Measuring the User Experience of a Task Oriented Software

Jonheidur Isleifsdottir
deCODE genetics
Sturlugötu 8, 101 Reykjavik
jonheidur.isleifsdottir@decode.is

Marta Larusdottir
Reykjavik University
Kringlan 1, 103 Reykjavik
marta@ru.is

ABSTRACT

In this paper a study on a web based tool is described that is used to keep track of attendance and work schedules by employees and managers in large companies. Ten users particpated in a think aloud test measuring the usability of a new version of the software and the user experience was measured before and after each user test.

The user experience results show that the group of questions that measure the personal growth of the user got the lowest scores for this product, but pragmatic attributes, hedonic identification and attraction got much higher scores. It is not surprising that pragmatic issues get high scores for a task-oriented software like this one, but it is an interesting result that the users value highly the attraction and hedonic identification.

Author Keywords

User experience, usability, think-aloud, user testing.

ACM Classification Keywords

INTRODUCTION

In the past decade user experience has become a popular field of study within the field of Human Computer Interaction. It challenges the past notion that task related features are the only ones that contribute to usability. User experience focuses on the user entire experience when using a software product, not just the ISO 9241-11 factors, effectiveness, efficiency and satisfaction. User experience introduces new concepts to the quality of software like fun, beauty and pleasure.

The problem with the concept of user experience is how vague it is and that it can be interpreted in many ways. More empirical results are needed to define the concept clearly [6] and these will only be obtained by making models and using them to measure user experience for different products [3]. There are many unanswered questions that need to be addressed. What is beauty? Are beauty and usability related? What contributes to the goodness and beauty of products? What effects how the users summarize experiences during usability evaluations? [5] To address some of these problems Hassenzahl [2] has proposed a model of user experience that divides the attributes of a product into pragmatic and hedonic attributes. Based on this model he has made the AttrakDiff 2 [4] questionnaire that can be used to measure the user's experience of these different attributes in the product. That questionnaire has been translated to Icelandic by Marta Larusdottir.

The goal of this study is to measure the user experience of users participating in a typical think aloud test of a task-oriented software by using the Icelandic version of AttrakDiff 2 questionnaire. Our goal was to compare the measurements of the expectations to the tool and the user experience measured just after taking part in a think aloud test. Secondly we wanted to analyse the Icelandic translation of the questionnaire.

BACKGROUND

In this section it is explained what is meant by User Experience, a clarification of the different concepts that lie behind the AttrakDiff 2 questionnaire is given and some of the studies that have used it to measure what attributes to attractiveness of products are described.

User Experience

User Experience is a relatively new field within the larger scope of Human Computer Interaction. It proposes a more holistic view of the user's experience when using a product then is usually taken in the evaluation of usability [2]. Until now usability evaluations have primarily focused on task-related issues such as efficiency and effectiveness. Stating that a product that is efficient and effective in allowing the user to solve the tasks needed, to fulfill the user's goals, makes the user satisfied [6].

But is it enough to have a satisfied user? The researchers leaning toward UX say the answer is no [6]. The user needs to experience more that satisfaction with the product for it to be marketable. Hassenzahl [2] proposes a model for the different attributes a product can have and make up a product character. He states that a product has both pragmatic and hedonic attributes. The former being the task related attributes we are used to from the classic usability literature and the later emphasizing the users well being

VUUM2008, June 18, 2008, Reykjavik, Iceland.

while using the product. Hassenzahl also introduces in the same chapter three different classes of hedonic attributes; stimulation, identification and evocation. Later Hasssenzahl decided to drop the evocation class from the model and that is not included in AttrakDiff 2.

AttrakDiff 2

AttrakDiff 2 [4] is a questionnaire that measures hedonic stimulation and identity and pragmatic qualities of software products [1]. The questionnaire was originally made in German but has been translated to English. AttrakDiff 2 has four, seven anchor scales, in total 28 questions. These will be described in more details in the following.

Pragmatic Manipulation

Pragmatic attributes are the ones that are associated with how easy the user finds it to manipulate the environment, in this case the product or software. It is what makes us able to fulfill our goals and what we have until now talked about as usability. If we think pragmatically the only requirement from a product to squeeze juice from an orange is that it actually squeezes the juice from the orange and that we can find out how to use it on our own. There is no beauty or design needed to make a product pragmatic.

Hedonic Stimulation

The attributes connected to hedonic stimulation are the ones that encourage personal growth of the user. People want to develop their skills and knowledge further and these are the attributes of the product that allow for that to happen. As an example Hassenzahl provides unused features of a software [2]. Those features that the user does not yet use are not a part of the pragmatic experience but are rather perceived as hedonic as they provide stimulation for further development. Stimulation can also be provided by presenting things in a novel way or by a new interaction style.

Hedonic Identification

Attributes connected to hedonic identity are the ones that make us identify with the product in a social context. What message are we communicating to other socially by using this product? These attributes are connected to the fact that all persons communicate their identity through things they use and own. An example of this would be a personal website where you can communicate who you are to the outside world. If a product communicates what we think to be advantageous to others we might prefer that product. Ipods would be a good example of a product that communicates a strong identity. There are several other mp3 players on the market that work the same but the brand name is so strongly connected to the product and its coolness that everyone has to have an ipod.

Attraction

When we talk about something as being attractive to us, we are usually summarizing the whole experience of the product. We judge the product as a whole and use words like good, bad, beautiful and ugly to describe things. In AttrakDiff 2, attraction is used to measure the global appeal of a product and to see how the other attribute affect this global judgment [3].

Scale Examples

The Pragmatic Quality (PQ) scale has seven items each with bipolar anchors that measure the pragmatic qualities of the product. This includes anchors such as Technical-Human, Complicated-Simple, Confusing-Clear. The Hedonic Quality Identification (HQI) and Stimulation (HQS) scales also have seven anchors each. HQI has anchors like Isolating-Integrating, Gaudy-Classy, Cheap-Valuable. HQS has anchors like Typical-Original, Cautious Courageous and Easy-Challenging. AttrakDiff 2 also has a seven item anchor scale for overall appeal or attraction (ATT) with anchors like ugly-beautiful and bad-good. The anchors are presented on opposite sides of a seven point likert scale, ranging from -3 to 3, where zero represents the neutral value between the two anchors of the scale.

Related Work

Hassenzahl used the AttrakDiff 2 questionnaire to study four different mp3 player skins [3]. He used the questionnaire to explore the effect of pragmatic and hedonic qualities on beauty and goodness. The four different skins had been pre tested and judged ugly or beautiful; two skins were judged beautiful and two ugly. In the first study 33 students were asked to look at each skin and fill out an AttrakDiff 2 questionnaire for each of the skins without using them. Two 2x3 ANOVAs were performed on the result data, one for the beautiful skins and another for the ugly ones. The ANOVAs had skin and attribute group (PQ, HQS, HQI) as within subject factors and score as dependent variable. This revealed that there was one skin in each group that was significantly more stimulating than the other and the other was thought to be significantly more pragmatic. Other results were that the identity communication factors, HQI were the factors that had the highest correlation with the beauty rating, i.e. the HQI scores were significantly higher for the more beautiful skins.

A study on the influence of hedonic quality on attractiveness was done by Schrepp, Held and Laugwitz [7]. They sent out e-mails to students and asked them to look at three different interfaces of business management software. Around 90 people responded and the response rate was 34%. AttrakDiff 2 was used to measure the user experience after the user had looked at 11 different screenshots, with an explanatory text before it, of the interface executing a part of a business scenario. Schrepp et al. expected, since they were testing business software that are in their nature meant to support people in their work, that pragmatic qualities would have greater influence on attractiveness than hedonic qualities. What they found on the other hand was that both HQI and PQ contributed evenly to the attraction and HQS also contributed significantly.

They also found, as they expected, that more attractive interfaces were preferred over the less attractive ones.

MATERIALS AND METHODS

In this chapter the study using the newly translated AttrakDiff 2 questionnaire is described, first the tool which was evaluated, then the usability tests, the participants and the measurements in the study.

The tool - Workhour

Usability tests were conducted on a new version of software called *Workhour* (Vinnustund in Icelandic) [8]. It is designed by the Icelandic software company Skyrr, see figure 1. An old version had been in use for serveral years, but in the new version the user interface was changed exstensively. There are four main user groups of Workhour; ordinary users that work on shifts and those that work regular hours. The other two main user groups are managers that work on shifts and those that don´t.

The main tasks for ordinary users working on shifts is to check there monthly plan for shifts, ask for a day off and check if they have fullfilled all their work obligations for that month. The main tasks for regular users are asking for holidays and check if they have been too many hours off work. The *Workhour* system is very useful to managers, because they can do much of their organizing work in Workhour like check if all timestamps for their employees are correct, insert information about an employee that is sick and get an overview of how many have been sick over a particular period to name a few.

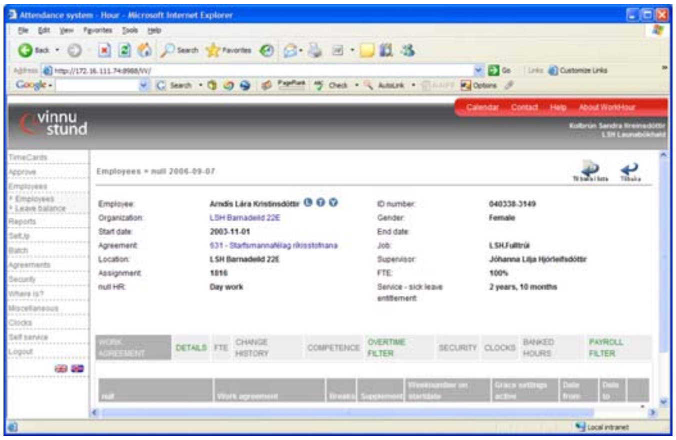
Participants

Ten individuals participated in the study, eight women and two men. Five of the participants were categorized as managers and the other five as ordinary users. The participants are employees of Landspitali – University Hospital and Financial Management Authority in Iceland, and were divided into two groups but all of them use *Workhour* as a part of their workday. One of the participants only filled out a very small portion of the pre-use questionnaire and none of the post-use so we did not use any data from him in the results.

The usability tests

The ten usability tests were conducted by two usability specialists on the new version of *Workhour* running on a test database two months before it was installed.

Each user solved six or seven tasks in think aloud tests which were adjusted to their ordinary tasks. The total number of tasks in the study was 17. The tasks were made by one of the developers of the user interface that has good connections to the users.



Employees » null 2006-09-07

Employee:	Arndís Lára Kristinsdóttir	ID number:	040338-3149
Organization:	LSH Barnadeild 22E	Gender:	Female
Start date:	2003-11-01	End date:	
Agreement:	631 - Starfsmannafélag ríkisstofnana	Job:	LSH.Fulltrúi
Location:	LSH Barnadeild 22E	Supervisor:	Jóhanna Lilja Hjörleifsdóttir
Assignment:	1816	FTE:	100%
null HR:	Day work	Service - sick leave entitlement:	2 years, 10 months

Figure 1. The new version of the software *Workhour*

The user tests were conducted at their ordinary working place, so a lot of contextual information was also gained. Two usability specialists conducted the tests; one was the organiser and one the data recorder. Additionally everything that was said was recorded on tape.

The AttrakDiff 2 questionnaire was administered before and after the think aloud test. First the participants were asked to answer the questionnaire according to their expectations to the new version of Workhour they would be trying in a minute. The users are all familiar with older versions of *Workhour* and were chosen to be in the study as typical users of the system. After the think aloud test was finished the user filled in the questionnaire again and now the participants were asked to base their answers on the experience of using Workhour to solve the given tasks. The reason for measuring both the expectations and the experience is that user experience a subjective factor that changes over time and we were interested to see what affect actually using the tool would have on the measurements of the user experience for the tool.

Unlike studies done by Hassenzahl on skins for mp3 players [3], our study had participants that are actual users of the software being evaluated. The tasks they performed were tasks they know and perform every day. The expectation of the user when answering the AttrakDiff 2 questionnaire before use was purely based on what the user expected of the new version, because most users had not seen the new version before answering the questionnaire.

Measurements

The AttrakDiff 2 questionnaire was translated from English to Icelandic by Marta Larusdottir and this was the first time it was used in the Icelandic format. There was one set of anchors that were left out when the questionnaire was translated. It was the HQI item with anchors Alienating-

Integrating. The translator was not able to find a suitable translation that differed from the translation of another pair of anchors. So the HQI only had six items.

The internal consistency of the HQI, HQS, PQ and ATT scores was measured both before use and after use and unfortunately it was not as high as previously measured by Hassenzahl [3]. Cronbach's α on the pooled values for the different scales before use was: PQ, α = .58; HQI, α = .57; HQS, α = .42; ATT, α = .43. These values for alpha are rather low and that indicates that there is not a high correlation between the answers with in each group. Usually the criteria of internal validity wanted from questionnaires is an α > .70 [3]. After use the scales internal validity was higher in three cases: PQ, α = .86; HQS, α = .55; ATT, α = .70. In the HQI, α = .46 scale the validity was lower than before. Both PQ and ATT were over the, α > .7 mark when measured after use, which is good but the internal validity of HQS is still too low.

RESULTS

In the following chapter the results will be described, first on the user experience and then the analysis of the translation of AttrakDiff 2.0 to Icelandic is described.

The user experience

We calculated the mean score of all user answers for each quality scale (each scale has 7 questions). As mentioned earlier each answer gets a value from -3 to 3, with zero as the neutral value between the anchors of the question.

As can be seen in Figure 2 all the quality scales have means above zero both before and after use of Workhour. The post-use line is also always beneath the pre-use line. A paired T-test that compared the pre and post-use scores for each participant and each category showed that difference in mean scores pre and post-use was significant in HQS(meanDiff = .67,t = 3.56, p = .007) and in ATT(meanDiff = .49,t = 3.43, p = .009) but not in HQI(meanDiff = .35,t = 1.62, p = .115) and PQ(meanDiff = .34,t = 1.78, p = .145). It is also noticeable that HQS score means are much lower than the other means both pre- and post use.

Since the internal validity of the scales was not very high we decided not to do any further analysis of the effects of the different qualities (HQI, HQS and PQ) on ATT as was done in other studies using AttrakDiff 2. One idea was also to test the correlation between different scales. At the AttrakDiff website there is a confidence square diagram made from the data which we did not attempt. There online experiments can be set up in German and in English [4].

Discussion

It is interesting to see that the post-use line (Experience) in Figure 2 is below the pre-use line (Expectations) and consistently so. Even though the difference is not statistically significant (This might be due to the small sample we had). We still believe it is relevant because it is found in all categories. We think that one reason for this is that users are optimistic that a new version is somehow better than the old and therefore have high expectations to its hedonic and pragmatic qualities, that are lowered when the product is used but not considerably. Participants might move their scoring one point closer to the middle.

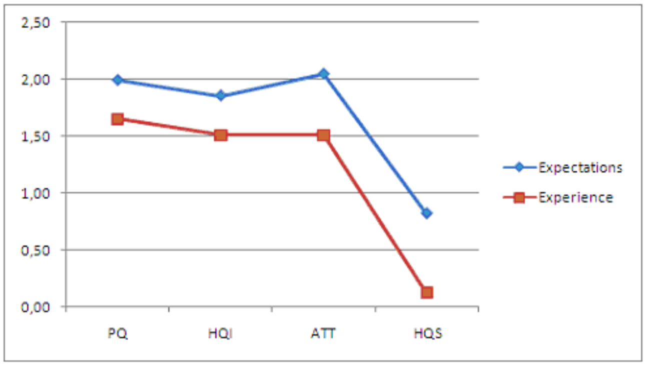


Figure 2. Mean scores for each scale of AttrakDiff 2

Moreover, even if the scores moved closer to zero they were over all pretty good compared with both the mp3 skin study and the business interface study mentioned earlier.

In Figure 2 we can see that HQS gives the lowest score. That category has a mean that is 1-1.5 lower that the other means.

This indicates that the software has less hedonic stimulation qualities than identification and pragmatic ones. This is an interesting result.

We think this is very understandable since we doubt that stimulation was one of the design goals of Workhour. It is software that is used to check on work schedules and to punch in and out of work. We think that the stimulation factor is more important when designing software for creative work rather than support software that most user use only for a short period of time each day and mainly just have to trust that it works.

Translation of AttrakDiff 2

As stated before the study above was also done to test how the Icelandic translation of the questionnaire was working. In the description of the variables and measurements above we showed that the internal validity of the scales was rather poor. Since other studies have shown much better validity scores our first thought was to look closer at how the scores were for each item on the scale. We started with the HQS scale because that gave the lowest internal validity in both the pre- and post-use study and also a much lower mean score than the others.

In Figure 3 we see that the third and sixth items of HQS give a very different mean than the others in this group.

This indicates that these items are not measuring effect that is similar to the others. The third HQS item has the word anchors bold - cautious where markings closer to bold give a higher score. The translation was *ótraust - traust* where *ótraust* gave a higher score. This is a somewhat misleading translation, that indicates that the software is not reliable and secure enough. A better translation would be *djarft-varfærnislegt*. The only problem is that *varfærnislegt* is already used in item four in HQS and we propose that *íhaldsamt* will be used there instead.

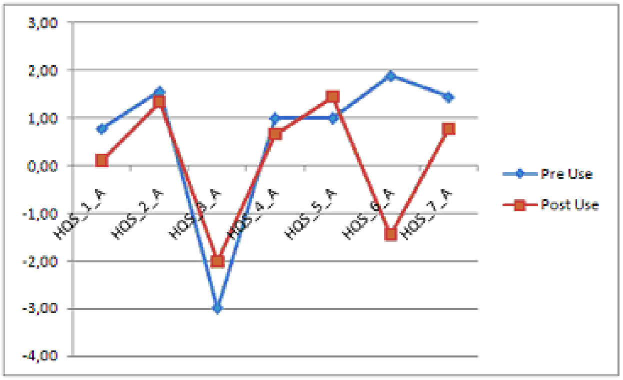


Figure 3. Scores of HQS: Pre- and post-use

The sixth HQS item has the word anchors undemanding - challenging where markings closer to challenging give a higher score. The translation was *auðvelt-krefjandi*. We don't think this translation can be improved and the reasons for the low score are most likely that users don't want software of this type to be challenging.

There was also one item missing from the translation. That was the fifth item in HQI that has the anchors alienating-integrating. The Icelandic translation for integrating was in use in the first HQI item.

Our suggestions are that we change item HQI_1 to *einangrandi-tengjandi* and item HQI_5 which was missing to *fráhverft-sameinandi*.

The other scales did not seem to be suffering from the same problems as the HQS scale. It is therefore our belief that if those changes are made the internal consistency of the scales would improve. If this does not happen it would raise the question whether the HQS scale simply does not apply in the same way as the other scales when measuring the user experience of very practical software. This is certainly a point worth studying further.

CONCLUSION

It was very interesting to see the scores for the different attributes of the AttrakDiff 2 Questionnaire on Workhour. It seems that hedonic stimulation is the least important factor in such software or at least what the participants thought deserved the lowest score.

It was surprising though how high the scores were in HQI considering that identification was probably not a part of the design. The good score in PQ, HQI and ATT is pleasant to see because that indicates that there is overall happiness with the software product.

It is dangerous to draw conclusions about the relationship of hedonic qualities and usability and goodness from the studies that have been done at present. A great deal of software is used every day for extended periods of time and not recreationally. It is questionable, if that kind of software would follow the same patterns as the mp3 player skins in Hassenzahl's study, and what about real users? Even though the people in his study probably use some kind of mp3 players frequently we do not know whether they are to be considered "real" users.

Further empirical studies are needed to be able to draw any conclusions about user experience and how it is affected. Hassenzahl's model is a step in the right direction and hopefully we will see a great increase in studies using that or similar models to evaluate user experience with software and other products. It is also our hope that translating the AttrakDiff 2 questionnaire to Icelandic will inspire more people to use it alone or as an addition to other user testing to gain more knowledge about what factors contribute to how attractive our software is to the user.

ACKNOWLEGDEMENTS

We want to thank the people at Skyrr for their very good co-operation in this project, the users, the developers and the managers.

REFERENCES

1. Hall, M., & Straub, K. (2005, October).
2. Ui design newsletter. Internet Resource http://www.humanfactors.com/downloads/oct05.asp. (Retrieved March 24, 2007)
3. Hassenzahl, M. (2003). The thing and I: Understanding the relationship between user and product. In M. A. Blyth, A. F. Monk, K. Overbeeke, & P. C. Wright (Eds.), Funology: From usability to enjoyment, 1-12 (chap. 3). Kluwer Academic Publishers.
4. Hassenzahl, M. (2004). The interplay of beauty, goodness, and usability in interactive products. Human-Computer Interaction, 19, 319-349.
5. Hassenzahl, M. (2007). AttrakDiff(tm). Internet Resource http://www.attrakdiff.de.
6. Hassenzahl, M., & Sandweg, N. (2004). From mental effort to perceived usability: transforming experiences into summary assessments. In Chi '04: Chi '04 extended abstracts on human factors in computing systems (p. 1283-1286). New York, NY, USA: ACM Press.

7. Hassenzahl, M., & Tractinsky, N. (2006, March-April). User experience - a research agenda. Behavior & Information Technology, 25, 91-97.

8. Schrepp, M., Held, T., & Laugwitz, B. (2006). The influence of hedonic quality on the attractiveness of user interfaces of business management software. Interacting with Computers 18 (5), 1055–1069.

9. Skyrr. (2007). Workhour. Internet Resource http://www.skyrr.is/vorur/vinnustund/.