Adapting User Testing for Evaluating a Multiplayer Game

Marta Kristin Larusdottir
Reykjavik University
Menntavegur 1, 101 Reykjavik
marta@ru.is
Tel: +354-599 6200

ABSTRACT
In this paper we describe how user testing was conducted in two cases; first how the evaluation of a work related software was conducted and secondly how an evaluation of a multi-user game was conducted. The goal of the study is to analyse what adaptations needed to be made when transferring a rather typical approach of conducting user testing in the first case to conducting user tests on two prototypes of the multiplayer online game in the second case. For the analysis we describe, which, who, what, why, when, where and how the evaluation was conducted.

This study shows that it is vital to describe all these different factors involved when describing a particular case of user testing to better understand the particular case and the effect that the context of the evaluation had on how the results can be interpreted.

Our analysis of the adaptations needed of the user testing approach shows that all the factors analysed needed to be changed in some way. The fact that this was a multi-user software still being developed had the biggest influence on the adaptations.

THE TWO CASES OF USER TESTING
Two cases of user testing are described in this paper. User testing of a web based software used for scheduling work hours at a hospital and in a public authority organization is described in the first case. A rather traditional user testing using the think-aloud procedure with real users was conducted in that evaluation. On the contrary an adapted user testing approach was used in the second case where two prototypes of a multiplayer online role playing game were evaluated with user surrogates.

In the paper we first describe each of these two cases by analysing those according to the 6Ws and the H approach [1]. Furthermore we describe the major differences between these two cases and how these affected the user testing in the second case.

Case 1 – The Work Related Software
In the following the first case will be described according to the 6Ws and the H approach [1].

Which – Traditional User Testing Approach
User tests were conducted with 10 users on a new version of a web based software called Workhour, which is used to schedule work hours and look up monthly plans for shifts for example. The users were asked to think aloud during the task solving session of the user tests. The evaluations were conducted in the users own work environment. During the evaluation, the participants filled-in two questionnaires, and got tasks to solve in the software while being observed by two researchers.

Who – Real Users
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.

Five regular users took part in the tests, four working on shifts in hospitals and one working on regular hours in a state institution. There were also 5 managers that took part, two working on shifts at hospitals and three working regular hours, two at a state institution and one at a software company. The users are all familiar with older versions of Workhour and were chosen to be in the study as typical users of the system.

What – The Work Related Software Workhour
User tests were conducted on a new version of software called Workhour. This version had been running on a test database for two months when the think-aloud evaluations were conducted.

Figure 1: The evaluated version of the software
An old version had been in use for several years, but in the new version the user interface was changed extensively to being more menu driven interface but still rather traditional design, see figure 1.

The main user goal for ordinary users working on shifts for using the software is to check their monthly plan for shifts, ask for a day off and check if they have fulfilled 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.

Why – For Researching Purposes - Summative
The reason for conducting the evaluation was to gather data for researching an approach for describing usability problems found in the user testing, described in [2] and to measure the user experience of the software, described in [3]. Furthermore, the results could be used for further design, but that was not the main purpose of the evaluation.

How – 6 Tasks and 3 Questionnaires
Ten usability tests were conducted by two usability specialists on the new version of Workhour. One of the usability specialists acted as a conductor and the other as a note taker and was responsible for the recordings of the sessions. In some sessions one of the developers of the system did attend as an observer.

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. The users took part one at a time and were observed by two researchers in their own environment. In some of the tests a developer of the system was also observing the users.

The participants were asked to fill in the AttrakDiff 2 questionnaire [4] before and after the task solving session. 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. After each think aloud user test was finished the user filled in the questionnaire again and now the participant was asked to base his/her answers on the experience of using Workhour to solve the given tasks.

Where & When – In the Users Own Work Environment
The user tests were conducted at their ordinary working place, so a lot of contextual information was also gained. The user tests were conducted two weeks before the delivery date, so the software was almost finished at that time. Still there was time to fix serious usability problems.

Case 2 – Multiplayer Online Game
In the following the second case will be described according to the 6Ws and the H approach [1].

Which – The Adapted Approach
Two prototypes for a massively multiplayer online role playing game (MMORPGs) made in a prototyping tool called CADIA-Populus (see http://populus.cadia.ru.is/) were evaluated in case 2. The prototypes were made to compare two possibilities of moving avatars around in the game and two possibilities of starting a conversation between the gamers using the avatars. A within subject user test was conducted with 14 participants in total to test the user experience while using these two prototypes.

Who - The Participants
The users of the game will be people who like playing computer games that include natural behaviour of avatars.

We organized sessions of user testing with two groups of participants using the prototypes simultaneously, the first with 6 participants and the second with 8 participants. We needed to ask many participants to participate at the same time because this is a multi-user game.

The participants in the user testing were actually user surrogates, because the game had not been launched. I the first user testing session there were 6 participants, 3 males and 3 females whilst in the second session there were 6 males and 2 females. Participants were between 20 and 32 years old, mainly students or researchers in at the university.

What - The Game Prototypes
One of the principle strands of research at CADIA at Reykjavik University is the agent simulation of the human ability to communicate verbally and nonverbally based on detailed models of human behaviour and cognitive processing. As a part of that principle two prototypes called System C and System S were made in the prototype tool CADIAPopulus for evaluating dynamic group behaviour in virtual conversation.

To be able to evaluate different issues two prototypes were made System S and System C. The differences between those can be resumed in 6 points:

- **Input devices**: what input devices are used when controlling an avatar in the prototype.
- **Navigation**: in which way the user can let the avatar move around and explore the environment.
- **Attention visual cues**: how the system provides cues about the avatar’s attention target.
- **Joining conversation**: in which way the user can join to a conversation.
- **Conversations**: how conversations are intended by the system.
- **Communication interface**: how the system presents communication features to users.
What - The System S Prototype

In System S the user controls his/her avatar forward and backward by the arrow keys. Additionally the user lets it rotate left and right using the respective arrow keys.

In order to start or join a conversation, the user had to approach another avatar, or group of avatars, and type something in the appropriated chat-box. All the avatars close enough to the user’s avatar received the message. Therefore, conversations were intended as a bunch of people close enough to each other. In System S there were not attention visual cues to let the attention’s target of an avatar stands out. Basically, the avatars in System S were agents with a very limited intelligence and not capable of being attentive or react to the environment, see figure 2.

Figure 2: The prototype named System S

What - The System C prototype

In System C user controlled the avatar by means of a point-and-click interface using the mouse, see figure 3.

Figure 3: The prototype called System C

Touching the screen’s border with the mouse pointer, the user could rotate the attention target and look around. In order to move the avatar, the user had to click on a desired destination point of the environment. Afterwards, the avatar would start to move in order to reach the chosen destination.

To start or join a conversation, the user clicked an avatar and typed something on the appropriate say-box. The message then appeared in a balloon above the avatar’s head. If a chosen avatar was too far away, the avatar would start to move in order to get close enough to its target. Basically, an avatar needed to get attentive to another avatar in order to say something to it. Therefore it needed to move towards the target to let it fall inside its field of attention, which was a triangular region in front of the avatar. The field of attention pointed toward the avatar’s attention target and changed colour and angular extension to visually describe different levels of attention focus. It is important to notice that an avatar was aware of the existence of another individual only if it did fall inside its field of attention and the avatar got attentive to it.

A Summary of the Differences Between the Prototypes

In SystemC avatars are smarter than in System S and are able to understand if they are in a conversation. Moreover, a chat-box will appear on the right side of the screen to let the user send and receive messages which belong exclusively to the conversation. The differences are summarised in table 1.

Table 1: Summary of the difference of the prototypes

<table>
<thead>
<tr>
<th></th>
<th>System S</th>
<th>System C</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Interface Control</strong></td>
<td>Arrow keys interface</td>
<td>Point-and-click interface</td>
</tr>
<tr>
<td><strong>Navigation</strong></td>
<td>Moving around the avatar</td>
<td>Setting a destination point</td>
</tr>
<tr>
<td></td>
<td>controlling its trajectory</td>
<td></td>
</tr>
<tr>
<td><strong>Attention visual cues</strong></td>
<td>No attention visual cues</td>
<td>Field of attention with different angular extensions, orientations and colors</td>
</tr>
<tr>
<td><strong>Joining Conversation</strong></td>
<td>Approaching another individual</td>
<td>Clicking on another individual and typing something</td>
</tr>
<tr>
<td><strong>Conversations</strong></td>
<td>Based purely on distance</td>
<td>Based on a model of motivational social forces</td>
</tr>
<tr>
<td><strong>Communication interface</strong></td>
<td>A common chat-box</td>
<td>Context-based interface and balloon system</td>
</tr>
</tbody>
</table>

Why – For Redesign Purposes - Formative

The reason for conducting the evaluation was to gather data for redesigning the CADIA populus prototype and to measure how the dynamic behaviour in virtual conversions affected the users. Additionally we wanted to measure the user experience of using each prototype and compare the
The goal was to evaluate the user experience of two issues:

1. **Avatar autonomous behaviour**: as an autonomous agent, each avatar is capable of reacting dynamically to other individuals. We used a model based on a potential field to drive the continuous action-reaction loop of an avatar participating in a conversation. We wanted to evaluate the validity of this approach in conceiving believability to users;

2. **Context-based interface**: our interface is able to understand in what social interaction the avatars are engaged and, consequently, to visualize parts of the interface without the user’s intervention. We wanted to evaluate the comfort and the efficiency of this approach.

**How – 6 Tasks and 3 Questionnaires**

The participants filled in three questionnaires, solved three tasks while using each prototype and took part in a debriefing session. The participants were asked to use the prototype for communication only and not talk to each other orally. So actually they were asked to keep quiet while participating in the user testing.

In order to avoid biasing, the first group evaluated first **System C** and then **System S**; vice versa for the second group. In order to evaluate all the differences between the prototypes, we provided the users with three tasks to accomplish in each prototype.

At the beginning each user in a session were assigned a color and a team. The exact texts of the tasks for the first prototype were:

**Task 1 - Find your way**: You should log on to the prototype and find the mark on the floor that has the same color as you and stand on that mark.

**Task 2 - Team up together**: You are a member of the Angels team. There are 4 members in your team. Go around, start to talk with everybody and try to find out your teammates. When you find one of them, stay together.

**Task 3 – Icebreaker – first prototype**: Let us say you have the whole evening off next Friday night. Your team is supposed to spend the night together so you have to decide on what to do that night. Please be precise, so for example, if you decide to go to the movies you have to say which movie you want to see. (If you have any questions you can send one member to the oracle in the game and ask.)

The participants were observed by two researchers where one of them was also a developer of the prototypes.

Before the evaluation each participant filled in a pre-evaluation questionnaire, collecting information on their background. After using each prototype the participants filled the AttrakDiff 2.0 questionnaire [4] on the user experience while using the prototype and questions about the comfort and expressiveness of the prototype interface and the believability of the conversations. After using both prototypes, each participant filled in a comparison questionnaire in order to understand strengths and weaknesses of both prototypes. After that there was a debriefing session which was audio recorded. In addition to that all the conversations between the participants through the prototypes were logged.

**Where & When – In Controlled Environment**

The user tests were conducted in a problem class room at the university. In that room there were at least 30 desktop computers. The user tests were conducted on running prototypes, see figure 2 and 3. The results were used for redesigning purposes and the work continued on the systems for over a year after the evaluation.

**THE ANALYSIS OF THE MAJOR DIFFERENCES**

When analysing the adaptions that needed to be made to be able to transfer the rather traditional approach of conducting user testing of work related software to the evaluation a multiplayer online game a comparison was made on the factors described for each case. The comparison of the factors is summarised in table 2. In the following the similarities and major differences are described.

The similarities are mostly in how the user testing is conducted in that sense that there were: a) 6 tasks b) the Attrakdiff 2.0 questionnaire was used, c) background questionnaire was used and d) there was a debriefing session in both the cases.

The main reason for the differences is that in the second case we were evaluating a multi-user game, so we needed many participants to be there at the same time. This made the observation during the task solving session much harder to conduct. We tried to get an overview of what was happening by standing behind all the participants and look over the area where the participants were sitting, but it was impossible to see everything that each of the participants were doing. Therefore we needed to rely much more on the logging of what happened and their comments in the debriefing session than in the think-aloud user tests in the first case.

Additionally, the goal of the game was to use it for communication, therefore we wanted the participants to use the game as much as possible for that purpose and keep quiet during the evaluation. Also, if they had been using the game by themselves they would not be communicating orally, so we wanted to mimic the real context in this way.
This was a bit hard for the participants and especially one of the groups did not really keep quiet, so they started commenting to each other which is understandable, because there was not that much distance between the participants, so this was easy for them.

Table 2: Summary of the Differences of the Cases

<table>
<thead>
<tr>
<th></th>
<th>Case 1 Work Related Tool</th>
<th>Case 2 a Multiplayer Game</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Which</strong></td>
<td>Think-aloud user testing</td>
<td>Adapted user testing</td>
</tr>
<tr>
<td><strong>Who</strong></td>
<td>Real users from two user groups, ordinary users and managers.</td>
<td>Groups of user surrogates using the two prototypes simultaneously</td>
</tr>
<tr>
<td><strong>What</strong></td>
<td>a detailed prototype of work related software</td>
<td>2 game prototypes</td>
</tr>
<tr>
<td><strong>How</strong></td>
<td>6 tasks, background questionnaire and 2 versions of Attrakdiff 2.0, debriefing. Users observed one at a time by two researchers</td>
<td>6 tasks (3 for each prototype), background questionnaire, Attrakdiff 2.0, comparison questionnaire, debriefing. Users observed by two researchers</td>
</tr>
<tr>
<td><strong>Why</strong></td>
<td>For researching purposes mainly (summative)</td>
<td>For redesigning purposes mainly (formative)</td>
</tr>
<tr>
<td><strong>When</strong></td>
<td>A very detailed prototype 2 weeks before launching</td>
<td>2 less detailed prototypes in the middle of development</td>
</tr>
<tr>
<td><strong>Where</strong></td>
<td>In the users own work environment</td>
<td>In controlled environment</td>
</tr>
</tbody>
</table>

Furthermore, because we wanted to evaluate two prototypes in the second case, the evaluation procedure was different from the traditional one. The users solved three set of two similar tasks during the evaluation, but that did not seem to affect the evaluation much though. Additionally the users were asked compare the prototypes in the end, which was of course not done the traditional user testing case.

Finally, because the game was still being developed, we could not involve the real users in the evaluations and the users were maybe a bit homogeneous group and we could not evaluate the game in real settings because, we were evaluating prototypes that we needed to run ourselves.

**CONCLUSION**

In this paper we describe the adaptations that needed to be done on different factors while transferring a traditional approach of conducting user testing on work related software to conducting user testing on a multiplayer game. Modifications needed to be done in all the factors described to fit this new context of conducting user testing.

We can recommend using the 6W and the H for describing the context which the evaluation was conducted in. It resulted in a clear presentation of what was done and was a good tool for comparing the two cases in this paper.

**ACKNOWLEDGMENTS**

We would like to thank all the participants that participated in both these cases, described in this paper.

**REFERENCES**