Seeing the System through the End Users' Eyes: Shadow Expert Technique for Evaluating the Consistency of a Learning Management System

Andreas Holzinger[1], Christian Stickel[2], Markus Fassold[2], and Martin Ebner[2]

[1] Medical University Graz, A-8036 Graz, Austria
Institute for Medical Informatics (IMI), Research Unit HCI4MED &
Graz University of Technology, A-8010 Graz, Austria
Institute for Information Systems and Computer Media (IICM)
a.holzinger@tugraz.at
[2] Graz University of Technology, A-8010 Graz, CIS/Department of Social Media
{stickel,martin.ebner}@tugraz.at

Abstract. Interface consistency is an important basic concept in web design and has an effect on performance and satisfaction of end users. Consistency also has significant effects on the learning performance of both expert and novice end users. Consequently, the evaluation of consistency within a e-learning system and the ensuing eradication of irritating discrepancies in the user interface redesign is a big issue. In this paper, we report of our experiences with the Shadow Expert Technique (SET) during the evaluation of the consistency of the user interface of a large university learning management system. The main objective of this new usability evaluation method is to understand the interaction processes of end users with a specific system interface. Two teams of usability experts worked independently from each other in order to maximize the objectivity of the results. The outcome of this SET method is a list of recommended changes to improve the user interaction processes, hence to facilitate high consistency.

Keywords: Consistency, Shadow Expert Technique, Usability Test, Methods, Performance, Measurement.

1 Introduction and Motivation for Research

It is generally known that the acceptance and usability of software is significantly related to its user interface quality [1] and the first of Ben Shneidermans' Golden Rule of Interface Design is "Strive for Consistency" [2]. Consistent sequences of actions are required in similar situations; identical terminology should be used; consistent color, layout, fonts, etc. should be used throughout the complete system. This not only enhances accessibility, but it is of benefit to every user.

Consequently, for a long time the most important design guideline has been: Build consistent human interfaces [3], yet often this guideline is ignored by designers [4]. Moreover, previous studies on Learning Management Systems (LMS, e-Learning systems) show that the user interface consistency has significant effects on the learning

A. Holzinger and K. Miesenberger (Eds.): USAB 2009, LNCS 5889, pp. 178–192, 2009.
© Springer-Verlag Berlin Heidelberg 2009

performance. Experienced end users make more errors than novices and their satisfaction level is lower when using a physically inconsistent user interface, whereas conceptually consistent user interfaces facilitate performance and satisfaction [5].

These facts have been the motivation for the work described here: At Graz University of Technology approximately 10,000 students and teachers work daily with the TeachCenter software, which is a large LMS, consisting of a number of applications, i.e. Classroom Pages, Chat rooms, Forums, Blogs, Polls, FAQ's, Plagiarism Tests etc. All these functions facilitate the daily teaching and learning of the students and teachers. However, the efficient use of those functions depends highly on the acceptability of the system and in previous usability studies it was found that the design of the system is lacking consistency. In order to carefully identify these inconsistencies, we decided to apply a new usability testing method called the Shadow Expert Technique (SET), which is basically a mixture of existing usability test methods.

2 Theoretical Background and Related Work

This section provides a short overview about the basic concept of consistency and describes two usability testing methods: Focused Heuristic Evaluation and the NPL Performance Measurement, which have been in use individually prior to their inclusion in the Shadow Expert Technique.

2.1 The Three-Dimensional Model of User Interface Consistency

Interface consistency has been studied for quite a long time, actually since Graphical User Interfaces (GUIs) began to be used widely, under the premise that a worker who is able to predict what the system will do in any given situation and can rely on the rules will be more efficient [6]. Consequently, the focus of research was on worker's productivity in order to achieve higher throughput and fewer errors. As a result of this goal, most early studies were on job performance of office workers, i.e. error rate and time to perform a task. The latter is the typical Human–Computer Interaction (HCI) approach and is usually considered in a transfer paradigm in which: *the higher the similarity between two tasks, the higher the transfer, hence the consistency* [7].

However, a strict establishment of the primary places of where consistency is most necessary, is difficult. Grudin (1989) [4] separated consistency into *internal* interface consistency and *external* interface consistency, wherein internal refers to consistency within a task and external means consistency among various tasks. Ozok & Salvendy (2000) [8] classified it into three sub types, establishing the *three-dimensional model of interface consistency:*

1) conceptual consistency (language, stereotypes, task concept, skill transfer, output consistency, hierarchical order of concept, etc.);
2) communicational consistency (moving between screens, menus, user conventions, between-task consistency, distinction of tasks and objects, etc.); and
3) physical consistency (color, size, shape, location, spacing, symbols, etc.).

Ad 1) *Conceptual consistency* can be defined as the consistency of metaphor applied to an interface feature or an action that is embodied within a feature. Frequent and

inconsistent use of synonyms, instead of using the same words for the same items, is unhelpful. Leaving something to students' conception and interpretation due to lack of explicitness is also regarded as conceptual inconsistency [4], [8].

Ad 2) *Communicational consistency* can be defined as the consistency of both input and output of the interface. It deals with how the user interacts with the computer interface and whether the means of interaction are consistent for fulfilling the same or similar tasks.

Ad 3) *Physical consistency* can be defined as the consistency of the visual appearance of an interface feature and indicates that the features are supposed to be consistent with the users mental models [9].

Although this has been known for quite a long time, research on the relationship between *consistency and human learning processes* has only recently been documented, and Satzinger & Olfman (1998) [10] pointed out that very few studies have investigated the effects of interface consistency on learning performance. Many questions still remain as to the effectiveness of interface design in enhancing learning and since designers are currently able to include a wide range of various visual elements, the complexity of design decisions is steadily increasing and most design decisions are not made on a scientific basis. To design an appropriate user interface demands insight into the *behaviour of the end users* and the application of user centered development [11], [12], [13], in order to achieve a true interaction. This is definitely important, since learning with interactive media is generally highly demanding from the perspective of the limited cognitive processing capabilities of the end users [14], [15]. Daily practice shows that many end users have difficulty learning with electronic systems, since they are often unable to form a mental model of the system and their current position within its complexity. However, when striving for a design following the "principle of the least surprise", we are faced with the problem that designers and developers rarely are able to predict exactly what the end users really expect (remember Steve Krug [16]: "Don't make me think!").

2.2 Focused Heuristic Evaluation

A Heuristic evaluation (HE) is an *inspection* method (for a overview on methods see [17]) to improve the usability of an interface by checking it against established standards [18], [19] [20]. The basic idea is to have usability experts examine the interface of a system according to a list of pre-determined heuristics. Each usability expert has to perform the evaluation with complete independence from each other.

They do not share their results or findings until they are all finished with their respective evaluations. The purpose of this independence is to assure an objective and unbiased evaluation. During an evaluation process each evaluator goes through the entire interface and checks if the interface complies with the pre-determined list of heuristics. The number of necessary evaluations is still debated, however, Nielsen (1992) [21] recommends to use from three to five evaluators. So far the outcome depends on the used list of heuristics. Usually these heuristics cover diverse or general design issues and system behavior. In the case of a **Focused** Heuristic Evaluation, one single issue (here consistency) is chosen and a list of appropriate heuristics is generated and used for evaluation [22].

2.3 NPL Performance Measurement Method

A performance test is a rigorous usability evaluation of a working system [5], [23]. In order for the performance test to gather realistic and precise information of the working system, the performance test must take place under realistic conditions (ideally in the real context, see figure 1) with real end users. Usually 12 to 20 test persons are sufficient to get reliable data [24]. Each trial of the test is video recorded and screen captures in order to provide a good documentation for the gathered data. A Performance Test identifies issues that influence the performance of an end user, including time, effectiveness and user efficiency [25]. This enables the comparison of different designs or design steps in an iterative development circle. Usually the Performance Test also addresses user satisfaction with the current system.

The NPL (National Physical Laboratory) Performance Measurement Method focuses on the quality and degree of work goal achievement in terms of task performance achievement of frequent and critical task goals by end users in a context simulating the work environment [26], [27], [28], [29], [30]. User performance is specified and assessed by measures including task effectiveness (the quantity and quality of task performance) and User efficiency (effectiveness divided by task time). Measures are obtained with users performing tasks in a context of evaluation which matches the intended context of use (figure 1) [29], [31]. Casually speaking, the test person

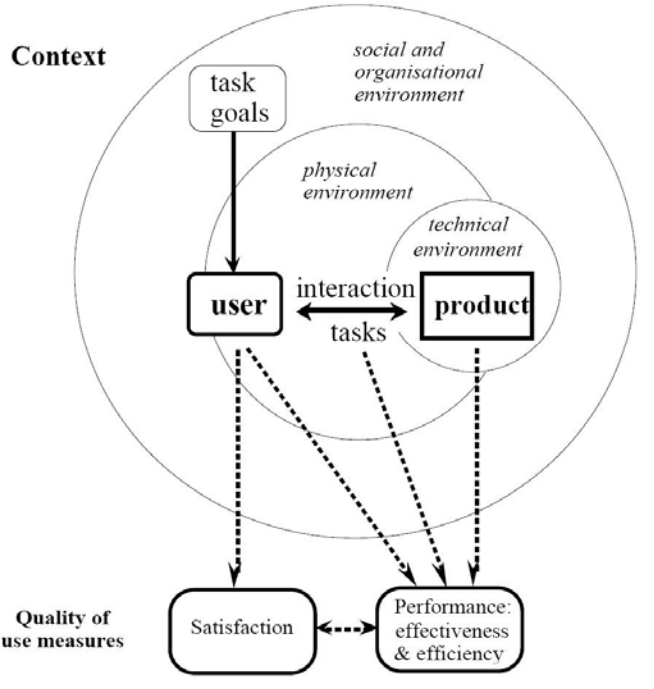

Fig. 1. According to Nigel Bevan (1995) [35] Performance consists of Effectiveness and Efficiency and is directly related with satisfaction

performs tasks, whereby the time is measured and a video is taken. Test persons are not allowed to talk with the facilitator, instead they are instructed to accomplish the tasks as fast as possible. Measures of core indicators of usability can be obtained, as defined in ISO 9241-11 [26], [32], [33], [34] e.g. user effectiveness, efficiency and satisfaction. It is possible to compare these measures with the 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, for example the SUMI (Software Usability Measurement Inventory) [36] or SUS (Software Usability Scale) [37].

3 Methods and Materials

3.1 Shadow Expert Technique Flow

The first step (see flowchart figure 2) was to analyze the interface of the system in the role as consistency experts in order to find as much inconsistencies as possible. The Focused Heuristic Evaluation (FHE) was applied - the focus being on consistency. The result of the FHE was a list of general inconsistency issues, ordered by severity.

Considering the most severe inconsistency issues, as well as the LMS main features, tasks for the performance test were gathered. Two independent teams derived these tasks. The reason for this independency is to obtain more objective results from the evaluators later on. Both teams conducted the performance test with seven different subjects each. Subsequently, they analyzed the tests to determine a subject with average results. For the SET an end user with average results is needed in order to

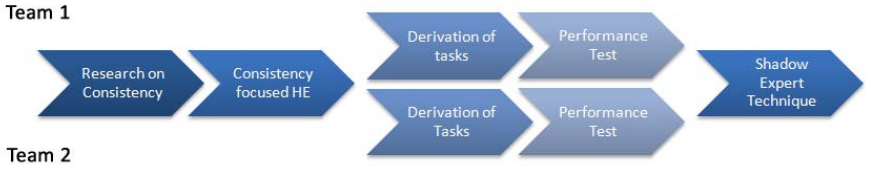

Fig. 2. The way from the research to the Shadow Expert Technique

provide a representative end user behavior. Once this was achieved the two teams exchanged the video and screen records of this subject and carried out the SET.

3.2 Focus on Consistency

Five evaluators examined the interface of the LMS based on a defined heuristic list. The output of this process was a list of issues that violated the consistency heuristics. The following section will outline the different relevant viewpoints on consistency, which lead to a set of applicable heuristics. A short table form of the heuristics can be found at the end of this section (see figure 3).

3.2.1 Principle of Least Surprise (POLS)

A golden rule in interface design is the Principle of Least Surprise (POLS). According to Geoffrey James "A user interface should be arranged in a way that the user experiences as few surprises as possible". It is a psychological fact, that humans can pay considerable attention only to one thing at one time (locus of attention) [38]. In many cases this is difficult for the software engineers, since they usually think in functional terms of the system and have difficulty putting himself in the end user's position (often the end users are not known). Moreover, different users have varying expectations.

3.2.2 Expectation Conformity (Playful Consistency)

This means that already learned techniques can likewise be applied anywhere in an application. One could also say that 'everything' must be conclusive (see DIN ISO 9241-10 Ergonomic demands for office activities with screen devices – part 10: principles of dialogue formation, expectation conformity) [39]. Accordingly a dialogue is conform to expectations if it is consistent and matches the characteristics of the user (user mental models, previous knowledge, expertise, education, literacy, experience etc.) as well as common acknowledged conventions. Operation cycles, symbols and the arrangement of information should be consistent within the application, match the gained knowledge of the user and should thus be conform to expectations. For example status declarations of the dialogue system are emitted at the same place or pressing the same button ends a dialogue. So, the dialogue behavior and information display are uniform within the dialogue system and thus consistent.

3.2.3 Browser Consistency

Concerning the UI design of websites and web applications, it is particularly the varied display in the different browsers, which plays an important role. The continuous uniform appearance as well as the same reaction to a given action (in all different

kinds of browsers) could be subsumed under the term browser consistency. These measures are applied to ensure that the user can handle the application or the site, no matter, which browser is used.

3.3 Materials for the Shadow Expert Technique

The output of the FHE was a list of issues that violated the consistency heuristics. From the list of issues we selected the most severe ones and used them to derive tasks for a performance test. This approach enables insight into the influence of these issues in a realistic context and get as much relevant results as possible. The subjects had to be also as realistic as possible, which is why we defined a test subject profile. As the learning platform is targeted at the use at university level, we set the profile for our test subjects according, to be students with some basic IT knowledge.

Consistency	Description	Evaluation
Expectation conformity (playful consistency)	I expect to be logged out if I click on "logout" – in the entire application, no matter in which context.	1 – 5 (fully met - not met)
Visual consistency	The design (typeface, colors, styles etc.) should be consistent within the application.	1 – 5 (fully met - not met)
Hyperlink consistency	Hyperlinks should not be a dead-end-street ("death links"). Furthermore, links should always behave the same way: this means, for example, that external links should always be opened in a new window, while internal links should always be opened in the same window.	1 – 5 (fully met - not met)
Linguistic consistency	If the application is available in several languages, there should be no mixture of languages.	1 – 5 (fully met - not met)
Browser consistency	The used browser should not limit the application. Furthermore, the browser should not significantly alter the appearance of the application (typefaces, colors, positioning of the different elements).	1 – 5 (fully met - not met)

Fig. 3. List of Consistency Heuristics

It is important, that the test users corresponded to the target group to avoid rough outliers in the analysis phase. Before deriving the tasks we divided our team into two groups in order to work completely independent from each other. This also meant that the two groups had a different set of tasks that were to be used on the performance test, and that both groups had absolutely no knowledge of each others tasks. Each group tested the system on seven subjects. The tasks were chosen so that the complete set of tasks would take the user no longer than 15 minutes. At the beginning of the test each subject was briefed on the process. They had a limited amount of time for each task, and did not receive any assistance whatsoever from our part. If a subject would exceed the time limit, they were told to move to the next task.

The user test environment was a Laptop with Windows XP, standard mouse and keyboard. As the learning management system is of course an online product, we used Internet Explorer 6 as browser.

All 14 trials of the performance test were recorded with webcam and screen recorder software. Techsmith Camtasia Studio [40] (http://ww.techsmith.com) was used. This tool synchronically recorded the computer screen during the test, as well as the test subjects upper body and face. Later on this allowed screening the subjects expressions with a Picture-in-Picture (PnP) view as can be seen in Fig. 2.

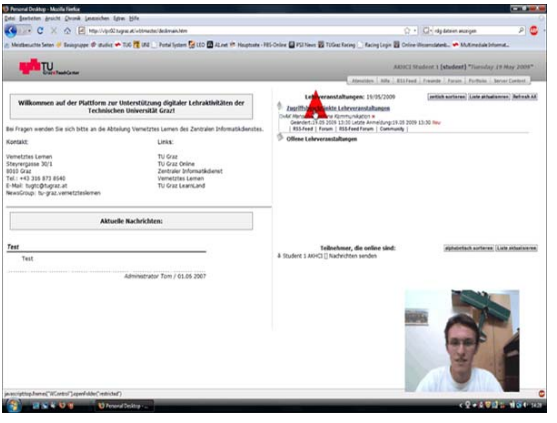

Fig. 4. Example screen recording with picture-in-picture

3.4 Methodology

The SET method has three phases: 1) preparation, 2) discussion and 3) analysis. SET is a multilevel-method in which two teams of experts try to understand the interaction processes of an end user, in order to derive suggestions for improvement. The primary goal of SET is to explore the users behavior by observation and discussion. Therefore synchronized screen recordings and video material of the user were used, without sound in order to recognize the end users inner state (e.g. frustration) from mimic, gestures and interactions with the systems. This enables the evaluator to see the system through the end users eyes and provides an in-depth understanding of occurring problems. Through iterative stop&go video review and discussion, this understanding grows to a point where simple and efficient solutions were gained during the discussion. The videos from the average users had a length of approximately 15 minutes. During the SET discussion session we produced four videos of the expert reviews and discussions (two reviews per group) that were transcribed in a final step. The length of the session videos varies, depending on the actual pass: The first round takes app. as long as the video from the average user (15 minutes), while the second round took in our case about 30 minutes. The entire SET discussion session lasted 4 hours including preparation time and debriefing.

3.4.1 SET Preparation Phase

In the preparation phase each expert-team must run a performance test as described in the section before. Both teams define the tasks for the performance test separately on base of a heuristic evaluation. The subjects should meet the requirements of the target user profile. Important is that screen records and videos of the subjects face are done during the performance test. We recommend using synchronized picture-in-picture screen recording.

These recordings will be exchanged between the groups later. After conducting the test both teams evaluate their results and try to figure out a user with average overall efficiency results. Using the data from an average user rather from a very good or very bad user is intended to provide more realistic data for the analysis than outliers and thus covers most problems that the biggest part of the target group would have. The average user was selected by the total user efficiency. For calculating the total user efficiency and selecting the data for the analysis we used the following formulas:

$$user\ efficiency\ task_x = \frac{effectivness}{t_{task}/t_{max}}$$

$$total\ user\ efficiency_x = \sum_{i=1}^{\sum tasks} user\ efficiency\ task_{x_i}$$

$$user\ efficiency\ intersection = \frac{\sum_{i=1}^{\sum user} efficiency_i}{number\ of\ users}$$

The *effectiveness* describes the success per task in % (e.g. Log in 30%, Find the chat and send a message via it 60%, Go back to the start screen 10% != 100%). If a user failed a partial task like "Go back to the start screen", he got a point deduction - in that case 10%.

The value t_{task} corresponds to the time the user needed for the task, and the value t_{max} corresponds to the maximum time for the task. The intersection task efficiencies results in the *total user efficiency* for every user. The *user efficiency intersection* calculates the average efficiency over all users. The subject with *total user efficiency* nearest to the *user efficiency intersection* was selected for SET screening. After these initiatory steps the discussion phase of the Shadow Expert Technique can be executed. For this purpose, both teams need to take a session together.

3.4.2 SET Discussion Phase

The technical requirements for the discussion phase are a laptop for replaying the recordings, a beamer for better review of the expert team and a video camera for recording the SET discussion session, thus collecting all discussed problems and solutions. Additional, we recommend a screen recording of the session on the laptop, in order to be able to understand what was spoken about, when the SET session is transcribed and analyzed later on. The roles of the discussion team are moderator, 2-4 experts and a scribe. These roles and tasks of the participants change during the discussion session.

In the first step, the teams exchange the videos of their average user. Now the first team starts with the review of the video from the other team. The review team must consist of 2-4 persons. During the whole session every reviewer first guesses the task, then commentates problems and anticipates following actions in the interaction process.

One external person must take over the role of the moderator in order to lead the discussion to a productive point. While the review team analyzes the video, at least one of the other team must log the main predicates of the session. The discussion is recorded on video for later transcription. As mentioned before, no one of the review team knows the tasks from the other team, and so in the first round the SET discussion is focused on the task detection. The PnP screen record of the average user is thereby viewed one time and without sound. If the review team can't identify a task in the first round, the other team explains the task in the next pass. It also must be said, that the number of the session passes is not limited, but at least two are required. The more passes are recorded, the more insights may be won. In our case, we recorded two passes per team. An important role during the sessions takes the moderator. In the first round, he has to guide the discussion towards the task detection, and to summarize the possible task.

In the second round the SET discussion should specifically pay attention to the behavior and expressions of the user, to get more information about his expectations, intentions and thoughts during the task completion - especially when he has a problem. Another question that arises at this point is: Has the user completed the task or not? If the second round doesn't generate satisfactory results, this process can be repeated as often as necessary. In this round(s), the moderator has to lead up the discussion towards some problem solving approaches. So the moderator has a continually steering function, and he is also responsible for ensuring that the team will not waste too much time on one problem.

A help for the reviewers to comment and anticipate the interactions may be first problem oriented statements like *The user does … because he thinks / perceives / feels …* These statements may then be reversed in order to generate solutions, e.g. *In order to prevent that the user … thinks / perceives / feels etc. …* Another solution oriented statement could be *In order to support the user … thinking / perceiving / feeling etc. a possible solution would be…* and so on.

The result of this iterative review process is a list of improvements for single steps in an interaction process. From the detailed review of single steps may also general improvements derived. However the generated solutions still need to be refined from the scribe's notes and the video recordings of the discussion session. The teams change their actions after team 1 finishes step 2 of the discussion, thus team 2 analyzing the data of team 1 now, while team 1 takes notes. Figure 5 summarizes the actions to be taken in the different steps.

3.4.3 SET Analysis Phase

The last step in the SET method is the analysis phase. Here the video recordings of the discussion phase are transcribed in order to refine and consolidate the findings from the discussion. Significant statements are investigated further. Each group transcribes the video of the other group, thereby analyzing their own data. The outcome is a list of problems and recommendations. This list should give the developers an insight into the thoughts of the user, to improve the system in a further consequence. In the most optimal case, developers are part of the evaluation teams.

Team	Roles	Discussion Step 1	Discussion Step 2	Analyse
Team 1	Expert 1	First review of video from team 2, find out unknown tasks	second review of video, comment behavior, anticipate interactions	
	Expert 2			
	Expert 3			
Team 2	Expert 1	scribe, take notes of guesses for tasks, no help for guesses at this point	scribe, help with task definitions, take notes of main discussion points	Transcribe session from video of discussion, refine results
	Expert 2			
	Expert 3			
	Moderator	focus discussion on task identification, summarize, conclude	focus discussion on problems and solutions, observable behavior, summarize, conclude	

Fig. 5. Summary of actions in the SET discussion and analysis phase

4 Results and Discussion

According to a predefined user profile it was limited to the areas and tasks, which are accessible by students. FHE revealed a comprehensive list of issues according to the predefined heuristics (see an sample extract in figure 6). Naturally these issues are based on assumptions and thus provide only hypothetical and limited insight into the underlying problems.

However, they provided a starting point for further investigation, as we used them for defining tasks for the following performance test. The performance test provided us with screen recordings and video records of the subjects during the trials. Selected records of average users were then reviewed, commented and analyzed.

While the Focused Heuristic Evaluation anticipates and explains general possible issues, the Performance Test shows real issues, however without explanation. The Shadow Expert Technique reveals the reason for real issues and provides solutions. Problems and flaws in the interaction design become visible. The art of discussing the users inner states while observing his behavior, combined with the possibility to review the actions again and in detail may prove valuable for user experience design. The outcome of the SET was a comprehensive discussion of problems and a list of tiny but effective improvements that should satisfy the needs and expectations of the biggest part of the users. We expected to discuss in the SET at least some of the issues detected in the Focused Heuristic Evaluation, however we found that the SET reveals problems and solutions on a deeper and more detailed level than the generalized assumptions of the FHE. Compared to the Thinking Aloud Method the SET reveals

Nr.	Title	Description	Image	Severity
1	Logout	The way to logout is always different.	1,2,3,4,5,6	3
2	RSS	The RSS button is in a different place in every room.	1,2,3,4,5,6	2
3	Position Sidebar	The Sidebar is not in a constant place.	1,2,3,4,5,6	3
4	Right Sidebar	The right sidebar is only available in the course main page.	1,3,4,5,6	1
5	SMS	The SMS button is only available in the course main page.	1,3,4,5,6	2
6	Mobiler Zugang	The Mobiler Zugang button is only available in the course main page.	1,3,4,5,6	1
7	Back Course Page	This link disappears on the FAQ room and on the chat room.	5,6	3
8	Hauptmenü Button	The button does the same, but is described differently.	5,6	2

Fig. 6. A sample extraction of the list of problems found

more objective data as tiny parts of an interaction process are reviewed in depth. While Thinking Aloud provides information on subjective experiences and recommendations from users, the SET method is based on objective observation of behavior and subjective simulation of the users inner states, in order to anticipate emotions and further courses of interaction.

5 Conclusion and Future Research

In this paper, we have presented a new usability test method called Shadow Expert Technique (SET). We described the path from our research on consistency to the results for the developers. One of the biggest advantages is the more natural and less intrusive interaction of the user with the system in comparison to more conventional methods, such as the thinking aloud test. Advantages include: its timesaving properties, the performance test only lasts 15 minutes per person, subjective properties, it enables you to see the system through the users eyes and to get a feeling for their expectations. Consequently, developers can gain helpful insight on user issues by stepping through single interaction processes of the system. Optimizing these interactions result in a significantly improved usability and acceptability.

Future research on the SET may include the data resulting from a Thinking Aloud Method as well as a Performance Test. Thereby enabling the comparison of comments by the experts calmly reviewing the video in retrospect, and the comments of the user thus revealing the degree of discrepancy between the users' perception and the expert reviewing the users' actions.

This combination of the users' and the experts' comments, issues and solutions, provides the foundation of improvement, leading to successful interaction, increased accessibility and an eventual increase in learners acceptance.

Acknowledgements

We cordially thank the following students of the lecture 706.046 "AK Human-Computer Interaction: Applying User Centered Design" for their enthusiasm in carrying out the evaluations: Claus Bürbaumer, Marco Garcia, Daniela Mellacher and Thomas Gebhard.

References

1. Chu, L.F., Chan, B.K.: Evolution of web site design: implications for medical education on the Internet. Computers in Biology and Medicine 28(5), 459–472 (1998)
2. Shneiderman, B.: Designing the User Interface. Strategies for effective Human-Computer Interaction, 3rd edn. Addison-Wesley, Reading (1997)
3. Rubinstein, R., Hersh, H.: The human factor. Digital Press, Bedford (1984)
4. Grudin, J.: The Case against User Interface Consistency. Communications of the ACM 32(10), 1164–1173 (1989)
5. Rhee, C., Moon, J., Choe, Y.: Web interface consistency in e-learning. Online Information Review 30(1), 53–69 (2006)
6. Nielsen, J.: Coordinating User Interfaces for Consistency. The Morgan Kaufmann Series in Interactive Technologies. Morgan Kaufmann, San Francisco (2001)
7. Tanaka, T., Eberts, R.E., Salvendy, G.: Consistency of Human-Computer Interface Design - Quantification and Validation. Human Factors 33(6), 653–676 (1991)
8. Ozok, A.A., Salvendy, G.: Measuring consistency of web page design and its effects on performance and satisfaction. Ergonomics 43(4), 443–460 (2000)
9. Satzinger, J.W.: The effects of conceptual consistency on the end user's mental models of multiple applications. Journal of End User Computing 10(3), 3–14 (1998)
10. Satzinger, J.W., Olfman, L.: User interface consistency across end-user applications: the effects on mental models. Journal of Management Information Systems 14(4), 167–193 (1998)
11. Norman, D.A., Draper, S.: User Centered System Design. Erlbaum, Hillsdale (1986)
12. Holzinger, A.: User-Centered Interface Design for disabled and elderly people: First experiences with designing a patient communication system (PACOSY). In: Miesenberger, K., Klaus, J., Zagler, W.L. (eds.) ICCHP 2002. LNCS, vol. 2398, pp. 33–41. Springer, Heidelberg (2002)
13. Norman, D.A.: Cognitive engineering. In: Norman, D., Draper, S. (eds.) User Centered System Design: New Perspectives on Human-Computer interaction. Erlbaum, Mahwah (1986)
14. 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), 279–290 (2008)
15. Holzinger, A., Kickmeier-Rust, M.D., Wassertheurer, S., Hessinger, M.: Learning performance with interactive simulations in medical education: Lessons learned from results of learning complex physiological models with the HAEMOdynamics SIMulator. Computers & Education 52(2), 292–301 (2009)

16. Krug, S.: Don't Make Me Think: A Common Sense Approach to Web Usability. New Riders, Indianapolis (2000)
17. Holzinger, A.: Usability Engineering for Software Developers. Communications of the ACM 48(1), 71–74 (2005)
18. Nielsen, J., Molich, R.: Heuristic evaluation of user interfaces. In: CHI 1990, pp. 249–256. ACM, New York (1990)
19. Kamper, R.J.: Extending the usability of heuristics for design and evaluation: Lead, follow get out of the way. International Journal of Human-Computer Interaction 14(3-4), 447–462 (2002)
20. Hvannberg, E.T., Law, E.L.C., Larusdottir, M.K.: Heuristic evaluation: Comparing ways of finding and reporting usability problems. Interacting with Computers 19(2), 225–240 (2007)
21. Nielsen, J.: Finding usability problems through heuristic evaluation. In: CHI 1992, pp. 373–380 (1992)
22. Javahery, H., Seffah, A.: Refining the usability engineering toolbox: lessons learned from a user study on a visualization tool. In: Holzinger, A. (ed.) USAB 2007. LNCS, vol. 4799, pp. 185–198. Springer, Heidelberg (2007)
23. Bailey, R.W., Wolfson, C.A., Nall, J., Koyani, S.: Performance-Based Usability Testing: Metrics That Have the Greatest Impact for Improving a System's Usability. In: Kurosu, M. (ed.) Human Centered Design HCII 2009. LNCS, vol. 5619, pp. 3–12. Springer, Heidelberg (2009)
24. Virzi, R.A.: Refining the test phase of usability evaluation: how many subjects is enough? Human Factors 34(4), 457–468 (1992)
25. Nielsen, J.: Usability Metrics: Tracking Interface Improvements. IEEE Software 13(6), 12–13 (1996)
26. Bevan, N.: Measuring Usability as Quality of Use. Software Quality Journal 4(2), 115–130 (1995)
27. Thomas, C., Bevan, N.: Usability Context Analysis: A Practical Guide. National Physical Laboratory, Teddington (1996)
28. Bevan, N.: Quality in Use: Incorporating Human Factors into the Software Engineering Lifecycle. In: 3rd International Software Engineering Standards Symposium (ISESS 1997), pp. 169–179 (1997)
29. Macleod, M., Bowden, R., Bevan, N., Curson, I.: The MUSiC performance measurement method. Behaviour & Information Technology 16(4-5), 279–293 (1997)
30. Bevan, N.: Extending Quality in Use to Provide a Framework for Usability Measurement. In: Kurosu, M. (ed.) Human Centered Design HCII 2009. LNCS, vol. 5619, pp. 13–22. Springer, Heidelberg (2009)
31. Stickel, C., Scerbakov, A., Kaufmann, T., Ebner, M.: Usability Metrics of Time and Stress - Biological Enhanced Performance Test of a University Wide Learning Management System. In: Holzinger, A. (ed.) 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian-Computer-Society, pp. 173–184. Springer, Berlin (2008)
32. Seffah, A., Metzker, E.: The obstacles and myths of usability and software engineering. Communications of the ACM 47(12), 71–76 (2004)
33. Seffah, A., Donyaee, M., Kline, R.B., Padda, H.K.: Usability measurement and metrics: A consolidated model. Software Quality Journal 14(2), 159–178 (2006)
34. Holzinger, A., Searle, G., Kleinberger, T., Seffah, A., Javahery, H.: Investigating Usability Metrics for the Design and Development of Applications for the Elderly. In: Miesenberger, K., Klaus, J., Zagler, W.L., Karshmer, A.I. (eds.) ICCHP 2008. LNCS, vol. 5105, pp. 98–105. Springer, Heidelberg (2008)

35. Bevan, N.: Usability is Quality of Use. In: Anzai, Y., Ogawa, K., Mori, H. (eds.) 6th International Conference on Human Computer Interaction. Elsevier, Amsterdam (1995)
36. Kirakowski, J., Corbett, M.: SUMI: The Software Usability Measurement Inventory. British Journal of Educational Technology 24(3), 210–212 (1993)
37. Brooke, J.: SUS: A "quick and dirty" usability scale. In: Jordan, P.W., Thomas, B., Weerdmeester, B.A., McClelland, A.L. (eds.) Usability Evaluation in Industry. Taylor & Francis, Abington (1996)
38. Raskin, J.: The Humane Interface: New Directions for Designing Interactive Systems. Addison-Wesley-Longman, Boston (2000)
39. Harbich, S., Auer, S.: Rater bias: The influence of hedonic quality on usability questionnaires. In: Costabile, M.F., Paternó, F. (eds.) INTERACT 2005. LNCS, vol. 3585, pp. 1129–1133. Springer, Heidelberg (2005)
40. Smith, L.A., Turner, E.: Using Camtasia to develop and enhance online learning: tutorial presentation. Journal of Computing Sciences in Colleges 22(5), 121–122 (2007)