Measuring User Experience of websites: Think aloud protocols and an emotion word prompt list

Helen Petrie
Human-Computer Interaction
Research Group
Department of Computer Science
University of York
Heslington, York
North Yorkshire UK YO10 5DD
Helen.Petrie@cs.york.ac.uk

John Precious
Human-Computer Interaction
Research Group
Department of Computer Science
University of York
Heslington, York,
North Yorkshire UK YO10 5DD
John.Precious@cs.york.ac.uk

CHI 2010, April 10–15, 2010, Atlanta, Georgia, USA.
ACM 978-1-60558-930-5/10/04.

Abstract
To develop simple yet effective methods for eliciting user experience of websites and other interactive technologies, we explored the use of two techniques: an emotional think aloud protocol and an emotion word prompt list (EWPL). A study of four websites with 16 participants found that a retrospective emotional think aloud protocol produced significantly more emotion words than an equivalent concurrent protocol; plus, with on average 40 emotion words per website, it appears an effective technique for eliciting users emotional reactions to websites. Surprisingly, the use of the EWPL did not produce more emotion words per website, but may still help users overcome their difficulties in expressing emotional reactions to websites when unprompted. Further research will explore the use of these methods with other interactive technologies.

Keywords
User experience, emotion, evaluation, website.

ACM Classification Keywords
H5.2User interfaces: evaluation/methodology.

General Terms

Experimentation, Human Factors, Measurement.

Introduction

Providing good user experience (UX) is rapidly becoming a key part of the development of interactive technologies, including websites. However, the scope of what constitutes UX is still a matter of debate. The new ISO standard [7] on human-centred design proposes UX as an all-encompassing concept, being "all aspects of the user's experience when interacting with the product … it includes all aspects of usability and desirability of a product … from the user's perspective". Of course, this then requires us to define and more importantly decide how to measure "desirability". Hassenzahl [6], on the other hand, proposes a definition of UX as a much more specific concept, being the "momentary, primarily evaluative feeling (good-bad) while interacting with a product or service". Hassenzahl appears to be trying to define UX as that aspect of a user's reaction to a product or service that is the emotional rather than the cognitive reaction (*feeling* as opposed to *thinking*). However, restricting this to only a good-bad reaction seems very narrow. Users can also have more subtle emotional reactions such as amusement, disappointment and frustration with technologies such as websites. These emotions can of course be categorized as good or bad (or perhaps positive or negative), but there is considerably more richness here that is potentially of interest to the designers of the technologies. As a result, our view is that UX can be considered to be all of the emotional components of a user's experience of interaction with a technology.

As part of an exploration of how to measure UX in simple but effective ways, we have adapted the think aloud technique [1, 2, 3] in order to elicit emotional reactions to interactive technologies rather than usability problems and users' cognitive understanding of the technology. In the *emotional think aloud*, users are asked to work through a series of tasks with the interactive technology (in this instance, a website) and talk about how they are feeling. As differences have been found between retrospective and concurrent protocols [4] we have employed this technique both as a concurrent and a retrospective think aloud protocol.

Initial research [5, 8] showed that some people find the emotional think aloud difficult to do (perhaps more so than doing a traditional think aloud). Users often report that it is difficult to articulate their feelings about a website. It is not clear whether this is because many websites do not elicit emotional reactions from users, or whether it is difficult for them to talk about their emotions per se, or in relation to a technology (for example, they may think it is inappropriate to react emotionally to a technology). In order to overcome this problem we have explored the usefulness of providing users with an Emotion Word Prompt List (EWPL), a list of emotion words commonly used in describing websites (see Table 1, Column 1), piloted in a previous study [8].

In order to investigate the effectiveness of the emotional think aloud technique, in both concurrent and retrospective forms, and the effectiveness of the EWPL in eliciting UX, we conducted a study with four websites. The websites were evaluated with both concurrent and retrospective emotional think aloud protocols, with and without the support of the EWPL.

After each task with a website, participants were also asked to complete a rating scale measure of their emotional reaction to the website, which consisted of the words from the EWPL presented as 7-point Likert items. This provided another measure of emotional reaction to the website to compare with the results from the think aloud protocols.

Methods

16 participants took part in the study, 5 women and 11 men. Their ages ranged from 24 to 66 years with a mean age of 36 years. On average they have used computers for 15 years and rated their expertise as "competent" (on a 5-point Likert item from 1 = novice to 5 = expert). On average they use the Internet for 15 hours each week. None of the participants regularly use any of the websites used in the study.

The design was a non-factorial one, with each participant evaluated two of the four websites, undertaking two typical tasks per website. On one website participants undertook a concurrent emotional think aloud protocol. Instead of talking about the usability problems they encountered and their understanding of the website, they were asked to express how they felt about the website whilst performing the tasks. On the other website, participants undertook a retrospective emotional think aloud protocol, with similar instructions.

For one website, participants were shown the EWPL, a list of 16 emotion words, 9 positive valence, 6 negative valence and one ambiguous (surprised) before they started the tasks (Table 1, Columns 1 and 2). This list was developed in a previous study (Petrie and Harrison, 2009) from words people produced in a think aloud exercise with websites. Participants were asked to read through the EWPL and told that, whilst they may find the words useful, they were also free to choose and use any other words to express their emotional reactions during the think aloud. The EWPL was placed so they could see it during the use of that website.

Table 1: Emotion Word Prompt List (EWPL)

Emotion Word	Valence	% total incidence
Amused	Positive	-
Annoyed	Negative	12.0
Bored	Negative	3.0
Confident	Positive	4.4
Confused	Negative	7.2
Creative	Positive	-
Curious	Positive	-
Disappointed	Negative	5.4
Frustrated	Negative	12.0
Happy	Positive	6.4
Interested	Positive	3.3
Hopeful	Positive	8.6
Pleased	Positive	2.5
Relieved	Positive	-
Surprised	Ambiguous	-
Unsure	Negative	3.8

After each task participants were asked to rate intensity of feelings during that task using the 16 EWPL words, presented as 7-point Likert items (1= low, 7= high).

The four websites used in the evaluation were:

- www.britishmusem.org: website of the British Museum in London
- www.visitbritain.com: a website to help people plan holidays in the UK
- www.british-towns.net: a website with information about how to find places in the UK
- www.uk-piano.org: website of the UK Association of Blind Piano Tuners

Results

The first analysis looked at the total incidence of emotion words per website for the different conditions (total incidence means all words including repeats of the word, for example if the participants says "I'm confused" three times during interaction with a website, that counts as three words). Overall, participants produced 29.7 emotion words per website. An analysis of variance showed a significant difference between the total incidence of emotion words in the concurrent and retrospective think aloud protocols ($F_{1, 15} = 5.89$, $p < 0.05$), with approximately twice as many words being produced in the retrospective protocols (Mean Concurrent = 20.7; Mean Retrospective = 40.4). On the other hand, there was no significant difference in the number of emotion words produced when the EWPL was available compared to when it was not available ($F_{1, 15} = 0.62$, n.s.). There was also no significant interaction between the protocol condition and the availability of the EWPL ($F_{1,12} = 0.5$, n.s., and $F_{1, 12} = 2.13$, n.s.). This pattern of results was exactly the same if the analysis was conducted on the number of different emotion words produced per website and on the number of emotion words produced per 100 seconds of task time.

A second analysis looked at the particular emotion words that were produced and whether they were part the EWPL. For this analysis all forms of a word were counted (e.g. if the participant said "I'm confused" and "that's confusing", these both counted towards the "confused" count). Clearly there is a semantic difference here, the first statement is about the participant, the second is about the website. However it was felt that these were both statements about the emotion experienced by the participant using the website. In the total incidence of emotion words, 73.9% of words used were part of the EWPL. There were no significant differences in the percentage of words which were part of the EWPL between the two think aloud conditions ($F_{1, 15} = 0.06$, n.s.), between the availability of the EWPL during the task and not ($F_{1, 15} = 0.17$, n.s.) and no interaction between think aloud condition and availability of EWPL ($F_{1,12} = 0.1$, n.s., and $F_{1, 12} = 0.5$, n.s.).

A third analysis looked at the frequency of use of the different words from the EWPL. This was conducted across all websites. Column 3 of Table 1 shows the percentage of the total incidence of emotion words for each of the EWPL words. 11 of the 16 EWPL words had percentages of more that 2.5% (more than 23 occurrences). In total, these 11 words accounted for 68.6% of the total incidence of emotion words. An analysis was also made of emotion words not from the EWPL, but none had percentages as high as 2.5%. The most commonly used emotion word not on the EWPL was "satisfied" (or "dissatisfied") which together accounted for 1.64% of incidences.

The final analysis looked at the relationship between the number of positive and negative emotion words

produced during the think aloud protocols and the ratings of the website on the EWPL words after completing each task. On average, participants produced 11.5 positive emotion words and 18.2 negative emotion words per website. Table 2 shows the correlations between the numbers of positive and negative emotion words produced during the think aloud protocols and the rating of the website on the EWPL words which were presented as 7-point Likert items. For positive emotions, there was a significant correlation on only one of the four tasks (although there was a strong trend on a second task), but for negative emotions there were significant correlations on all four tasks.

Table 2: Correlations between numbers of positive/negative emotion words produced during think aloud protocols and ratings of EWPL words

	Positive emotion words	Negative emotion words
Website 1	r = 0.48	r = 0.55
Task 1	p = 0.06	p = 0.02
Website 1	r = 0.53	r = 0.49
Task 2	p = 0.03	p = 0.05
Website 2	r = -0.33	r = 0.49
Task 1	n.s.	p = 0.05
Website 2	r = 0.09	r = 0.56
Task 2	n.s.	p = 0.02

Discussion

The results showed that both think aloud protocols produced a considerable number of emotion words (30 per website, over the two tasks), but contrary to our expectations, the retrospective think aloud produced significantly more words. This may be in part that the retrospective think aloud does not place such a cognitive workload on participants, as they are not trying to do the task and talk at the same time. It may also be that the retrospective nature of the protocol allows them to reflect on their feelings more.

It was surprising to us that providing the EWPL did not help participants produce more emotion words, and that the availability of the EWPL had no significant effect on the number of words produced. One might argue that once the participants had seen the EWPL words to rate them after the first task, this made them aware of these words, but the pattern of significant and non significant effects showed even in the results of the first task undertaken, before the EWPL had been seen by participants who experienced the non EWPL condition first. Future research will explore this (lack of) effect further.

However, a high percentage of emotion words produced by participants were from the EWPL (nearly 75%), which shows that this set of words does capture a lot of the emotions about websites that participants wish to express. This is encouraging, given that the EWPL was originally developed with one set of websites and one set of participants [8] and now tested with a very different set of websites and participants. Although the EWPL did not increase the number of emotion words produced by participants, we feel it still has a role in evaluations of UX to indicate to participants the kinds of words they might use. However, this study suggests that the EWPL could be reduced to a list of 11 words (as indicated in Table 1, column 3).

The results also showed that there was a significant relationship, particularly for negative words, between the number of words used during the think aloud protocol (regardless of whether that was concurrent or retrospective) and the EWPL rating scales completed after the tasks. This latter measure was included in the study to provide some validation of the emotional think aloud protocol method, but may also provide a quicker way for practitioners to measure emotional reactions to websites and other technologies. Asking a participant to complete an 11-item rating scale measure after interacting with a website is a much quicker and easier way of measuring emotional reaction, both for the participant and for the evaluator. Further research on the use of the EWPL as a rating scale measure of UX is planned.

Conclusions

This study has shown that the *emotional think aloud* protocol is an effective method for eliciting participants' emotional reactions to websites, an important part of UX. The retrospective emotional think aloud protocol was found to be significantly more effective than the concurrent think aloud protocol. Surprisingly, the use of a prompting list of emotion words, the EWPL, did not increase the number of emotion words produced by participants in the emotional think aloud. Nonetheless we think the EWPL has a useful role in evaluations of UX in indicating to participants the kinds of words they might use in evaluating a website emotionally, as participants do indicate that they find this kind of task difficult. Our results also suggest that the EWPL can be effectively used as a rating scale measure to be completed after interaction with a website, which is a very efficient method of measuring emotional reaction to a website. Future research will investigate the use of the emotional think aloud protocol with other interactive technologies.

References

[1] Birns, J.H., Joffre, K.A., Leclerc, J.F. and Paulsen, C.A. (2002). Getting the whole picture: collecting usability data using two methods – concurrent think aloud and retrospective probing. *Proceedings of UPA Conference*, July 8 – 12, Orlando. Available at: www.usabilityair.org/publications/christinepaulsen/upa_thinkaloud_paper.pdf

[2] Dumas, J.S. and Redish, J.C. (1999). *A practical guide to usability testing*. Exeter, UK: Intellect.

[3] Ericsson, A. K. and Simon, H. A. (1993). Protocol Analysis - Rev'd Edition: Verbal Reports as Data. The MIT Press, revised edition

[4] Haak, M., de Jong, M., and Schellens, P. J. (2003). Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. Behaviour & Information Technology, 22(5):339–351.

[5] Harrison, C. (2009). *Exploring emotional web experience: more than just usability and good design*. PhD Thesis, University of York, UK.

[6] Hassenzahl, M. (2008). User experience (UX): towards an experiential perspective on product quality. In *Proceedings of IHM08 (20th French-speaking Conference on Human-Computer Interaction – Conference Francophone sur l'Interaction Homme-Machine)*.

[7] International Organization for Standardization. (ISO). *ISO/DIN CD 9241: Ergonomics of human-system interaction. Part 210: Human-centred design for interactive systems*. Geneva, Switzerland: ISO.

[8] Petrie, H. and Harrison, C. (2009). Measuring users' emotional reactions to websites. In *Proceedings of CHI 2009 (Extended Abstracts)*, April 4 – 9, 2009, Boston Massachusetts, USA. New York: ACM Press.