Usability metrics of time and stress -
Biological enhanced performance test of a university wide learning management system

[1] Christian Stickel, [2] Alexei Scerbakov, [3] Thomas Kaufmann, [4] Martin Ebner

Zentraler Informatik Dienst / Vernetztes Lernen
Graz University of Technology
Steyrergasse 30/I,
A-8010 Graz
[1] stickel@tugraz.at, [2] alexei@sbox.tugraz.at, [3] kaufma@sbox.tugraz.at, [4] martin.ebner@tugraz.at

Abstract: This paper describes the modification and outcome of a performance test applied to a university wide learning management system under realistic conditions to identify usability problems and to compare measures such as success rate, task time and user satisfaction with requirements. Two user groups with 20 test users each took part in this study. During the whole test psycho-physiological parameters of the test persons were monitored and recorded, in order to find event related stress symptoms. Modifications of the original test allowed a faster analysis of relevant quantitative metrics and the collection of qualitative information.

Keywords: Usability Test, Performance Measurement, Self-Assessment, EEG

1 Introduction

In a world where information systems are designed to support particular transactions and workflows, a primary concern should be the effect on the efficiency of the work they support. This paper describes a modified performance test applied to a university wide learning management system developed at the IICM of Graz University of Technology (Ebner & Walder 2007). Actually this system hosts about 200 lectures and serves 10000+ users, thus the requirement for efficient workflows and productivity is given, especially as the system has to support short and longtime routines, as well as enhance learning and teaching behaviors. Targets are only meaningful when they can be expressed quantitatively, so this study measured the efficiency, respectively productivity of existing workflows and practices. The applied performance test is mostly based on the NPL Performance Measurement Method proposed by Rengger et al. whereby the modification was a small two-question survey after every task. The test persons gave thereby statements on the subjective

difficulty of the tasks and their emotional state. Surprisingly a correlation was found between the hard to extract performance data and the easy to analyze survey data, which led to some ideas for economic web based performance tests. The approach included further the combination with a light Thinking Aloud Method and additional psycho physiological measures in order to research in the enhancement of UE methods with biological data. This enabled the collection of quantitative performance data on the one hand and qualitative feedback on the other hand, thereby balancing the advantages and disadvantages of both methods.

2 Motivation

The primary intention for conducting this study concerns the improvement of the tested software. TeachCenter is a learning management system that combines course management, digital content distribution and interactivity between students and teachers. The software is actually running a huge amount of lectures and is subject of continuous improvement and trend setting features. Because it is a daily used system with 10000+ users, lecturers and students it seems obviously that improvement of usability is of highest interest. Graz University of Technology had a high expertise in usability testing and learning objects (Holzinger & Ebner, 2003; Holzinger, 2004; Holzinger et al., 2005).

3 Methods and Design

The NPL Performance Measurement Method was used, in order to derive the effectiveness and efficiency of some main tasks, respectively workflows. Further it was of interest, which other issues the persons would find and what would be the underlying reasons. Therefore a combination of two Methods, the NPL Performance Test and the Thinking Aloud Test was designed, in order to provide performance data on the one hand and deeper insight on user interaction processes on the other hand. Further motivation came from the hypothesis, that the system has a good learnability. Therefore a biological rapid usability approach (Stickel, Holzinger & Fink, 2007) was applied whereby the psycho physiological parameters EEG, SCL and HR were recorded. As cross check to the above-mentioned goals, the original performance test has been further enhanced by a user self-assessment of the difficulty and the arousal after each task.

3.1 NPL Performance Measurement Method

The NPL Performance Measurement Method (Rengger et al., 1993) focuses on the quality and degree of work goal achievement. It is a rigorous usability evaluation of a working system under realistic conditions to identify major usability problems and areas. The test person fulfills tasks, whereby the time and video are recorded. As this method depends on realistic conditions, participants are not allowed to talk with the

facilitator, instead they are asked to accomplish the tasks as fast as possible. Measures of core indicators of usability can be obtained, as defined in ISO 9241-11 (ISO, 1998) e.g. user effectiveness, efficiency and satisfaction. It's then possible to compare these measures with requirements. These measures are directly related to productivity and business goals. In this study the metrics Task Effectiveness, User Efficiency, Relative User Efficiency and User Satisfaction were derived. *Task Effectiveness* (TES) determines how correctly and completely the goals have been achieved in the context of the task. In most cases there's more than one way to accomplish a task and every task has several steps, as it's not meaningful to test single click actions - instead the use of the systems main functions is compiled in a task. TES is a function of quantity and quality of the task. Quantity is measured objectively as the percentage of the control parameters, which have been altered from their default values by the end of the task. Quality consists of the definition of an optimal path, with weighted alternatives and penalty actions (e.g. help or explorative search). Quantity and Quality are measured as percentage values, so the resulting TES is also a percentage value. The value of TES is obtained by measuring quantity and quality and application of the formula TES = 1/100 (Quantity x Quality). *User Efficiency* (UE) relates effectiveness to costs in terms of time, e.g. if a task can be completed in a high quality AND fast, then the efficiency is high. UE provides here the absolute measure for the comparison of the five tasks of this study, carried out by the same users, on the same product in the same environment. It is calculated as the ratio between the effectiveness in carrying out the task and the time it takes to complete the task using UE = Task Effectiveness / Task Time. The *Relative User Efficiency* (RUE) is a metric that can be employed by the relation of a particular group of users compared to fully trained and experienced user of the product being tested. It is defined as the ratio of the efficiency of any user and the efficiency of an expert user in the same context RUE = (User TES / Expert TES) * (Expert Task Time / User Task Time) *100. The User satisfaction is derived with a standardized questionnaire like SUMI or SUS.

3.2 User Self Assessment web questionnaires

One question of this study was how the subjective user assessment of single tasks compares to the objective performance measurement. Therefore a web interface was developed, which was present for the participants on a second screen. At the beginning of every task a "Start" button had to be pressed and when the task was finished a "Stop" button. In this way the duration of the tasks were recorded. After pressing "Stop" the user had to rate the difficulty and his arousal, each on a five point scale. After rating the "Start" button appeared again, ready for the next task. All results were logged on a web server, with a unique id for every participant.

3.3 Thinking Aloud Method

The Thinking Aloud (TA) method reveals hidden thoughts and gives an insight in the users mental model. It helps understanding how the user wants to use the system and what kind of features might be optimized. Therefore the user is asked to verbalize

all thoughts and actions during the test. The test is usually designed of different tasks, which represent the major application and functionality of a system. The whole procedure is recorded with a video camera and the session is transcribed afterwards. The generated protocol can be analyzed in order to reveal information about the users reasoning sequences and goal structures. This study used the thinking aloud technique in a block right after the main tasks. Thereby questions were asked that aimed to determine how the system works. All of the so-called "Minitasks" could be easily answered by analog transfer from the recently used functions. The intention of this approach was the compensation of the main disadvantage of the NPL Performance Measurement Method, which will not reveal the reason for problems.

3.4 Psycho-Physiological Measurement Methods

Biological measures of emotional states have been used by several researchers in Human-Computer Interaction (Picard, 1997), (Picard, 2000), (Riseberg et al., 1998), (Murgg & Nischelwitzer, 2004). Muter et al. (1993) found that psycho-physiological measures can be regarded useful for Usability; especially the Skin Conductance Level (SCL) seems to be a good indicator for the overall usability of software, as they found a correlation between user-hostile systems and an increase of SCL. In this study a combination of Electroencephalogram (EEG), SCL and the Heart Rate (HR) was applied.

4   Design of the study

This chapter covers the design of the study, thus the schematic design, the user profiling and recruitment, the setup and the standardized procedure are described.

4.1 Schematic Design

Every trial of the test was split in three main parts, the control condition (K1), the major tasks performance test (L1) and the mini tasks TA test (L2). Figure 1 on the next page shows the schematic design of the test. In the beginning all test persons were asked to fill out a profile and at the end of the test a feedback form for user satisfaction. For user profiling the scheme from the Performance Measurement Handbook was used. The German version of the System Usability Scale (SUS) questionnaire from Brooke (1996) was used to derive User Satisfaction afterwards. During the blocks K1, L1 and L2 the psycho physiological parameters EEG, SCL and HR were recorded. In L1 and L2 additional videos and screen recordings have been done.

4.2 Test procedure

In the control condition the test persons were asked to relax in order to get some basic biodata. The relaxation process was supported by a Brainlight system (http://www.brainlight.com). The second block was the actual performance test. First the users were given a sheet of paper with the task. When they understood the task they pushed the "Start" button, otherwise they were allowed to ask the moderator comprehension questions. Each task had a goal and sub goals. When either the user thought that the complete task had been accomplished, or the moderator finished the task due to completion or time out, the user had to push the "Stop" button and was instantly asked in the web interface, to rate the difficulty of the tasks and his state of arousal. This self-assessment took place after each major task in the L1 block.
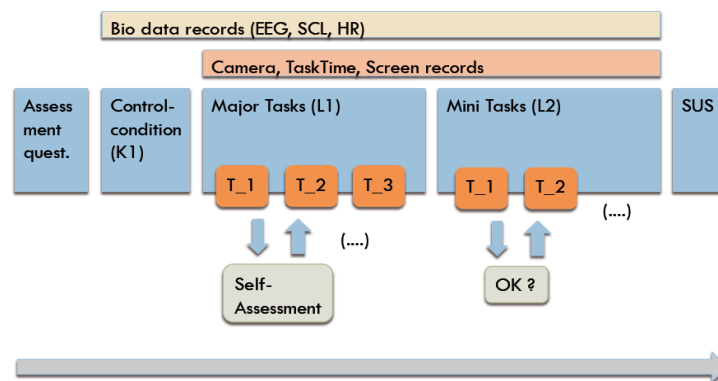

Fig. 1 Schematic design of the test

In the "Mini Tasks" L2 block, users were asked questions and had 30 seconds each to accomplish the task respectively answering the question and showing how to do. According to the TA Method the users were thereby asked to verbalize their thoughts and actions. As last step User Satisfaction was derived with the System Usability Scale (SUS), a standardized psychometric questionnaire for a high-level subjective assessment of Usability. The outcome is a rate on a scale of 0-100. Benefit and drawback is the general nature of the questions, however this allows the comparison of diverse systems (Brooke, 1996) and different parts of the system.

4.3 User Profiles

The profiling as described in the NPL PM handbook (Rengger et al., 1993) was applied. Thereby two major user groups were derived, which are students and lecturers. Target for both groups was getting novices, who had no or just little experience using the learning management system. The students had an average experience of 2 month, while the teachers had an average experience of 17 month, however the variance inside the teachers group concerning this variable was high. Beside this the profile contained several usability relevant issues on skills, trainings,

mental-, physical- and job attributes. For the two user groups profile questionnaires were generated, which were then completed by every participant.

4.4 Hardware and Software setup

Figure 2 shows the setup for the students group (left) and the teachers group (right). The test user (T) is sitting in front of two screens (M1, M2), on his left side is the moderator (M) Right beside the screens a mirror was positioned in order to capture facial expressions and the actions on the main screen (M1). The scribe (S) was operating the camera (C), Laptop L2 and taking notes. The operator (O) took care of the psycho-physiological recordings on laptop L1 and took notes too. The user test environment was a Shuttle PC with Windows XP, standard mouse and keyboard. As the learning management system is an online product Internet Explorer 6 was used as browser. Techsmith Camtasia Studio (http://ww.techsmith.com/camtasia.asp) was used for the screen recordings. A further laptop (L2) with the same software configuration was used to simulate online users. Video recordings were done with a HD Cam on a tripod right behind the user, capturing his actions on the screen, as well as the facial expressions in a mirror beside the screen.
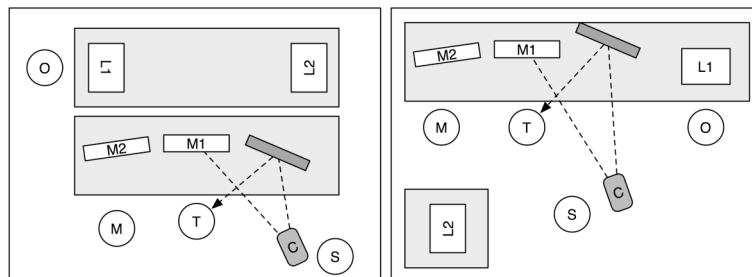

Fig. 2 Test settings for students (left) and teachers (right)

A Brainlight system (http://www.brainlight.com) was used as stimulation unit to induce relaxation by Steady State Visual Evoked Potentials (SSVEP) in a frequency range, changing between 8 and 12 Hz for 10 minutes. This system is a standalone product, which was programmed separately. It's based on SSVEPs, whereby a flickering light source elicits potentials of the same frequency in the brain, while the subject shifts the gaze to these stimuli (Müller-Putz et al, 2005). It works visually on the same base as the auditory frequency following-response, which states that most periodic complex sounds evoke low pitches associated with their fundamental frequency, also called periodicity pitch (deBoer, 1976; Evans, 1978; Moore, 1989). As different frequencies are linked to different mental respectively physiological states, relaxation can be induced using according frequencies. The EEG recordings were done with an IBVA 3 electrode headband EEG from Psychiclabs Inc. (http://www.psychiclabs.net) at a rate of 512 Hz. The equipment can be used in laboratory and field settings as well. As IBVA's headband uses only 3 electrodes, it can record an EEG of the frontal lobe only. A Lightstone from the wild divine project (http://www.wilddivine.com) was used for SCL and HR recordings. It acquires the

data non-invasive by sensors, which are put on the fingertips. Unfortunately this might interfere with tasks that depend on extensive keyboard input.

5 Results & Discussion

The metrics Task Effectiveness (TES), User Efficiency (UE) and Relative User Efficiency (RUE) were derived for both groups to identify the problematic tasks. TES determines how correctly and completely the goals have been achieved, while UE relates effectiveness to costs in terms of time. The Relative User Efficiency (RUE) is the ratio between the efficiency of a user and an expert. First the RUE was calculated for every user, and then an average for all users was calculated per task. Most participants were new to the system, so the driving question for deriving this metric was the gap between experts and novices. TES, UE and RUE metrics are measured on a percentage scale, with 1 for the lowest and 100 for the highest value.
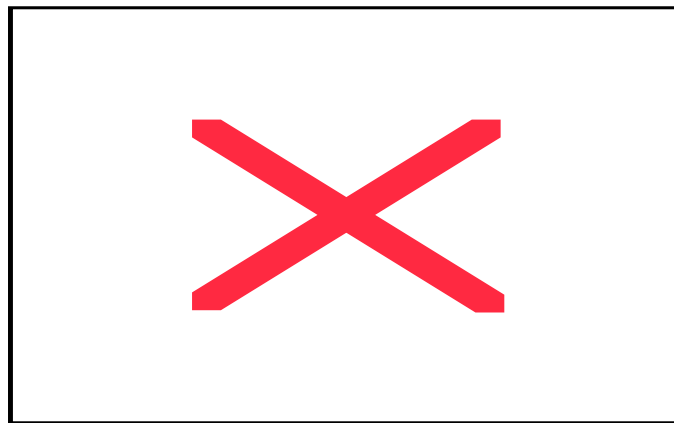

Fig. 3 TES, UE and RUE of the student trial

Figure 3 shows TES, UE and RUE for each task of the student trial. Tasks with low values determine usability problems. The results of the student trial show extremely decreased TES and UE values in the second task, which leads to the conclusion that the students had the most problems with the second task. The goal of this task was gathering a set of information that was spread throughout the whole course, ranging from download files to discussion board threads. The use of the search feature was mandatory here in order to find all information concerning the specific topic. However, most test persons used explorative search, found one or two information and thought the task being finished. Actually the performance of the search window should have been tested; instead it was surprising to find that the search feature was hardly used.

This led to the conclusion to make the search more visible in terms of generalizabilty and place a search field with an according button in the page header. Figure 3 shows also a high RUE value for Task 2, which should usually be low. This is because the expert user wasn't able to solve the task, while some of the novice user accomplished the task.
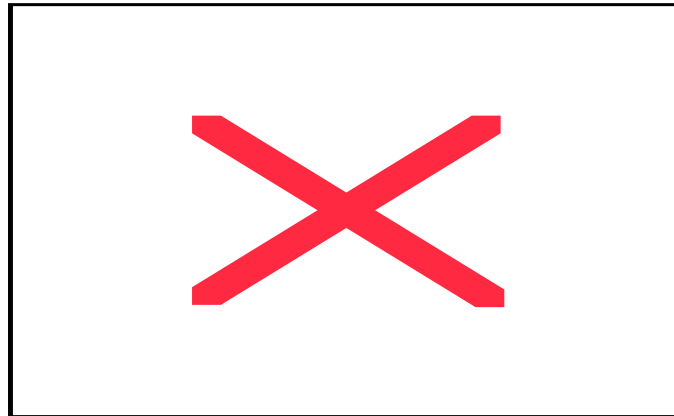

Fig. 4 TES, UE and RUE of the teacher trial

Figure 4 shows TES, UE and RUE for each task of the teacher trial. It can be seen that all values decreased in the fourth task so obviously this task contained an issue. Task 4 concerned the import of a description. The bottleneck in the workflow was the import button, which is just an icon with an upward arrow and not obvious to find. The fact that the screen recording software averted the tooltips made this task even harder. The second important issue in this case was a choice, where it was not obvious that the system expected further input. Overriding this choice cancelled the whole import process. Two changes were proposed in order to solve these issues. First the replacement of the import icon by an "IMPORT" button; second a clear message from the system that it expects another choice of the user.

Figure 4 shows also the strong decrease of RUE in Task 4 of the teacher trial as expected. This is because the expert user was able to solve the task much faster and with a higher quality. All other tasks show high values for RUE, even though they are averaged, this can be interpreted positive, because it mans that the average teacher user with one-year experience is able to accomplish main tasks with 80-90% efficiency of an expert user. Overall can be stated that the system has a very good learnablity. Simple functions are intuitive, while advanced functionality requires more training.

From Figure 3 and Figure 4 the question arose, why the Task Effectiveness was always higher than the User Efficiency. The results can be interpreted that in most cases the users of both groups were able to solve the tasks, however in a lower quality and taking much more time as expected and possible.
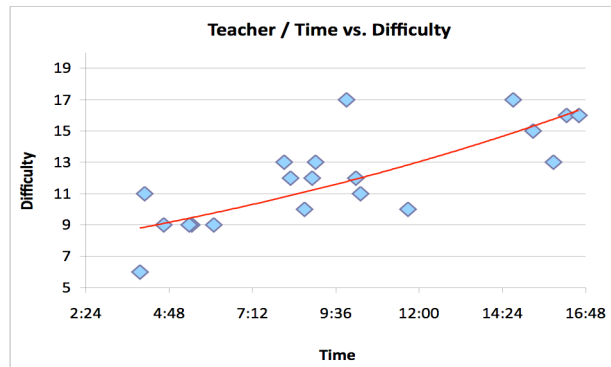
Fig. 5. Teacher assessed the difficulty according to the task time

Figure 5 shows that there's a positive correlation between task time and self assessed task difficulty, within the teachers group. The longer the tasks took the more difficult they were rated. Within the students group this can also be pointed out, but not significant. We suppose that it didn't occur within the students group, because the students have been faster, due to simpler tasks in the front end.
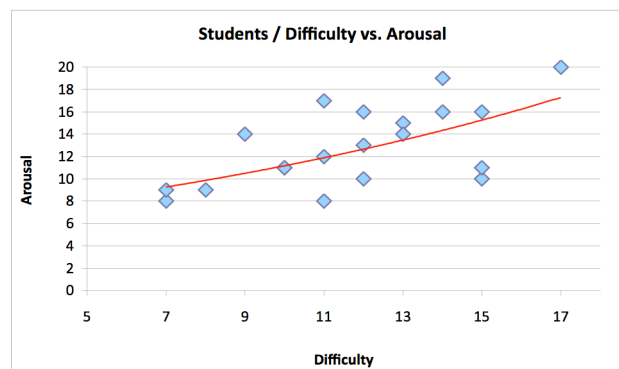

Fig. 6. Students were more exited on higher difficulty

Figure 6 shows a positive correlation between the self-assessed task arousal and the task difficulty was found in the students group. The more difficult the tasks were rated the higher was the arousal. This was not found in the teachers group, although the distribution of the difficulty was similar. The positive correlation between difficulty and arousal in the students group could be interpreted as stress. We suppose students are more sensitive to test related stress than teachers.
One of the most important questions concerning the modification of the test was if there is a negative correlation between the subjective task difficulty assessment and the performance data. We hypothesized that the task difficulty should be high for low task performance. In order to visualize this, average performance and task difficulty

data were normalized. Additionally the difficulty was inverted to show a positive correlating curve, which can be seen in figure 7 for the students group and in figure 8 for the teachers group. The curves for TES/UE and the subjective inversed difficulty (SID) correlate, so does also the curve for the subjective Arousal assessment per task, which was spared in the charts.
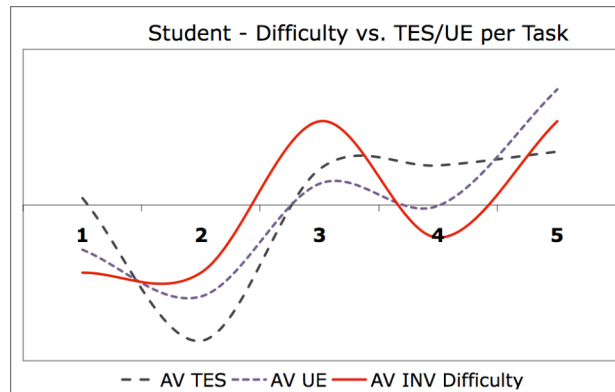

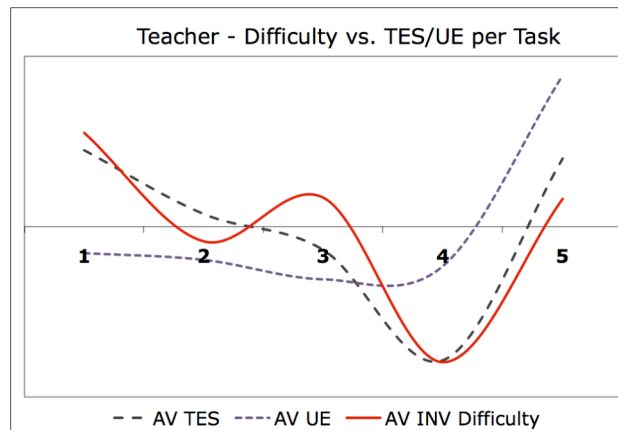Fig. 7 Students SID vs. TES/UE per task

Fig. 8 Teacher SID vs. TES/UE per task

It is important to note that, as these curves correlate, they provide the same information on the usability of the system, although obtained in a different way. TES/UE values have to be generated by a standardized extensive procedure, while the SID is just the result of a simple question after fulfillment of a task. This leads to interesting web testing paradigms, which may provide similar conclusions as extensive testing procedures in a faster and more economic way.

The System Usability Scale (SUS) rating for the system was equal for both groups (Students: 70, Teacher: 66), although they had different tasks, according to their role.

The average difficulty of the self-assessment after each of the main tasks was similar (Students: 11.8, Teacher: 12 on a 25 point scale, whereby 25 is the highest difficulty). The SUS rating of the system can be considered as "usable". As there were two different groups with different tasks, it provides the feedback that most parts of system are satisfying and usable. Figure 9 shows a Gaussian distribution for user satisfaction is around 70 - 80 %.
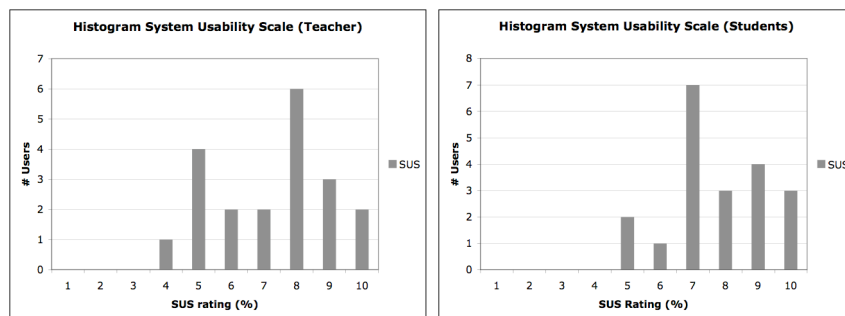

Fig. 9 Histograms of the SUS questionnaire for both groups

5.1 Further observations

SCL peaks occurred when the user was confronted with problems, e.g. problems with uploading, login problems, explorative problems (can't find). Further peaks were noted when the users read the task and tried to understand what they had to do. With some users there were also peaks for every question in the Mini Task block. The SCL during working condition in L1 and L2 was relatively doubled compared to the SCL recorded in the relaxation condition K1.

5.2 Conclusion

The modifications of the original NPL Performance Measurement Method provided the performance metrics and additional qualitative data on reasons of problems. The self-assessment after each task gave insight into task specific user perception of difficulty and arousal. Furthermore, it was shown that the normalized performance and subjective inversed difficulty data correlate positive. So far we suppose that a simplified, more economic version of the NPL PM Method, can also provide data on the performance of a system. A further interesting paradigm can be the application of this procedure as automated web survey. As the subjective arousal assessment also correlated with the performance data, it will be interesting to analyze how this data relates to the recorded biological data and reveal further objectives and approaches to biological rapid usability testing.

6 References

1. Brooke, J.: SUS: a "quick and dirty" usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester & A. L. McClelland (eds.) Usability Evaluation in Industry. Taylor and Francis, London (1996)
2. deBoer, E.: On the residue and auditory pitch perception. In: Keidel, W.D., Ne¡, W.D. (Eds.), Handbook of Sensors Physiology. Springer-Verlag, Berlin, (1976) 479-583
3. Ebner, M, Walder, U.: e-Learning in Civil Engineering – Six Years of Experience at Graz University of Technology. In Bringing ITC Knowledge to work, Proceeding of 24th W78 Conference Maribor 2007 & 14th EG-ICE Workshop & 5th ITC@EDU Workshop, Danijel Rebolj (ed.), (2007) 749 - 754
4. Evans, E.F.: Place and time coding of frequency in the peripheral auditory system:some physiological pros and cons. Audiology 17, (1978) 369-420
5. ISO The international Organization for Standardization : Ergonomic requirements for office work with visual display terminals (VDTs). Part 11: Guidance on usability (ISO 9241--11) (1998)
6. Holzinger, A. & Ebner, M.: Interaction and Usability of Simulations & Animations: A case study of the Flash Technology. Proceedings of: Interact 2003, Zurich, (2003) 777-780.
7. Holzinger, A.: Usability Engineering for Software Developers. Communications of the ACM, 48(1) (2005) 71-74
8. Holzinger, A., Kickmeier-Rust, M., Albert, D.: Dynamic Media in Computer Science Education; Content Complexity and Learning Performance: Is Less More?. Educational Technology & Society, 11(1) (2008) 279-290
9. Holzinger, A., Nischelwitzer, A. & Meisenberger, M.: Mobile Phones as a Challenge for m-Learning: Examples for Mobile Interactive Learning Objects (MILOs). Proceedings of: Third IEEE International Conference on Pervasive Computing and Communication (PerCom 05), Kauai Island (HI), (2005) 307-311
10. Macleod, M.: Draft of chapter to appear in P Jordan (ed.) 'Usability Evaluation in Industry'. London, Taylor and Francis. Crown publishing (1994)
11. Moore, B.C.J.: Introduction to the Psychology of Hearing, 3rd edn. Academic Press, London (1989)
12. Müller-Putz, G. R., Scherer, R., Brauneis, C. and Pfurtscheller G.: Steady-state visual evoked potential (SSVEP)- based communication: impact of harmonic frequency components, Journal of Neural Engineering, vol. 2, no. 4, (2005) 123-130
13. Picard, R. W.: Affective Computing. MIT Press, Cambridge (MA) (1997)
14. Picard, R. W. & Healey J.: Affective Wearables. MIT Press, Cambridge (MA) (1997)
15. Picard, R. W.: Perceptual user interfaces: affective perception. Communications of the ACM, 43, 3, (2000) 50-51
16. Rengger, R., Macleod, M., Bowden, R., Drynan, A. and Blayney, M.: MUSiC Performance Measurement Handbook, V2. NPL, DITC, Teddington, UK (1993)
17. Riseberg, J., Klein, J., Fernandez, R. & Picard, R. W.: Frustrating the user on purpose: using biosignals in a pilot study to detect the user's emotional state. Conference on Human Factors in Computing Systems, Los Angeles (CA), (1998) 227-228
18. Murgg E., Nischelwitzer A.: *Physiological Usability Testing: A Biological Approach to Detect and Measure Usability Problems*, Multimedia Applications in Education Conference (MApEC) Proceedings 2004, (2004) 122-127
19. Stickel, C., Fink, J., Holzinger, A.: *Enhancing Universal Access – EEG based Learnability Assessment,* In: Universal Access to Applications and Services. Lecture Notes in Computer Science (LNCS 4556). Berlin, Heidelberg, New York, Springer, (2007) 813–822