and researchers have worked 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, as *think aloud, heuristic evaluation* and *task analysis*(c.f. Nielsen 1993). The scope of artifacts and interfaces tested are wide but recently, usability engineering built the ground for *web usability evaluation*. The same or similar techniques as traditionally have been used. However, this new evaluation context requires new approaches (Borges et al, 1996, 1998; Bevan, 1998; Nielsen, 1999; Spool, 1999; Olsson, 2000a, 2000b; Kaasgard, 2000).

In the field *affective computing* (Picard, 1997; Höök et al, 2000), research focuses upon issues related to interfaces and systems that imitate human behavior, as robots, interface agents and assistants and more. Questions like how they should be built as well as how people's reactions are, are of concern. Some studies of interest for us have been conducted where designed affective systems were evaluated on users. Here interesting discussions arouse.

The experience economy is discusses according to the fact that more and more companies make money on experiences instead of on the product itself, or more, the experience is the product. Examples are theme parks as Disney World and Universal Studios, restaurants like Planet Hollywood and Hard Rock Café and web sites like www.whatisthematrix.com, a support site for the film Matrix.(Pine II and Gilmore 1999) The discussions in this field are fruitful and valuable for us. However, they are a bit too focused on more general phenomena like trends and tendencies in the experience economy. This type of literature does not actually inform us how to evaluate these experiences on a concrete level. However, some of the frameworks presented here are of interest and could be found useful as frameworks of guidance of different kinds.

Recently, discussions about the need for new types of measures have been raised. A more holistic view, compared to the cognitive and physical view of products, is emerging and different types of human-product relationships are explored. Jordan (2000) builds an explaining framework and talks about four different types of pleasures; Physio-Pleasure, Socio-Pleasure, Psycho-Pleasure and Ideo-Pleasure. These are of more general kind then only applicable on the web. However, they may well be used as a support in design of tests and analyze of data.

Overall, in the above mentioned related work, there is a lack of focus on methodological considerations regarding evaluations in general as well as entertainment web site more specifically. This is what this article aims to discuss.

3. Method

The tests were designed in order to give guidance in how to combine different evaluation techniques when testing entertainment web sites. The sites were chosen according to different parameters. The sites should be (1) up and running, (2) have different target groups, (3) provide a unique type of entertainment compared to the other sites in the test, and (4) won a prize or got other type of recognition. We selected the sites according to the criteria.

3.1 The sites

Below, the three sites are described according to overall purpose of the site and target group. Also, a brief overview of the design is given.

Eurovision Song Contest

This was an event site for a TV-event with the same name. In 2000 the contest was held in Stockholm, Sweden, and Paregos AB built the web site The target group of this site were people interested in schlager music in general and more specifically the Eurovision Song Contest. The purpose of the event site was:

> *"Swedish Television and Aftonbladet 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)*

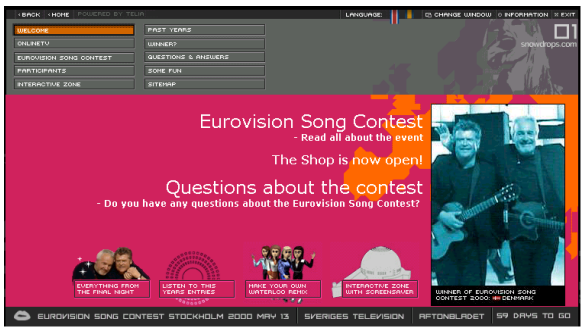

Figure 1: The ESC home page

Figure 2: The 'Waterloo remix' page

Mosquito

The Mosquito web site was created by Paregos AB, as a support site of the Swedish TV-show with the same name. The target group for this site were children 7-15 years old and people interested in design and technology. The audience is probably found in the target group of the TV-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 with several prizes in the media business, chosen the 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)

Figure 3: The Mosquito home page.

Figure 4: The Hong Kong yoyo page.

Totalförsvaret (Total Defence)

This site was released in 1999 and has got a lot of attention as well as won prizes for 'best information site' in Sweden. The main purpose of the site stated on the web site of Paregos Media design AB:

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 politics. Everything concerning total defence, for instance 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)

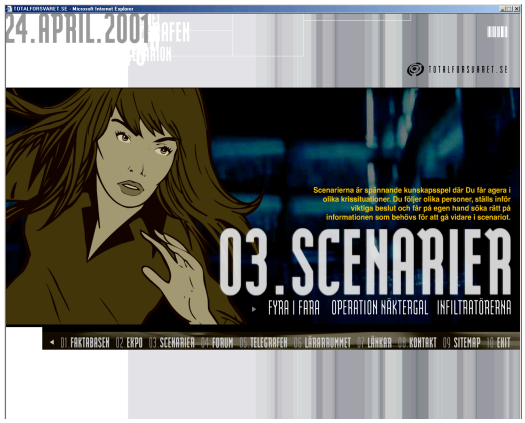

Figure 5: The Totalförsvaret Web Site: The Scenario Section

Figure 6: The scenario 'Fyra i fara' (four in danger).

As the web site of *Total defence* is very big and the fact that the target group is children, a specific part of the site was chosen to work as test bed for the tests. The chosen part was one out of three scenarios, called *Four in Danger*. Below, a brief description of the two above screen shots is given.

The Scenario page (see figure 5) contains three different interactive "knowledge games" that the visitor can play. In the scenarios, the visitor is introduced to various situations that requires some degree of problem solving. Most of the problems can be solved using the knowledge base. In the scenario 'Fyra i fara' *four in danger* (see figure 6) the visitor helps four persons in various problematic situations caused by power failures. The situations are presented as short, animated movies that stops in specific states where the visitor must participate and help the persons. In order to do so, the visitor answers a number of questions. Each question has three alternative answers, and the correct answer can be found in the knowledge base. In some questions, a link to the knowledge base is offered as an explicit help.

3.2 Subjects

Selection criteria for the subjects participating in the study were different for different target groups.

- ***Eurovision Song Contest:*** Here, e-mail were sent to students and staff at department of Informatics at a Swedish university and some randomly chosen acquaintances of the test team that matched the target group described above. Approximately 80 recipients received the e-mail and 20 of them replied with a positive respond. All of those were tested. No one in the group that answered that they wanted to be tested got rejected.
- ***Mosquito:*** Two types of subjects were used; 1) 'Adults' were recruited to be between 20 and 30 years and the tested group were mostly students. 2) Children were recruited to be from 7 to 14 years old, which was the target group of the site. This group was found at a local school. The main group of these children was 9 – 10 years old
- ***Total Defence:*** This site had a similar target group as Mosquito, and the same school was used to find children. However, different subjects were used in the two tests. The age of the children was 9 – 10 years.

Table 1 shows data of the tests according to different sites. ESC corresponds with *Eurovision Song Contest*, M with *Mosquito* and T with *Total Defence*.

Table 1: An overview of the tests

	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, age 7-14	Children, age 7-14
Group(s)		A=Adults; age 20-30	

3.3 The test setting

- Most of the tests consisted of three phases; 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, where questions about the site and the test were asked.
- In all of the sessions, two experimenters were present, one of them was coordinator, who ran the test and made all the decisions according to test strategy. The other experimenter present took notes.
- A digital video camera was used to record all the tests. The focus for the camera was an 'overall view' where we tried to catch both the screen as well as the face on the subject. To have a clear understanding of what happened on the computer screen, the experimenter used follow-up questions of the actions the subject did, like *'and now you clicked ... (yes) and where in the site did you end up then.. '*.
- The tests had been conducted in a variety of settings, such as computer labs, meeting rooms and more. The criteria for the setting were a desk with a computer, a digital video camera on a 3-pod, pointing towards subject and computer screen from beside. Further, slightly behind the subject, the coordinator of the test had his place on a chair and finally the second experimenter were sitting somewhere near, not to aside, as we got comments that it was unpleasant to have someone 'over in the corner, just staring'.
- One issue in user testing is the level of intervention from the experimenters. On one hand such an intervention could be seen as a big disturbance when intervening. However, the situation could be awkward if experimenter is totally quiet, even when obvious occasions occur when support or feedback would be normal, both pedagogically and social. Some difficulties that arise are not within the scope of the test and there the experimenter should be available to help subjects deal with such problems. The level of intervention was not decided in advance. The decision was situated and the level of intervention was elaborated upon due to circumstances. The person that did not test for the moment took notes about outcomes concerning level of intervention.
- Each subject received a ticket to the cinema as thank for his/her time and effort.

3.4 Design of the study
The study was designed to provide a comparison of the following conditions described below:

Pairs vs. Singles
The intention was to see whether the number of simultaneous subjects tested was of significance. How should the tested persons react when being tested with a friend? Would the think-aloud protocol work better with pairs, or is web entertainment better experienced and tested alone?

Table x: 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

Structured vs. unstructured user tasks
Task analysis is seen as central when doing 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, time required, mistakes committed, subject's comments and more. What happen when we evaluate entertainment? Is task analysis a proper way? How does a more free technique work where the subject is let loose? Approximately one third of all tests were conducted with structured user tasks where users got assignments to solve. One third used a mixed approach with both structured and unstructured tasks, where the latter could be described as 'free surf'. For the rest of the tests we used unstructured user tasks, and 'free surf' were used through the whole test.

Testing children vs. adults
When the designers of the 3sites told experimenters of the span of age of subjects in the target groups they got somewhat reluctant. This because 2 out of 3 sites had children from 7 to15 years old as main target group. To test children is different from testing adults and also. The experimenters were not clear, however, exactly what would be different. However, a decision was made to stick to the target group for the most part, but also do a split test on Mosquito, with both adults and children. This was in order to get a feeling for the differences in the results between adults and children. Instead of becoming 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 a common technique for testing subjective satisfaction for subjects is interviews (c.f. Nielsen, 1993). What is the most successful approach when asking about entertainment and experiences? Should subjects be asked in writing or should the questions be included into the oral interview? In the test we used both approaches. At ESC, interviews were mostly oral. At *Mosquito*, the same questions were given to the subjects in writing. At *Total Defence* most of the interviews were oral, as all the subjects were children.

4. Results

Overall, the tests produced numerous results relevant to both evaluation and design of entertainment web sites. However, below the focus is upon results concerning evaluation, beginning with some overall findings and continuing with more specific findings, sorted by the different parameters elaborated.

Testing the *fun of use* compared to the *ease of use* may differ a great deal. However, the ease of use must not be forgotten in some circumstances. For instance, subjects do not find it amusing when there is a chance they are being fooled. In one of the sites one of the buttons in the navigation bar was marked 'Fun Stuff'. Subjects interpreted it as the area where games, e-postcards and other fun stuff can be found on this type of site. Instead, subjects got a commercial film from sponsors with a high level of sound. Almost all of the subjects got surprised and/or annoyed with this part. None of the subjects showed interest in the content of commercials after being fooled.

The level of intervention varied. If using a scale from 1 to 5, where 1 is no intervention, on the average the tests rated around 3. The varying of the level of intervention followed two patterns; (1) Situated level, and (2) in advance set level. Here, the first situated approach was much more reliable in order to get useful data. Setting the level in advance only disturbed the test in an unnatural way. Having fun on one's one could be embarrassing when two silent people watch. It they smile along and ask follow-up questions it is more natural. This was at

least clear for the single tested subjects. Pairs required less intervention and more *traditional Usability Engineering* type of guidelines for the level of intervention could be practiced in these cases. One example of this is to give firm guidance around difficulties not tested, like hardware problems, for instance.

4.1 Pairs vs. singles

Entertainment fits well to be tested in pairs, as entertainment and exploration are activities well suited for groups. This aspect is also connected with the level of intervention from the experimenter. If tested in pair, subjects have less need for feedback on their experiences from the experimenter. It fits well with the idea that if you share the experience with someone else it gets even better. However, tendencies like 'show off', competition, domination etc.e could also be observed, especially during the tests of the children.

4.2 Structured vs. unstructured user tasks

As said earlier, three approaches were used; structured user tasks, unstructured user tasks and a mixed version. This due to the fact that many researchers pointed out the difficulties in giving subjects well defined assignments when testing web sites in general and especially this type of sites. Our findings show that depending on type of entertainment, an approach with structured user tasks might not be such a bad idea after all. For instance in order to test features where subjects build and send postcards, or where they mix their own song, these activities were assigned to the subjects. For some subjects, the experience and entertainment took over and they got somewhat immersed and therefor forgot the test situation. Here, the test was successful. However, in some of the tests the subjects performed the task as quickly as possible as if time measuring was being used. In the last type of tests structured user tasks did not support at all in testing sites.

These situations also occurred when the mixed approach, with both structured and unstructured user tasks were used. When suddenly new circumstances appeared like coming from given tasks to get recommendations to perform free surf the subjects were confused. The opposite order worked better, but the truth was that for some sections, structured user tasks was a poor choice of evaluation technique. Once subjects were put to perform structured tasks, they worked hard to fight the clock instead of spending time as they probably would, if not tested.

In some circumstances, like passages of games or riddles, subjects often wanted to do some exploration and not get it right at once. Traditional usability here would just spoil the fun. Here, the levels of support and on built-in keys were critical. Too much or too little help would spoil the entertainment. The examples of this were numerous.

However, overall through all of the sites, navigation was not something subjects found amusing when not found useful. Even if people wanted to explore, experience and be entertained on all these sites, this did not include navigation. This should stay *easy to use* and could very well be tested as such, with traditional *Usability Engineering* approaches, such as, for instance, *Task analysis*.

4.3 Testing children vs. adults

The experimenters testing children was really enthusiastic after the tests. Children have a totally different pattern overall in usability testing. Dealing with entertainment, some aspects were of specific significance. The children seem to 'play around' with, and explore interfaces more frequently compared to the adults we tested. This highlights an advantage, as children need fewer incitements to explore. Also, it was more common that children became immersed into the activity at hand compared to adults, who got less 'carried away'. In some ways, children could be more spontaneous, as when a pop-up message states they are wrong - without giving any correction at all and they, as talking to themselves, said things like *'strange… how am I to learn then'*. This spontaneity was more rare with adults, which were more 'well behaved' culturally.

However, adults could verbalize their actions more easily compared to some children who got completely quiet during the test. Also, when testing adults in the single tests the tests were overall more successful in the way that it gave results. This was not always the case with single children. Some of the children got afraid during tests, and one child had to interrupt the test to get the teacher to come and sit on the test as the child got frightened.

Emerging results regarding design were found, especially when testing the children, i.e. the true target group of the sites. Examples were level of language, complexity of tasks and some basic assumptions regarding knowledge in free text search made some tasks and scenarios almost impossible for children to handle the site.

Finally a comment concerning testing children and that is regarding the context of the usability tests. Soon the experimenters realized that the children answered to their questions with the fact in mind that if they responded negatively on questions they would not be able to get more exiting computer experiments to the school ever again. Some children also had comments like *'oh, I missed once again – do I not get a ticket to the cinema now…'*. These aspects must also be taken into consideration when testing children.

4.4 Written or oral answers on questions regarding entertainment.
As mentioned, the tests contained different types of questions, on which the subjects answered by writing or orally. The subjects answered in writing to the pre-test questionnaires and to some of post-test questionnaires, and orally during the think aloud test and also during the post-test interview. When answering questions regarding background information as well as question about errors, mistakes, effectiveness and other more common *Usability engineering* related questions, the subjects had no problems. The results were similar to other studies, and guidelines for those questions could be found in literature upon that subject (c.f. Nielsen, 1993).
However, to ask question regarding experience, fun, entertaining and such, the situation got more complex. One question was specifically used in order to get the subjects to, in a way, free their minds. The question was:

> '*If you should describe the site as a person (alternatively as a car) and you should describe this person to a friend you meet. How would you describe this person?'*

This question went splendid orally. The subjects first got quite silent, but after have thought for a while the most elaborating descriptions appeared. Here, the experimenter had a chance to follow up the answers with questions regarding what came up, as the answers were very subjective. The character of the site really got revealed here. However, this question, as well as other questions around entertainment, fun, experience and such, got very quick and short answers, most quite difficult to relate to. The experimenters just did not understand what the subjects had meant. When testing entertainment, questions and answers should be kept oral if possible, in order to be able to make follow-up questions.

5. Conclusions
Are we able to use old and traditional test methods from, for instance, *Usability Engineering* when testing entertainment web sites in the future? Concerning the implications of our study for web site evaluations, the findings can be summarized as follows:

- *Pairs vs. singles* – Testing entertainment with pairs works well, especially when testing kids. Look up, however for, domination, 'showing-off' and competitions in the pairs.
- *Structured vs. unstructured activities* - Depending on type of entertainment, a structured activity with assignments is not so bad after all. One might think it would be a disaster to use it in these environments. It is not. However, in highly explorative sections – use unstructured activities and let the subject surf more freely.
- *Testing children vs. adults* – Children are more spontaneous and explorative as subjects. If lucky, you get data of high quality. If children is target group – test children basically and also some adults to get comparison. However, be aware of the fact that adults might be better in thinking abstract and verbalize more frequently.
- *Written vs. oral answers on questions regarding entertainment* – For all traditional usability engineering based questions in the tests – use traditional guidelines. When asking questions regarding entertainment oral answers are to prefer, due to subjectivity in answers and the ability to do follow-up questioning

Finally, in order to get fruitful results when testing entertainment our tests showed upon the importance of being situated and intuitive as an experimenter. As a subject, to laugh in a silent crowd is difficult!

6. Future work
This paper describes and reports the first out of three iterations in a project, *Joyride*. The project has the purpose to find implications for entertainment web site evaluation. Within the project, three iterations of 60 usability tests on entertainment sites will be conducted, with totally 180 tests done. The different phases have different purposes: The first phase focuses upon how traditional usability evaluation methods can be used and combined and how to design conditions evaluated. This, in order to give guidance concerning usability aspects regarding entertainment. The second phase will continue based on insights from the first phase and also come up with new approaches for evaluation. The last phase will try to bring the above insights about evaluation into the design process and ongoing designs will be evaluated.

7. Acknowledgements
Thanks are due to Viktor Kaptelinin and Mikael Wiberg for valuable comments on earlier versions on this paper. I would also like to thank Kalle Jegers, Klaas Tojkander and Johan Tufberg for their endless enthusiasm for the usability evaluations. This work was conducted within the project *Joyride*, a part of the research group *Entertainment Services* in *Center for Digital Business*, Umeå University. This research was partly funded by European Union's regional development fund.

References

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

Bevan, N. (1998). Usability Issues in Web Site Design. In Proceedings of UPA'98. Washington DC.

Borges, J. A., Morales, I., Rodrigues, N. J. (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., Morales, I., Rodrigues, N. J. (1998). Page Design Guidelines Developed Through Usability Testing. In Forsythe, C., Grose, E., Ratner, J., Human Factors and Web Development. Lawrence Erlbaum Associates, Publishers. Mahwah, NJ, USA.

Höök, Kristina, Persson, Per, and Sjölinder, Marie (2000) Evaluating Users' Experience of a Character-Enhanced Information Space, In Journal of AI communications, pp. 195-212, Vol. 13, No. 3, 2000.

Jordan, P. W. (2000). *Designing Pleasurable Products.* An Introduction to the New Human Factors. London, Taylor & Francis.

Kaasgaard, K. (2000). *Software Design & Usability.* Copenhagen, Copenhagen Buisness School Press.

Nielsen, J. (1993). *Usability Engineering.* San Francisco, CA, Morgan Kaufmann Publishers, Inc

Nielsen, J. (1999). User Interface Directions for the web. In Communications of the ACM. Vol.42, No.1.

Olsson, C. (2000a). 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. (2000b). 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.

Picard, R. W. (1997). *Affective Coputing.* Cambridge, MA, The MIT Press.

Pine II, B. J. and J. H. Gilmore (1999). *The Experience Economy*. Boston, MA, Harvard Business School Press.

Rizzo, P. (2000). Why Should Agents Be Emotional for Entertaining Users? *Affective Interactions. Towards a New Generation of Computer Interfaces.* A. Paiva. Berlin, Spriger Verlag**:** 166-181.

Spool, J., Scanlon, T., Schroeder, W., Snyder, C., DeAngelo, T. (1999) *Web Site Usability: A Designer's Guide.* Morgan Kauffman Publishers Inc.

.