Netflix vs Viaplay

Group 3
Assignment 5
Usability Evaluation Project
Evaluation Methods in HCI

Niklas Blomqvist 900817-0939 nblomqvi@kth.se
Erica Bronge 920327-0344 bronge@kth.se
Annika Strålfors 881207-0442 stralf@kth.se

January 7, 2015

Abstract

In this paper a comparative study is presented that compare the usability between the on-demand streaming media services Netflix and Viaplay concerning effectiveness, efficiency and user satisfaction. The aim of the study was to evaluate if Netflix's interface is more user-friendly than Viaplay. Through a within-group design study in a laboratory environment, both qualitative and quantitative data was collected regarding the users ability to reach different targets and manage to perform a defined set of tasks. The results indicate that the total time that the participants took to complete all the tasks differed greatly between the two services. Some of the tasks stood out as being much harder to carry out with Viaplay than Netflix. The results from this study indicates that Netflix is more user-friendly than Viaplay due to some minor flaws in Viaplay's interface. The results are presented together with recommendations about how to improve the usability of both of the services.

Contents

1 Introduction

On-demand streaming media have in the recent years become a popular tool for watching TV-series and movies online. There are several actors on the market that are competing for the same customers. In order for an application to be successful, a user-friendly interface is of great importance. Netflix and Viaplay are two of the most widely used on-demand streaming media services available today. This study's aim was to explore the usability of Netflix and Viaplay by conducting a comparative evaluation of their web applications. The goal was to answer the following research question:

- *Is Netflix's interface more user-friendly than Viaplay's?*

Netflix and Viaplay were evaluated through a focus on effectiveness, efficiency and user satisfaction as usability measure. Effectiveness refers to the extent to which a product fulfill users expectations on how the product should behave and how easily the product can be used as intended [1]. This is often measured quantitatively through error rate. Efficiency describes the extent to which time, effort and cost is well used to accomplish the user's goals accurately. Efficiency is often evaluated by a measure of time. User satisfaction refers to the users perception and their feelings and thoughts about the interaction with the product, often measured through different kind of questionnaires. In this study, effectiveness, efficiency and user satisfaction were compared between Netflix and Viaplay regarding the users ability to reach different targets in the interface and manage to perform a set of tasks.

Through conducting a within-group design study in a laboratory environment, both qualitative and quantitative data was collected. This paper describes the design of the evaluation as well as previous research in the topic area. In section 4 the results are presented and followed by a subsequent discussion.

2 Related research

There have been many research studies that evaluate usability regarding effectiveness, efficiency and user satisfaction. Together with factors such as usefulness and learnability, it covers the concept of usability. Video streaming services have also been the objects for evaluations, although one might expect that the research area will increase with an expanding market. The following are some of the previous research in the topic area that have been taken under consideration during the process of this study.

The comparative research field contains studies with a wide variety of comparisons across different interfaces. A study published 2008 in the International Journal of Human-Computer Interaction, "A Comparative Study between

Tablet and Laptop PCs: User Satisfaction and Preferences", evaluated user satisfaction and preference aspects of Tablet PC in comparison to laptop PC [4]. By comparing the usability between different media the study aimed to explore users attitudes towards PC tablets and how to improve their usability. They identified common computer tasks and conducted a within-group design experimental study. Each task was followed by a subsequent questionnaire where user satisfaction and preference aspects were measured. The questionnaire evaluated multiple usability factors such as perceived task efficiency and effectiveness, overall satisfaction, perceived number of errors and enjoyment factors.

The rapid development of video streaming services creates needs for being able to evaluate the customers subjective perception on video streaming, named as Quality of Experiences (QoS). In the article "QoE-based Evaluation Model on Video Streaming Service Quality", a QoE evaluation model is proposed to predict the end users' perception on video streaming service considering different video content types [2]. The QoE model, specifically named Video-Mean Opinion Score (VMOS) focuses on measuring the end users' feeling.

Furthermore, in the article "Towards a combined method of usability testing: an assessment of the complementary advantages of lab testing, pre-session assignments, and online usability services", C. Jewell and F. Salvetti bring to light weaknesses and strengths of usability testing in a lab environment [3]. The main strength mentioned is that researchers get the opportunity to closely observe and understand the users' behaviour on a relatively detailed level when they are dealing with well-defined tasks. Drawbacks with testing in a lab environment are according to the authors the risk that people behave and use the tested interface differently because the lab environment isn't a natural setting and that it is a very resource-demanding evaluation method.

3 Method

The design of the evaluation was a comparative study that was carried out in the usability lab of KTH. The participants consisted of six KTH students that did not use any of the services on a regular basis but had good computer skills. Their task was to interact with both Netflix and Viaplay by performing a set of predefined tasks to compare the services. The Morae usability software was used to record and log the data during the test sessions. Prior to the usability test, a pilot test was conducted to ensure that the test instructions were perceived correctly in order to to optimize the design and to get some ideas about the time required.

3.1 Target group

The target group used in this evaluation consisted of six KTH students who did not use either of the services on a regular basis. Since on-demand streaming have become very popular in the recent years, finding participants that have not used any of the services before was not possible in the time given. By recruiting participants that are not frequent users it was possible to prevent biased results through top-down knowledge.

3.2 Tasks

The same set of tasks were used in both of the services and for all participants. The study used a within-group design where each participant started by conducting a set of tasks on one of the services before performing the same set of tasks on the other service. Since the services had some similarities in the interface, a risk of biased results through learning factors had to be taken into consideration. The order in which the participants tested the services was therefore equally divided so that half of the participants started with Netflix first and the other 3 tested Viaplay first.

The tasks consisted of three main tasks with different subtasks. The tasks were selected to represent some of the most common functionality used in video streaming media services. This included functions such as finding specific movies, changing subtitles, adding movies to the watch list, sort by genre and finding customer service information. The complete list of the tasks together with the complete test process can be found in the Appendix B in section 9.2.

3.3 Process

Each participant had at most 30 minutes in completing the test. Each participant was booked on a specific time slot in the usability lab so the total time of the experiment was about 3 hours (6*30=180). Prior to each test, the participants had to read the instructions and sign a certificate of consent to the experiment.

After the participants had read the instructions, they performed the tasks that appeared on the screen. They had the option to ask questions since there was an instructor in the same room. The instructor was there to avoid possible misconceptions and did not try to influence the participants performance of the tasks.

After each participant had completed the tasks of one service they filled in a questionnaire where they answered if they perceived that something was particularly good or confusing with the interface. This aimed to collect qualitative data for the comparison of the user satisfaction. After each participants had completed the set of tasks in both of the services, they were instructed to decide which of the services they preferred. By using a within-group design it was possible to log which service each person thought was the most user-friendly. The setup for logging the data with Morae software is explained in the sections below and the results are represented through diagrams in the results section.

3.4 Data to log

The data was collected with Morae usability software which gave the possibility of logging both quantitative and qualitative data for measurement of effectiveness, efficiency and user satisfaction. By recording the time it took for the user to finish each task it was possible to measure efficiency. Effectiveness was measured by the test supervisors taking notes on when the participants made errors while finishing each task. Since the screen also was recorded on video, it was possible to analyze how the tasks were performed in detail. Morae also allowed to collect qualitative data for the evaluation of user satisfaction through the use of questionnaires. This made it possible to collect subjective data on the users perception. The questionnaires were constructed as open questions that highlighted what the user liked about the interface and what they perceived as problematic aspects.

4 Results

The time for each task was recorded and their corresponding success rates noted. The scores were labeled from 0 to 2 (completed with ease, completed with difficulty and failed to complete). The answers from the surveys were saved and summarised (see Appendix A) where similar opinions were labeled together to give an overview of the participants opinion of the services.

4.1 Diagrams

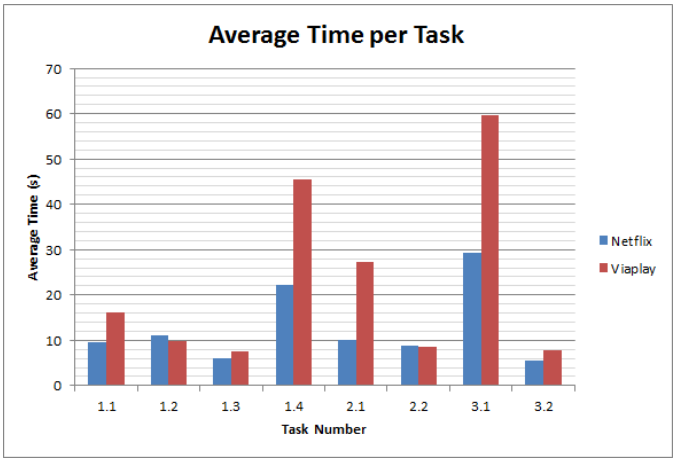

Figure 1: Average Time per Task

There were two tasks that took significantly more time for both services. These tasks took roughly twice the time when performed in Viaplay than in Netflix; Task 1.4: Add movie to your watch list and Task 3.1: Find contact information for customer support. A third task took longer time when participants tested the Viaplay service, Task 2.1: Look for movies by the genre "horror". Due to the three tasks standing out from the rest the standard deviations are high, especially for Viaplay. The standard deviation for Netflix was 8.4 and 19.9 for Viaplay.

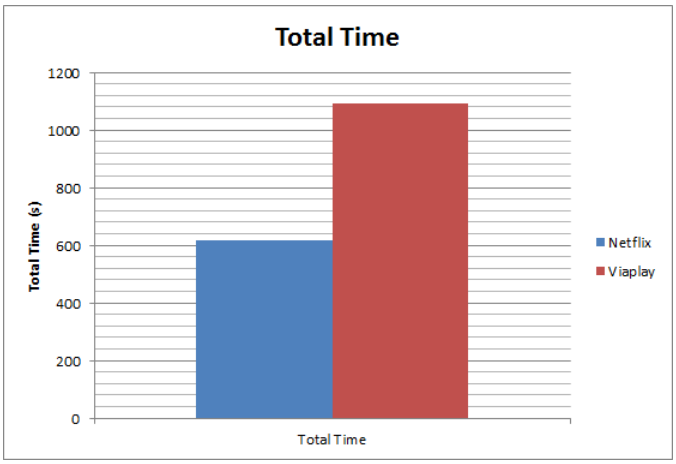

Figure 2: Total time per service

The total time for Viaplay was 1093 seconds (average 182 seconds per participant) and for 616 seconds for Netflix (average 103 seconds per participant).

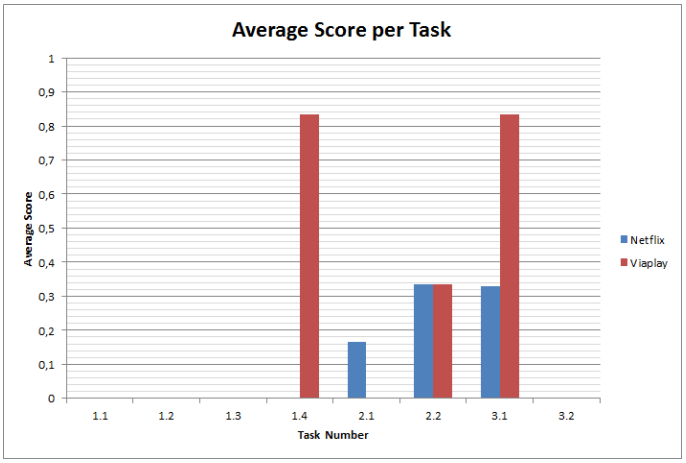

Figure 3: Average score per task

The scores were labeled from 0 to 2 (completed with ease, completed with difficulty and failed to complete). Task 1.4 and 3.1 when performed in Viaplay had significantly higher scores. The standard deviation was 0.15 for Netflix and 0.38 for Viaplay.

4.2 Examples of problems with the services

Below are some examples of problems with the services that was observed during the experiment and derived from the survey questions.

4.2.1 Viaplay

Adding movies to favourites was hard and not very intuitive. The marker for favourites (the star) was grayed out when clicked and white otherwise, many participants had problems knowing if they had clicked it or not. Participants had problem finding the customer service. The menu for customer service was in the footer of the page but the website had an "infinite scroll" so that movies were added to the screen when scrolling down. Furthermore, some participants thought it was unintuitive having to hover on a movie to get more information about it.

4.2.2 Netflix

Adding movies to the watch list (same as favourites for Viaplay) was somewhat challenging because you had to go back from the movie to get that option. One of the six participants had problem finding the customer service. It was furthest down on the page (same as Viaplay) but the website did not have an "infinite scroll" function so it was more intuitive to find it. Some thought it was unintuitive having to hover on a movie to get more information about it.

4.3 Answers from the last survey

In our final survey we asked "Which service did you like the most? Why?". Below are the answers where PX means the answer from participant X.

P1 *"Viaplay. It was a little harder to navigate but looked better. Netflix felt grainy somehow."*

P2 *"Netflix. Simpler, felt more intuitive (some parts). More logical structure (for me). Less steps to achieve results (some situations)."*

P3 *"Hard to tell. Netflix seemed to have a better 'recommend-for-you'-function, but I liked that Viaplay showed ratings from IMDb instead of an internal rating"*

P4 *"Netflix. felt like it had a greater range of movies as well as being the better service"*

P5 *"Netflix. Because it was easier to click the different menus"*

P6 *"Without involving the range of movies, Viaplay seemed slightly better"*

5 Recommendations

5.1 Both services

Adding a movie to the watch list (Netflix) or favourites (Viaplay) could be more intuitive. The option should be available when watching the movie, the user shouldn't't be forced to go back from the movie to find that option.

Some of the participants tried to use the search function when navigating on the site but it was lacking functionality. Adding so that users could find customer service and other things from the search function would increase the usability of both services. Another example is to be able to find your watch list or your favourite movies by searching for keywords like "favourites" or "my watch list".

5.2 Viaplay

Make the favourites button more intuitive. A simple solution is to color it yellow when clicked to let the user know they clicked it.

The "infinite scroll" function on the website is a nice touch and according to most participants they found the design on Viaplay's website to be superior. But having the menu for navigating to customer service and other places in the footer and having an infinite scroll function is maybe not the best combination. Adding functionality to the website so that the user can find the information elsewhere or having a "go to bottom"-button could be a solution to this problem.

5.3 Netflix

Netflix had no apparent problems that Viaplay did not have.

6 Discussion

6.1 Results

First of all, it is worth mentioning that the similarities in problems encountered by the users on the two sites were expected since their respective designs were very similar to each other. Even so, it is interesting to see that they still differed in some important cases, making a notable difference for the users with respect to site usability. Also worth noting is that more than half of the tasks were performed without any problem at all for the users, indicating that the sites overall are not disastrously designed. For our results, the features which lacked in measured efficiency and effectiveness in general coincided with the features users reported dissatisfaction in.

As we can see in the results, the total time users took to complete all tasks differed greatly between the two services. This difference mainly originated from three of the tasks; 1.4 which was to add a movie to watch list, 2.1 which was to sort movies by the genre "horror" and 3.1 which was to find contact info for customer support. Of these three, tasks 1.4 and 2.1 are the most relevant to discuss with respect to our research question since they represent the type of action users would be expected to take on a very regular basis when using the service. Task 1.4 can also be seen to have a pretty high score indicating that users were not even able to finish it as intended which is a serious problem for usability. Finding customer service is of course also important, but it is not something every user would like to do every time they use the service, and so the total time it takes (the efficiency with which the task can be solved) is of less importance for the essential parts of the service. So for tasks 1.4 and 2.1, why did the time differ between the two services affecting their respective usability and why did users have such trouble completing task 4.1?

Starting with task 1.4, add a movie to watch list, the task could be divided into several steps that the user had to perform to complete it. The first steps were the same for both services (leave the movie being watched, search it up again, bring up info box by hovering) and caused equal trouble, so the difference in time can be explained by the last step of actually adding the movie to watch list. This was supported both by user feedback and by our observations.

For Viaplay, it might not have been just as obvious that the star represented the action of saving the movie to your watch list as a text explicitly saying "Add to watch list". It is important to note however that our task formulation here might have had a lot of influence here since we asked the users to add to watch list/favourites. If we had asked to "star-mark" or something like that, maybe the result would have been different. Additionally, users expressed difficulty

with knowing if the star button on Viaplay actually worked since it appeared that not much happened when clicking it. This appears to be an issue with providing proper feedback to the user. For example Susan M. Weinschenk writes in "100 Things Every Designer Needs to Know About People" (2011) that it is important to take note of the salience of graphical cues since it is more likely that people will pay attention to more salient things [5]. Thus, the change in color from white to gray on a star was probably not salient enough for users to realize something important had happened, hurting the usability.

One thing which was also interesting is that half of the test participants found having to hover over a movie for information and alternatives unintuitive and took note of this. Even so, no one seemed to have any particular problems with this step. For task 2.1, sort by genre horror, none of the user expressed that this was difficult on either service, the result is therefore slightly unexpected. As can be seen in Figure 3, this task did not have a particularly high score indicating bad effectiveness either. Here, a mistake in our task design might also be a contributing reason to the difference since we asked users to sort by genre "skräck" although it was called "rysare" in Viaplay.

6.2 Method

We are satisfied with having chosen within-group design for our test. The only drawback was the learning effect which definitely was present, but since we let half of the participants use the services in reverse order, it did not affect the end result notably. If we would have had a little more time guaranteed in the usability lab, this and our other statistical data on efficiency and effectiveness could have been further improved by a slightly larger number of participants.

Concerning the recruitment of participants, this was also a limiting factor for the number of participants we could have. We set up a doodle poll with time slots people could chose from, and sent it out to people of our target group. Since the test was planned to take up 30 minutes, it was hard to find people willing to spare the time. When the test day came, we also had an issue with people not turning up on time. Thankfully, it didn't pose too much of a problem since most participants didn't use their entire time slot, so we were able to work it out on the go.

To get a better idea of how Netflix and Viaplay are used normally, we could have chosen to try to resemble their target groups more accurately when selecting participants for our test. Instead we chose to include people who we knew were part of the target group but who did not frequently use the services, and who had good computer skills. In "Handbook of Usability Testing", Jeff Rubin and Dana Chisnell warns for for inadvertently testing only expert users who will eas-

ily accomplish all tasks of the test, but since we are aware of this we have used it to our advantage instead [1]. Testing only people with good computer skills helped us find the most severe problems with the service, since if a person with good computer skills have trouble using a web site for the first time, it is very likely that other people doing the same will have problems too. It was also a positive thing that this group of people was fairly easy for us to get in contact with.

Earlier mentioned was Jewell and Salvetti's conclusion that testing in a lab environment could make users behave in a way they would not do at home. This was certainly something we kept in mind when designing our test. However, since we wanted to see how good usability the services' interfaces showed for normal use tasks we designed the test with a specific set of tasks that the users were to perform. Having done this, we found that testing in a lab environment would not matter that much since we already possibly had forced the participants into behaving they would not normally do. Instead, testing in a lab environment seemed more beneficial since we also got the benefits Jewell and Salvetti mentioned, possibility to closer watch the participants and their interactions with the interfaces.

The use of the usability testing software Morae is another important thing to discuss, because although we are overall content with our choice, we ran into some small issues during the test. We had planned our test under the assumption that we would be able to manually control when task instructions were showed to the user on the screen, but when setting up the test environment we found out that this was not possible. We therefore ended up having to ask the users to themselves click a button to start each task (and thereby start logging the time), and then click it again to end the task when they thought they were done.

First, the above mentioned method led to people sometimes forgetting to start the task before performing it. Even though the test supervisor told them that they forgot as soon as he or she saw it, this may still have had a small impact on the task time measurements. Second, one or two participants accidentally double clicked the button so that they moved on to the next task without even performing the previous one. This was a more serious problem because of how we had set up the scoring for the tasks. As mentioned previously, we only had "completed with ease", "completed with difficulty" and "did not complete" leading to us tagging skipped tasked with "did not complete". Therefore, our effectiveness score may be inaccurate, leading to a serious weakness in the study. We should have had a fourth tag for marking tasks which were not completed because of this type of error rather than because of deficient usability of the tested interface.

Finally, as mentioned in the results discussion, some of our wording may have been creating a bias for one or the other of the services. Therefore, if we would

repeat the test we would make sure to look over the wording a bit more carefully to prevent this kind of bias. The wording also proved to be a problem for one participant in particular, who wasn't a native Swedish speaker. This participant did not know that the word "rysare" was equal to "skräck" since young people rarely use it in common speech. For future testing, this could be solved by making sure that all participants are native speakers of the language of the test or by changing the wording in the questions to include both synonyms in cases like this.

7 Conclusions

Our most important conclusion would be the answer to our research question; Yes, Netflix's interface is more user-friendly than Viaplay's. However, the issues which cause the difference in user friendliness are small and quite easily solvable with the recommendations we have put forth in this report. This also goes to show that even very small details can radically worsen usability of an interface. We can also conclude that we had some issues with the test design that calls for caution with reading too much into the test results. The best thing would be to confirm the results by redoing the test while taking into consideration the issues learned about here.

8 References

[1] Jeffrey Rubin, Dana Chisnell *Handbook of usability testing: How to plan, design, and conduct effective tests* 2008

[2] Yun Shen, Yitong Liu, Nan Qiao, Lin Sang Dacheng Jang *QoE-based Evaluation Model on Video Streaming Service Quality* Beijing Univ. of Posts Telecommunication, 2012

[3] Christopfer Jewell, Franco Salvetti *Towards a combined method of usability testing: an assessment of the complementary advantages of lab testing, presession assignments, and online usability services* 2012

[4] Ant Ozok, Dana Benson, Joyram Chakraborty Anthony F. Norcio *A Comparative Study between Tablet and Laptop PCs: User Satisfaction and Preferences* International. Journal of Human Computer Interaction, 2008

[5] Susan M. Weinschenk *100 Things Every Designer Needs to Know About People* New Riders, Berkely, 2011

9 Appendices

9.1 Appendix A: Summary of the surveys

- Was there something with the service that you experienced as confusing or bad? If so, what?

Netflix

Couldn't add the movie you were looking at to the watch list, you had to go back and then add it. The process could have been more intuitive. (2 participants)
Customer support was hard to find. (1 participants)

Viplay

Hard to add a movie to favourites, the star-marker became gray when clicked and it was white otherwise and it was hard to see if you had clicked it. (5 participants)
Customer support was hard to find. (2 participants)

- Was there something you thought was especially good with the service?

Netflix

Easy to navigate, simple interface. (5 participants)

Viplay

Easy to find, simple interface (2 participants)

Other comments (Netflix)

Unintuitive that you needed to mouse-over a movie to see information and options. (2 participants)

Other comments (Viaplay)

Unintuitive that you needed to mouse-over a movie to see information and options. (1 participant)

- Which service did you like the most? Why?

P1 *"Viaplay. It was a little harder to navigate but looked better. Netflix felt grainy somehow."*

P2 *"Netflix. Simpler, felt more intuitive (some parts). More logical structure (for me). Less steps to achieve results (some situations)."*

P3 *"Hard to tell. Netflix seemed to have a better 'recommend-for-you'-function, but I liked that Viaplay showed ratings from IMDb instead of an internal rating"*

P4 *"Netflix. felt like it had a greater range of movies as well as being the better service"*

P5 *"Netflix. Because it was easier to click the different menus"*

P6 *"Without involving the range of movies, Viaplay seemed slightly better"*

9.2 Appendix B: Test process

1. Task 1
   (a) Find a movie you like and start watching it
   (b) Forward it to minute 40
   (c) Put on subtitles in Swedish
   (d) Add the movie to your watchlist
2. Task 2
   (a) Look for movies by the genre "Horror"
   (b) Find the highest rated movie in that genre
3. Task 3
   (a) Find contact information for customer support
   (b) Log out of the service

9.3 Appendix C: Consent form

Syftet med detta experiment är att utvärdera användarvänligheten hos två webb-baserade tjänster för video-on-demand. Experimentet kommer att genomföras under 30 minuter då du som deltagare kommer ombes att använda de två tjänsterna för att genomföra ett antal uppgifter. Du som deltagare kan när som helst välja att avbryta experimentet. Under experimentet kommer data att samlas in på två sätt: genom inspelning av skärmen och genom de skriftliga svar som du som deltagare ger på ett antal frågor rörande de tjänster som utvärderas. All samlad data kommer att vara anonym och kommer inte utlämnas till tredje part eller användas i vinstgivande syfte.

Jag intygar härmed att jag läst ovanstående och samtycker till att delta och att den insamlade datan kan användas i forskningssyfte:

Signatur:

Namnförtydligande: