A Case Study of Software Replacement

Marta Kristín Lárusdóttir
Reykjavík University
School of Computer Science
Ofanleiti 2, 103 Reykjavik, Iceland
marta@ru.is

Sigrún Eva Ármannsdóttir
Hugur ltd.
Grjótháls 5, 110 Reykjavík, Iceland
sea@hugur.is

Abstract: This paper describes a case study, where the impact of introducing to users a new software system to replace an existing one was measured in the users' own environment. Workers at an aluminum company in Iceland used an old character-based system for keeping track of spare parts for maintenance. When planning to introduce new Windows software for solving these tasks, a decision was made to measure the effect of the replacement by doing usability evaluations with the users. First, the old system was evaluated, then the new one shortly after introducing it to users and finally, the new one again after six months usage. The results show that the new system does not in all cases give the users a more effective and efficient system for solving their tasks. Still the users were more satisfied with the new system. Furthermore, tasks that scored below a somewhat unidentified margin in the usability evaluation did not improve in terms of effectiveness and efficiency after the six months usage of the new system.

Keywords: Usability, user testing, user's environment, think-aloud, replacing software.

1 Introduction

In September 2001 the IT manager at a subsidiary aluminum company located in Iceland contacted us, two usability experts, in order to measure the impact of replacing a software system. The company was using a locally custom made character-based system, which had been in use for 13 years, however, the mother company had decided that all its aluminum companies around the world should from now on use the same system. Therefore, the workers in Iceland had to switch to the new system developed in USA, with the user interface translated to Icelandic. "We want to measure the impact on our users switching from the old systems they know, to the new system they must use from now on", the IT manager explained. The challenge was clear and a decision was made to measure the usability of the systems by testing users in their own environment during three different occasions. First, the old system to get reference points, second, the new system after two weeks of use and third, after six months use, when all users were expected to have learnt to use the new system. This allowed the new system to be compared to the old one in a fair manner. No changes were made to the new system during the six months of use so this case study gives a very good insight into what impact it can have on users replacing an old system with a new one and how it can be measured.

The ISO-definition on usability [4] is used in this study as a basis for measurements, where the three usability factors; effectiveness, efficiency and satisfaction are described as:

- *Effectiveness:* ‘the accuracy and completeness of which users achieve specific goals’. Focus is only on completeness and accuracy of task completion, regardless of how much effort is made to complete a task, e.g. time;
- *Efficiency:* ‘the accuracy and completeness of goals in relation to the resources expended’. Resources being e.g. time or mouse clicks.
- *Satisfaction:* ′the comfort and acceptability of the system’.

In this study usability was measured by doing think aloud test as described in many textbooks on human-computer interaction, e.g. by Nielsen, Faulkner, Preece et. al. and Molich [7, 2, 9, 6] to name a few. The users solved predefined realistic tasks in their own work environment and were asked to think aloud, that is, say everything they thought of while working. An observant registered all their comments. Additionally, information was gathered through questionnaires, both for background information, between task satisfaction and after the test satisfaction. A recent survey by Gulliksen, et. al, [3] on the usability profession in Sweden, shows that the think-aloud method is rated as the best method professionals used or had used for user involvement. This is also the result in a survey that Rosenbaum, et. al. [10] conducted where usability evaluation in a lab and usability evaluation without a lab or outside of a lab facility get the best ratings from usability professionals that had used some of the 29 methods and approaches mentioned in the study.

The remainder of the paper is structured as follows. In section 2 the methodology of the case study is explained, introducing the systems involved, the testing methods, the procedure of the study and the participants. In section 3 the results according to effectiveness, efficiency and satisfaction are described. A conclusion is in section 4.

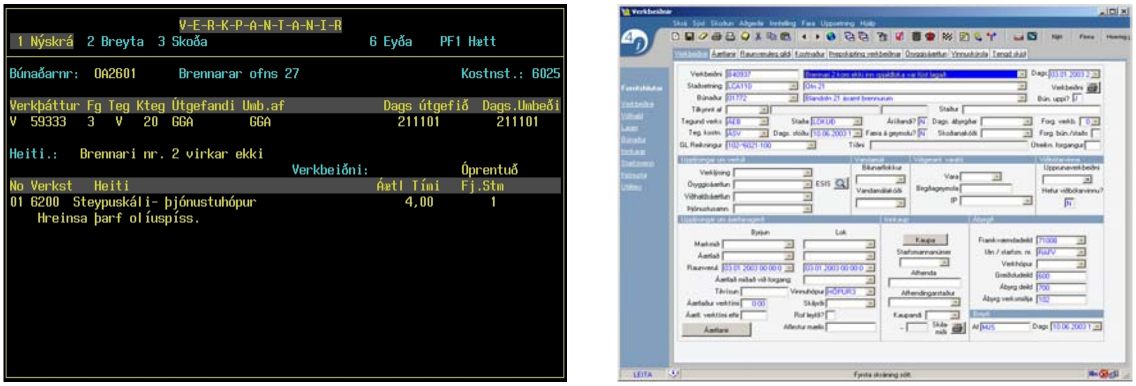
2 Methodology

In the following the systems evaluated are described, and examples of typical user tasks are given, then the procedure of the evaluation is described, the participants and finally the data gathering methods used.

2.1 The systems evaluated

CSD system: The CSD system was a character based system running on PC, see Figure 1. It was in use at the aluminum company from 1989 to 2002 that is for 13 years. It was a custom made software and the user interface was in Icelandic. The users had to remember a lot of codes from one screen to another and almost all of them had a writing block beside the screen for writing down codes. Three supplementary systems had been made to make the use of CSD easier, mainly for looking up information and codes. So the users used CSD in combination with the other three systems, switching many times between them while solving one single task. In the remainder of the paper, the CSD system is called the old system.

Maximo: Is a windows system with a WIMP (*W*indows, *I*cons, *M*enus and *P*ointing device) interface, see Figure 1, developed by MRO Software (www.mro.com) in USA. The user interface was translated to Icelandic, but the structure of the system was not changed. The users could more or less solve all their tasks in the system, so they did not have to use the supplementary systems that had been made for the old CSD system. In the remainder of the paper this system will be called the new system.



```
                     V-E-R-K-P-A-N-T-A-N-I-R
 1 Nýskrá  2 Breyta  3 Skoða                     6 Eyða    PF1 Hætt

Búnaðarnr:  OA2601      Brennarar ofns 27                Kostnst.: 6025

Verkþáttur Fg Teg Kteg Útgefandi Umb.af            Dags útgefið  Dags.Umbeði
V  59333   3   V   20  GGA        GGA                 211101        211101

Heiti.:   Brennari nr. 2 virkar ekki
                                         Verkbeiðni:           Óprentuð
No Verkst   Heiti                                   Áætl Tími   Fj.Stm
01 6200  Steypuskáli- þjónustuhópur                    4,00        1
    Hreinsa þarf olíuspíss.
```

Figure 1: Screenshots from the CSD (the old system) to the left and Maximo (the new system) to the right

2.2 Examples of user scenarios

A user scenario using the old system: For inserting information in the old system, the user opened it and started working. When going to insert an item number, the user had to insert the letters in the code. However, the user did not get any assistance within the system, so the user opened a complementary look up system on the intranet and searched there for the item code and after that switched back to the old system again. Then when planning to insert the cost-code the user looked at a list on paper over the cost codes laying beside the computer and inserted that into the old system. Many users commented that they were not allowed to use that paper any more, because some outdated codes were on the paper, but it was so convenient, that they did it anyway. When inserting the workshop code, many of the users remembered the code correctly and completed the task.

A user scenario using the new system: For inserting requests in the new system, the information on the item code was structured in a tree layout, and many users used that to find the item code. The user inserts all the data, having drop down lists or other guidance for inserting the codes. So the user did not enter other systems, especially when having used the new system for some time. However, the user was asked to insert more entries for inserting a request in the new system than in the old one.

2.3 Procedure of the study

The usability evaluation was carried out on three different occasions, i.e. in three parts.

Part I: The old system was evaluated by running think-aloud tests by 6 users in their environment at the aluminum company. First the users completed a pre-evaluation questionnaire for gathering information on their background. They also completed three task related questions after each task and a post-evaluation questionnaire for finding the benefits and drawbacks of the system evaluated. These measurements were carried out in December 2001. Figure 2 illustrates the procedure of the whole study.

Part II: The new system was evaluated in the same way as the old one was in part I, after two weeks usage with four users from part I and 4 new users. These measurements were carried out in April 2002. All the users got the same training in using the system, before the user tests were made.

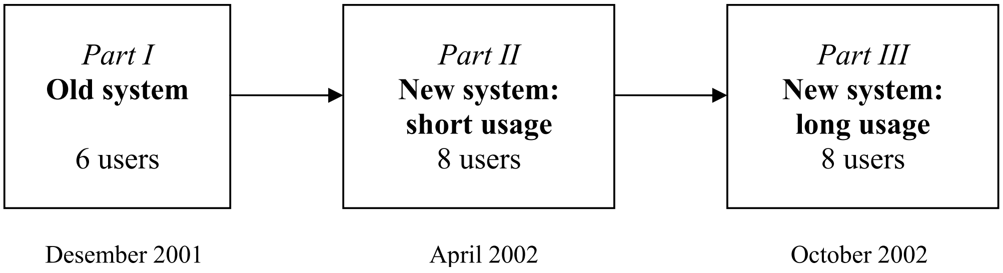


Figure 2: The procedure of measuring the quality of use of two software systems

Part III: The new system was evaluated again after six months usage with the same 8 users as in part II, following the same procedure as in part I and II. These measures were carried out in October 2002.

2.4 Participants

The 6 participants in part I were all male employees and all foremen. The participants were chosen by the technical manager at the aluminum company. Their job was to fix defected equipment of the aluminum company, order spare parts and keep track of how much time had been used for each request. All had worked there for more than a year and two of them for more than 15 years. They used the old system up to 4 hours a day. Seven said they were intermediate computer users and one beginner. The users worked in three different work shops: the main work shop, the vehicle work shop and in the cast house work shop.

In part II both foremen and regular workers were tested. Four of the six foremen in the first study took part in part II, and 4 new users. The 8 participants were all male employees; all had worked there for more than a year and four of them for more than 15 years. All participants had joined a course on how to use the new system and had now used it for two to four weeks. Seven used the system up to 4 hours a day, thereof four for less than an hour and 1 for up to six hours a day. Five were intermediate computer users and 3 beginners.

In part III the same users were involved as in part II.

2.5 The data gathering methods

Data was gathered by performing usability tests, namely think aloud tests with supplementary questionnaires and informal interviews.

2.5.1 The think aloud tests

The think aloud tests were done at the users' ordinary work place. One of the usability experts functioned as a conductor and kept track of the completion time of the tasks. The other usability expert functioned as a secretary, registering if the task was completed or not, the usability problems, the users' comments, the data entered and the number of changing from one system to another. So this person functioned as a live video camera.

The users performed the same 6 tasks in the think aloud tests in all three parts of this study. The tasks were made by a project manager at the aluminum company, which knew exactly what tasks are relevant to the specified users. There were two tasks for inserting a work request for fixing defected equipment, one for an ordinary case, and one for a claim event. An example of a task is: 'A compressor nr. 1 in a compressor facility does not work. You need to change the filter. Choose your own workers group.' Two tasks were made for looking up information; one for checking if some work had been done on a particular work request and the other for finding information in the work request form. Finally two tasks were made to work on spare parts, one for reserving spare parts from the stock and the other for checking how much of this item was in stock and where it was situated in the stockroom.

After the tests in part I, the tasks needed to be modified to represent relevant products and items for each of the three work shops involved. The users got very frustrated, if they were asked to solve a task that involved some equipment, they were not used to handle. So in part II and part III, the data included in the tasks was adjusted to each work shop, but the goals of the tasks were still the same.

2.5.2 The questionnaires

There were three types of questionnaires: a background questionnaire, after-task questionnaire and after-test questionnaire. The background questionnaire contained 10 questions collecting information on personal data and the technical and professional experience. The after-task questionnaire contained 5 questions on: 1) how realistic the tasks were, 2) how easy the tasks were, 3) how often they solved similar tasks, 4) if it took more or less time to solve the task than usually and 5) how sure they were that they had completed the task in the correct way. Finally there was space for user's comments. The after-test questionnaire contained 3 questions, the first asking how much the user used the system tested in his daily work, the second asking the user to state the strengths of the system and the third asking about the weaknesses.

3 Results

In this section the results of the usability evaluations are described, first according to effectiveness, then efficiency and finally satisfaction.

3.1 Effectiveness

As stated earlier, effectiveness is: 'the accuracy and completeness of which users achieve specific goals' according to the ISO-definition of usability [4]. In this study the user's goal is to complete a predefined task with the correct data inserted.

In Table 1, the percentage of users completing a task with the correct data, with incorrect data and not completing the task is shown for the three parts of the study. As measurement for the completeness and accuracy factors in the effectiveness definition, the percentage of users completing the tasks with the correct data can be used. When completing a task with incorrect data one has to assess how serious it is for the user to insert incorrect data and how much attention should be given to that measurement. In this study it was not dangerous to insert incorrect information, but of course the task was not completed in the perfect way, and the data filed was incorrect.

Table 1 shows that when comparing the new system after 6 months of use to the old system, the new one is better for the users, when solving the two look up tasks. In *look up task – 1* half of the users give up solving the task in the old system and 17% give up while solving the other look up task, but all users complete both the look up task in the new system. But when working with spare parts, more than half of the users give up while solving both tasks with the new system, which is much worse than in the old system.

Table 1: The percentage of users completing successfully, completing with incorrect data and not completing tasks

User tasks	**Completed with the correct data**			**Completed with incorrect data**			**Not completed**		
System used	**Old**	**New: short usage**	**New: long usage**	**Old**	**New: short usage**	**New: long usage**	**Old**	**New: short usage**	**New: long usage**
Insert request -1	33%	0%	25%	67%	88%	75%	0%	12%	0%
Insert request -2	50%	0%	0%	50%	75%	100%	0%	25%	0%
Look up -1	50%	100%	75%	0%	0%	25%	50%	0%	0%
Look up -2	83%	75%	100%	0%	0%	0%	17%	25%	0%
Spare parts -1	50%	25%	38%	17%	17%	0%	33%	50%	63%
Spare parts -2	100%	25%	25%	0%	0%	13%	0%	75%	63%

It was surprising that the users had to give up finishing tasks in the old system. One explanation could be that in a think aloud session things are not the same as in the user's daily work, even though everything was done to make the sessions as realistic as possible. The users comments in the think aloud sessions showed that the data that the users were asked to work with in the think aloud sessions in the old system were not the data that they were acquainted with and used in their daily work, so that could also explain some of the troubles the users had. If the users had been alone they would probably have asked someone for help, or even asked somebody to solve the task, but when being observed they did not want to do that. Efforts from the IT management had been made to make these look up tasks easier by making supplementary systems for the users, but still half of the users did not complete one of the look up tasks.

To summarize the results from Table 1, task effectiveness is calculated according to definition by Bevan and Macleod [1]. This way of calculating the effectiveness is further explained by Macloud, Bowden and Bevan [5]. There effectiveness is described there as both quantitative and qualitative, where quantitative effectiveness is 'the proportion of task goals represented in the output of a task which have been attempted'. The qualitative effectiveness is described as: 'the degree to which the task goals represented in the output have been achieved'. Task effectiveness is then defined as:

*(1) Task Effectiveness = (Quantity *Quality)/100 %*

In this study, the quantitative effectiveness is defined as 100%, if the user completes the task (with the correct or incorrect data), but 0%, if the user does not complete the task. The qualitative effectiveness is defined as 100%, if the task is completed with the correct data, 50% if completed with incorrect data and 0% if not completed. This is in harmony with Nielsen's calculations of success-rate [8]. To take an example of the calculation of task effectiveness for the task *Spare parts -1* in the old system is shown in Table 2.

Of course, the output of the users tasks could have been analysed further giving more detailed quantitative and qualitative effectiveness, but that needs further studies of the data. Examples in the literature of how to exactly calculate these numbers are very few, so this case study could function as one example.

Table 2: The task effectiveness for the solving task: *Spare parts -2* in the old system.

User	Completed the task	Quantity	Quality	Task effectiveness
V1	**Yes**, with the correct data	100	100	100%
V2	**Yes**, with incorrect data	100	50	50%
V3	**Yes**, with the correct data	100	100	100%
V4	**No**	0	0	0%
V5	**Yes**, with the correct data	100	100	100%
V6	**No**	0	0	0%
Total				58,3%

In Figure 4 the task effectiveness is shown for the all the think aloud tests for all the six tasks. Two of the tasks have higher task effectiveness in the new system than in the old system, one task has similar task effectiveness in the new system after six months use, and three tasks have lower task effectiveness in the new system even after 6 months of use.

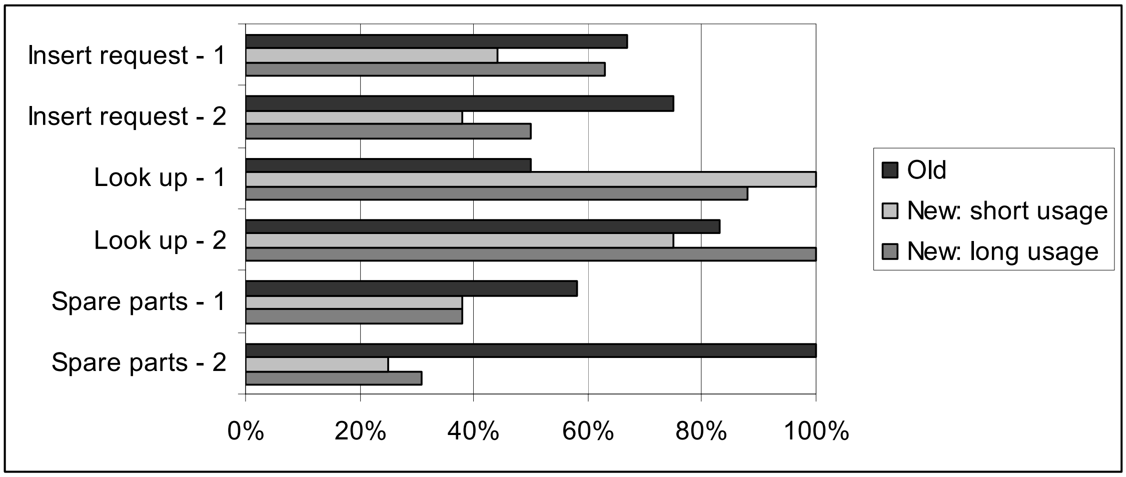


Figure 4: The task effectiveness calculated for the old system and the new after short usage and long usage

Learnability can be measured by comparing the effectiveness of the new system after two weeks usage and after 6 months [1] as seen in Table 3. This is done to see how users improved by using the new system for 6 months. After two weeks usage, the two spare parts tasks scored badly, the efficiency was 38% and 25%, so did also the inserting tasks, but the look up tasks had good effectiveness, 100% and 75%.

Table 3: The learnability of the new system measured by comparing the effectiveness

User tasks	Effectiveness		
System used	**New: short usage**	**New: long usage**	**Differ-ence**
Insert request -1	44%	63%	19%
Insert request -2	38%	50%	12%
Look up -1	100%	88%	-12%
Look up -2	75%	100%	25%
Spare parts -1	38%	38%	0%
Spare parts -2	25%	31%	6%

There is a very interesting trend in the data, the tasks that scored badly after the two weeks still did after 6 months. It was as if the users had decided that the system was so badly designed for solving these tasks that they did not want to try to learn how to solve them. We got comments in that direction from the users while thinking aloud and in the informal interviews. This could point to an existing margin within the completion rate which could indicate that a task will not better itself in terms of usage if the rate is below that margin during the time when users are starting out with a new system. To find out what that margin is and if it truly exists, further studies would have to be carried out. A lesson to learn from this is that it is very important to design the quality into the system, if the system is not effective in the beginning, it does not matter how much time you give users, the effectiveness will not improve.

3.2 Efficiency

*A*ccording to the ISO-definition of usability [4], efficiency is: 'the accuracy and completeness of goals in relation to the resources expended'. To measure the resources expended the average time to complete each task is calculated and can be seen in Table 4. These results are based on the time for all completed tasks both with the correct and incorrect data. The time users used on trying to solve the tasks that they did not complete is not included in the calculations. Task time is often used as a measurement of quality of use. This is very suitable, especially if all the tasks were completed with 100% effectiveness.

Table 4: The resources used for each completing each task (both correct and incorrect data)

System used	Old system		New: short usage		New: long usage	
	Average time	(Stdev, N)	Average time	(Stdev,N)	Average time	(Stdev, N)
Insert request – 1	2:58	(0:58, 6)	5:59	(4:45,7)	3:24	(0:46,8)
Insert request – 2	3:39	(1:41, 6)	5:32	(2:23,6)	5:38	(4:21,8)
Look up – 1	1:02	(0:17, 3)	0:38	(0:24,8)	0:37	(0:18,8)
Look up – 2	1:35	(1:43, 5)	2:50	(1:05,6)	1:28	(0:39,8)
Spare parts – 1	3:50	(1:16, 4)	11:10	(3:13,4)	6:06	(2:35,3)
Spare parts – 2	2:18	(0:32, 6)	11:55	(n/a, 2)	9:04	(3:23,3)

In Bevan and Macleod [1] the user efficiency is defined as:

(2) User efficiency = Task effectiveness / Task Time

This gives the opportunity to combine the accuracy, the completeness and the resources expended in one measurement. In Figure 5 the user efficiency for each task is shown for the three evaluations by combining measures on how many users completed the task with correct and incorrect data and divide with the average task completion time. The advantage of the efficiency measure is that it provides a more general measure of the usage of the system by trading off the quantity and quality of the task completion against time of completion.

In Figure 5 we can see that users got a much better tool for solving the look-up task 1 using the new system than using the old system. The look-up task 2 was less efficient in the new system after two week of use than in the old system, but after 6 months usage of the new system, the efficiency was better there than the efficiency of that task in the old system. For inserting a request the efficiency is similar in the old system and in the new after long usage. The remaining three tasks have much higher rate for the efficiency in the old system than in the new. So when combining the measures on effectiveness and task time, we can see that the users got a worse system after six months usage of the new system for solving three tasks, similar for solving one and better for solving two tasks.

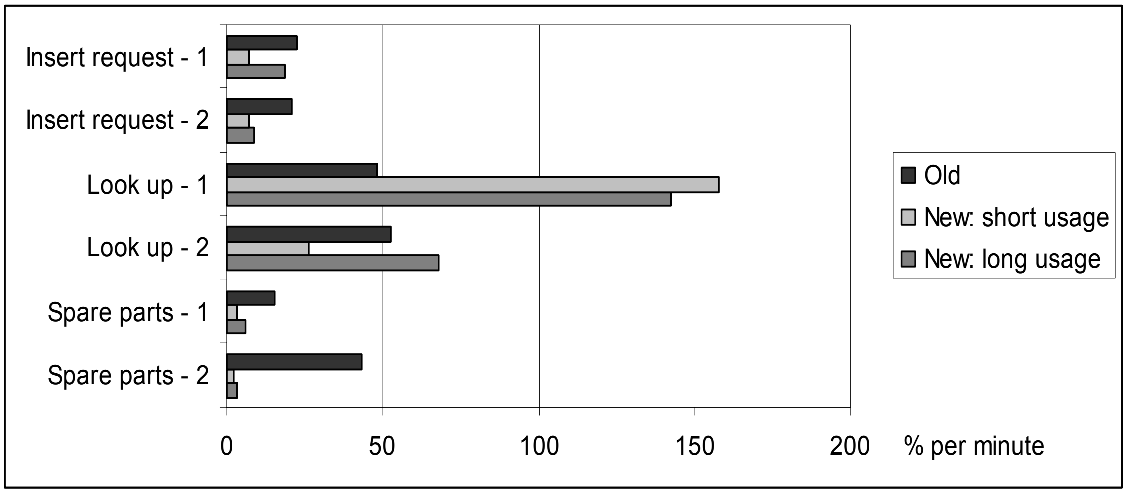


Figure 5: The user efficiency - The % of effectiveness per minute comparing the systems

Another way of looking at the resources involved is to count how many systems the users had to use to solve one task. Many users complained that when using the old system, they had to use a lot of supplementary systems to be able to solve one task including writing codes on paper. In Table 4 we see that the user's mental load was less in the new system, because they could solve the tasks in many cases in the new system alone.

Table 5: The average number of systems used for solving the tasks

System used	**Old**	**New: short usage**	**New: long usage**
Insert request – 1	3,50	1,57	1,25
Insert request – 2	3,00	2,17	1,00
Look up – 1	1,00	1,00	1,00
Look up – 2	1,00	1,00	1,00
Spare parts - 1	3,75	3,00	1,37
Spare parts - 2	2,33	2,00	1,12

***1 means that the user used one system to solve the tasks**

When using the old systems all the users switched often between the main system and supplementary systems, when solving four out of the six tasks. After two weeks usage of the new system, the users were used to use the supplementary systems, even thought they did not need to, but after six months the users used the supplementary system much less.

3.3 Satisfaction

To measure user's satisfaction, the users were asked to fill out a questionnaire when all task were completed, writing down the benefits and drawbacks of the system in use. The results are listed in Table 6.

Table 6 The benefits and drawbacks of the system in use

Old system	
Benefits	**Drawbacks**
I know how to use it (2 users)	Too many systems that we have to use to solve one task (2 users)
It makes it easy for me to do my job and advances that the work is successful	It takes too much time to change between systems
Simple and relatively easy to use	When I have inserted a request and printed it out, and something is wrong I can't print it again
	I would like to see this in the Windows system. The progress of the task-request is missing. The amount of work shows up at a late stage in the system. It is hard to change the type of work request.
New system: long usage	
Benefits	**Drawbacks**
It is easy to work in the system	The data on spare parts are missing
The main advantage is that it combines many other systems that we used to use in one system. Furthermore I think it is much easier to register my work in the new system.	The main drawbacks are that data on the equipment in stock is missing and I find it sometimes hard to find some spare parts. The description of spare parts does not have the correct wording, the brand name that the producer uses should also be in the description
One system	Information equipment and the placement in stock is missing
I find it easy to use, the part I know in the system	It takes too long time to insert into tables.
I don't know	The system needs improvements
Easier to use	
It is better than the old one, (but can be improved)	.

Overall, the users seemed quite happy with their new system, most drawbacks they mentioned could be blamed on the data available to the system, not the system itself. So the results on user's satisfaction were better than the results shown within the other two factors measured, effectiveness and efficiency. User's satisfaction is not affected by the other factors as much as expected.

The same results on satisfaction were obtained by looking at the comments given by users during the think aloud testing. The comments, both those omitted spontaneously and those more carefully worded by users, revealed that users were quite happy and excited about the new system and very ready to forgive any faults the new system had, both after two weeks of use and more so after six months of use. It was as if they felt that the company really appreciated their work, since an effort was taken to introduce a new up-to-date system for them. This gave the users feeling of importance an extra boost.

4 Conclusion

This paper has given an account of a case study, where the impact of introducing to users a new software system to replace an existing one was examined. Usability evaluations were carried out in the users' own environment. First, the old system was evaluated, then the new one shortly after introducing it to users and finally, the new one again after six months of usage.

The main findings of the study were that the new Windows interface did not in all cases evolve into a more effective and efficient system for users to solve their tasks with. Half of the tasks were less efficient and effective to solve than in the old character-based system and only one task was clearly easier to solve. It did, however, benefit users in that they eventually used supplementary systems much less than before and were happier with the new system than the old one.

When considering learnability between the two evaluations of the new system, the solving of the tasks did not improve much. Especially those tasks that proved very difficult for users to solve to begin with. They had improved to some extent but much less than expected. Therefore, it seems extremely important that each task is usability tested and redesigned if necessary before handed out to users, otherwise the effectiveness of the system will not improve with time.

For future work, this case study may also have given indications of an existing margin within the completion rate. This margin could indicate that a task will not better itself in terms of usage if the rate is below that margin during the time when users are starting out with a new system. However, this study being way too small, both in terms of number of users and number of tasks, to give conclusive evidence and to find out what that margin is and if it truly exists, further studies would have to be carried out .

Acknowledgements

We would like to thank Ásrún Matthíasdóttir, assistant professor at Reykjavik University, for good comments on an early draft of the paper, Björn Þór Jónsson and Yngvi Björnsson, associate professors at Reykjavik University for their very valuable comments on the final draft and Guðmundur Valsson, consultant at Ráðorka ehf for his inspiration and good critique. Finally, a special thank you to Þórunn Pálsdóttir, for giving us the opportunity to carry out this work and for approving of the final version of the paper.

References

1. Bevan, N. and Macleod, M. *Usability measurement in context,* Behavior and Information Technology 13, 132 – 145.
2. Faulkner, C. (2001): *Usability Engineering,* Palgrave, Hampshire.
3. Gulliksen, J., Boivie, I, Persson, J., Hektor, A. and Herulf, L. (2004): *Making a difference – a survey of the usability profession in Sweden,* Proceedings from NordiCHI '04, October 23-27, 2004 Tampere, Finland.
4. ISO/IEC (1998) 9241-11*, Ergonomic requirements for office work with visual display terminals, (VDT)s – Part 11, Guideance on usability,* ISO/IEC 9241-11: 1998.
5. Macloeod, M., Bowden, R. and Bevan, N: *The MUSiC Performance Measurement Method,* Behavior and Information Technology, 16, 1997.
6. Molich, Rolf (2003): *Brugervenligt webdesign,* Teknisk Forlag, Copenhagen.
7. Nielsen, J. (1993), *Usability Engineering*, Academic Press, Inc. San Diego.
8. Nielsen, J. (2001), *Success Rate: The Simplest Usability Metric,* Alertbox from February 18, 2001. (http://www.useit.com/alertbox/20010218.html).
9. Preece, J., Rogers, Y. and Sharp, H. (2002), *Interaction design,* John Wiley & Sons, Inc. New York.
10. Rosenbaum, S., Rohn, J. A. and Humburg, J (2000): *A Toolkit for Strategic Usability: Results from Workshops, Panels and Surveys,* Proceedings from CHI 2000, The Hague, Amsterdam.