Shadow Expert Technique (SET) for Interaction Analysis in Educational Systems

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

[1] Social Learning / Computing and Information Services, Graz University of Technology, Steyrergasse 30/I, A-8010 Graz, Austria
tickel@tugraz.at, martin.ebner@tugraz.at
[2] Institute of Medical Informatics, Statistics and Documentation Medical University Graz, Auenbruggerplatz 2/5, A-8036 Graz, Austria
andreas.holzinger@meduni-graz.at

Abstract. This paper describes a novel usability method called Shadow Expert Technique (SET), which was applied on the learning management system of the Graz University of Technology in two different trials, with focus on consistency and visual complexity. This is the summary of the development of this new method and the approach to generalize it as a new way to get deeper insight into interaction processes.

Keywords: Shadow Expert Technique, Usability Test, LMS, Methods.

1 Introduction

This paper describes the work done on developing a technique for analyzing tiny bits of user interactions, in order to improve the interface as well as the interaction processes within our self-developed university wide Learning Management System (LMS) at Graz University of Technology. We will show how this technique was applied at two full scale usability tests and then discuss the lessons learned, thereby generalizing our so called Shadow Expert Technique (SET) [5]. Studies have shown that the consistency of user interfaces in LMSs plays a significant role on the learning performance [16]. The application of SET on our LMS is among these lines, as we try to reduce learner's cognitive load, by examining potential influential variables. So far the presented work was focused on improving consistency and visual complexity. At this point it might be important to mention that the development of SET as well as the experiments took place as part of a lecture course in advanced chapters of human computer interaction at Graz University of Technology during two consecutive years. The Shadow Expert Technique (SET) consists of several steps with two groups who strive to decompose the interaction processes of an end user, in order to derive suggestions for improvement. The primary advantage of SET is thereby the in-depth understanding of the end users behavior and actions by mirroring his emotional state, anticipating his thoughts as well as his next actions and discussing the observations and estimations within the group.

C. Stephanidis (Ed.): Universal Access in HCI, Part IV, HCII 2011, LNCS 6768, pp. 642–651, 2011.

2 Theoretical Background and Related Work

This chapter will roughly outline some topics which are related to the SET and provide some insight for a better understanding of the technique. SET is based on three foundations. The first is that emotions are elicited during the interaction of the user with the system. Emotions are a fast and fluid process, but if they can be detected by the analyst, they provide good markers for usability issues. The second foundation is the theory of mirror neurons, which are said to provide a human being with the ability to simulate the inner state of others, thereby providing also a basic level of empathy. This makes it even for 'untrained' persons possible to detect basic emotions like frustration or happiness. The third foundation is the dynamics of decision making and solution generation in moderated groups. Beside these three topics, this chapter provides a short description of the UE methods, Performance Testing, Thinking Aloud and Focused Heuristic Evaluation, as they were used to provide the necessary input material for the SET. A general description of SET can be found in chapter 3.

2.1 Emotions and Usability

Emotions are fast and an ever changing floating process during an interaction. They are created every time when an important change in the environment is perceived. Emotions provide us with the readiness to act in a certain way [4]. They are a psychological state that helps managing the achievement of goals, such that the relevance of events towards these goals is evaluated and eventually rewarded. Positive emotions occur when the goal is advanced and negative emotions appear when the goal is impeded. In usability engineering especially negative emotions provide valuable markers for issues. However they are far more important as they influence and interrupt ongoing interactions and even tend to change the course of action towards positive emotions. No user likes to achieve a goal with software that repeatedly plays with his frustration tolerance.

There are several approaches of labeling and categorizing basic sets of emotions [15]. Following a base like facial expressions and the well renowned coding system FACS [3] would provide six distinct categories, which are happiness, sadness, anger, fear, disgust and surprise. For the studies described within this work especially the negative emotions are of interest. In particular the detection of frustration is an interesting topic. Frustration is related to anger and disappointment, which is a negative form of surprise. It can be considered as problem-response behavior and will eventually lead to further negative behavior like narrow-minded problem solving approaches. The level of frustration can task wise be measured by using a combined recording the subject's arousal and valence [19] or using visual self-reports like SAM [10] or PrEmo [2]. However in the discussed studies we used a straight forward approach for detection by letting the observers 'mirror' the subjects, thereby becoming the users 'shadows'.

2.2 Mirror Neurons and Empathy

Mirror Neurons are a certain class of neurons for understanding the actions, intentions and emotions of other people, as well as the social meaning of their behavior. They

were found by Rizzolatti [18] in 1995 and are since then subject of much controversial discussions. Rizzolatti states that mirror neurons allow human beings to grasp the mind of others by direct simulation instead of conceptual reasoning [17]. As this mechanism helps also to understand the emotions of others, it can also be seen as neuronal basis for empathy. A simple example might be that when a person observes another person pulling back an arm, as if to throw a ball, there would be a copy of this behavior in the observer's brain that helps understanding the goal. Additionally the facial expression gives clues about the inner state of other person, e.g. to solve the question if the ball is thrown to hurt the observer or just to play. In this way the mirror neurons help to anticipate the next steps by simulating the actions and emotions of those we observe.

For SET the step of anticipation is interesting in particular, as we want to find out what happens to a person, who clicks a button for the tenth time, or just gets cryptic error messages. Finally this leads to the conclusion that the (screen record) playback observation of a subject's actions on an interface together with a video of the face is sufficient to give insight into the interaction process as well as identify important emotional states like frustration or happiness, without the explicit knowledge of an emotion coding system like FACS or EmFACS [3]. The efficiency of this simulation can be even increased by several replays of the situation and by asking the observers to physically mimic the subject.

2.3 The Power of Groups

Group decision making refers to a process in which multiple individuals act together in order to analyze problems and generate solutions. The efficiency of groups is influenced by their size, demographic makeup, structure, contextual settings, goals and actions. In such a process the group utilizes the diverse strengths and expertise of its members, thereby generating a greater number of alternatives than an individual. A simple example for this is a brainstorming session. In order to work effective there's the need of a leader for a group. In case of the SET this role is taken by the investigator respectively moderator, who determines tasks and structures the discussion. The moderator clearly frames the problem and encourages the group to develop several creative options, without indicating preferences for any particular solution [11]. In our experience the structured discussion within the group generates a collective understanding of issues on the one hand and diverse solutions on the other hand. The exact role of the moderator will be explained in chapter 3.

2.4 Focused Heuristic Evaluation (FHE)

The Heuristic Evaluation is an informal usability inspection method that allows identifying potential problems in the design of user interfaces by following recognized usability principles [13, 9]. In this method, a list of pre-defined heuristics is provided to a small team of evaluators. The number of evaluators is still debated; however, Nielsen [14] recommends using three to five evaluators. Each evaluator works independently and tries to find out potential problems and positive aspects of the interfaces according to the checklist of provided heuristics. The purpose of this independence is to ensure an objective and unbiased evaluation. In simple Heuristic

Evaluations, these heuristics usually focus on discovering various general and diverse issues regarding the design and behavior of the system. However, in the case of a Focused Heuristic Evaluation, only a single issue is chosen, and a list of appropriate heuristics is generated accordingly for evaluation [5, 8].

2.5 Performance Test (PT)

Performance testing helps to identify issues that influence the performance of an end user while using a working system [16]. The performance of users are usually evaluated in three dimensions i.e. time, effectiveness and user efficiency. In order to gather reliable and precise information about the system for a performance test, the test must take place under realistic conditions with real end users. As suggested by Virzi [20], usually from 12 to 20 test persons are sufficient to gather reliable data. Each test is usually video and audio taped to help developers to compare and evaluate different system designs in an iterative development cycle. The core indicators of a user's performance such as Task Effectiveness, User Efficiency, Relative User Efficiency and User Satisfaction can be obtained as defined in ISO 9241-11 [1, 7].

2.6 Thinking Aloud (TA)

The "Thinking Aloud Method" was introduced by Lewis in 1982 to understand cognitive processes [12]. It encourages test users to speak verbally about whatever they are looking, feeling, and thinking, while performing any task. Test sessions are usually recorded on video such that the actions of participants can be tracked back to see what participants did, and how they reacted. This method helps observers to understand the cognitive processes involved in completing any task. By engaging actual or intending users as participants, the thinking aloud method provides a closer view of how users use the system and reveals several usability problems in performing any particular task [6].

3 Methods and Materials

This chapter provides a general description of the Shadow Expert Technique (SET) with detailed information on the three phases. It will further discuss the experimental design and argue the changes for optimizing the technique.

3.1 General Description of the SET

The Shadow Expert Technique (SET) is an experimental multilevel usability method that utilizes the input of well renowned methods like Heuristic Evaluation (HE), Thinking aloud (TA) or Performance tests (PT). The core is a structured discussion which aims at getting insight into the inner state of a subject during an interaction, thus revealing already slight negative emotions, like beginning frustration. This in turn provides a deeper insight into occurring issues and generates interesting solutions. The approach can be seen as stepwise thinking into the user and will be described in detail within this chapter. The Shadow Expert Technique consists of three distinct phases [5] as can be seen in fig1 below. Every phase has specific goals and actions.

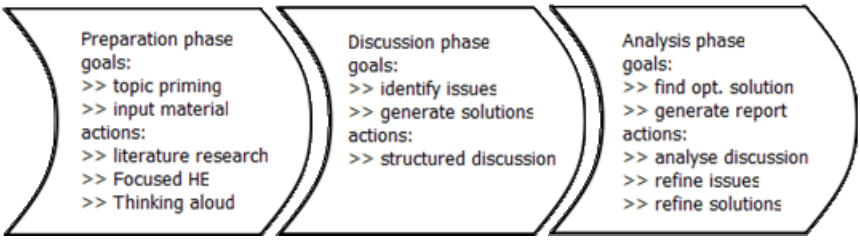

Fig. 1. Goals and actions in the three phases of SET

The first part is the preparation phase. In this phase the evaluators are primed on a specific topic (e.g. consistency) and the interface, and then split up by the investigator into two groups, which generate the input material for the discussion. Two groups are necessary here in order to later benefit from the effect of 'fresh eyes'. In the second phase, synchronized screen recordings and video material from the first phase are exchanged between the groups and reviewed in the actual discussion, in order to identify the issues and generate solutions. The discussion is recorded for the final analysis phase, which is meant to describe all issues in detail, elaborate optimum solutions and generate a report. Therefor the record of the discussion is analyzed and all issues and solutions are refined. The final report contains a list of potential problems and alternative/optimum solutions to improve the system.

3.2 Preparation Phase

The preparation phase aims at priming the evaluators on a specific topic on the one hand and on getting them to know the interface on the other hand. A literature research can provide the background on the topic and a list of specific heuristics. However it's not necessary to use a specific topic like consistency or visual complexity, as the method should also work in a more general way. In case that there is no specific topic, the Focused HE is unnecessary and we'd suggest to rather use a general HE according to Nielsen [6]. The investigator will the split the evaluators in two groups which derive tasks for a TA based on the most severe problems discovered during the FHE or HE. Thereby the tasks for are created by each group independently with the obligation not to let know the other group. Then the investigator should select the tasks for every group, such that the groups use different tasks for the TA. The TA should be run with at least seven subjects.

The outcome of the test should be a synchronized video of a screen recording, showing the interactions of the subject and a video showing the subjects facial expressions. The group will then need to evaluate the whole trial to determine a video of an average user for every task. The selection of an average user should provide more realistic data than choosing an extreme good or bad user. Further considering the rather intensive discussion phase the decision of discussing just one video per task helped to reduce the workload. However if there's enough time and resources all videos should be analyzed.

Equations (1), (2) and (3) were used to select the videos such that for each task, the user whose efficiency in a task (1) is closest to the task efficiency intersection (3) of this task was selected for the next phase. As there were five tasks, this selection

resulted in five videos as input for the discussion. The advantages and disadvantages of this selection approach are discussed in chapter 5.

$$user\,efficiency\,task_x = \frac{effectiveness}{t_{task}/t_{max}} \quad (1)$$

$$total\,task\,efficiency_x = \sum_{i=1}^{\sum users} user\,efficiency\,task_{x_i} \quad (2)$$

$$task\,efficiency\,intersection_x = \frac{total\,task\,efficiency_x}{number\,of\,tasks} \quad (3)$$

3.3 Discussion Phase

In the beginning of the discussion phase both groups exchange their material, such that group 1 will discuss the output of group 2 and vice versa. Here we use the idea of 'fresh eyes' which can discover things that have been unrecognized due to intensive work on the material. During the discussion the group that provided the material will act as scribes, thus taking notes and eventually asking questions for clarification. The investigator will take the role of the moderator and guide the discussion. The whole session should be recorded with a proper microphone and a camcorder. It will be processed together with the notes in the last phase.

In order to proper review the tasks one or two projectors and computer for playback will be needed. An optimal setup would project the screen record synchronized right beside the subjects face. Showing the videos picture-in-picture is not optimal as the face is small and eventually overlaps areas which are important during the interaction. Composing and ordering both videos with an editor program beforehand is a proper solution to this problem. The whole discussion phase consists of two consecutive steps, which can be seen in fig2.

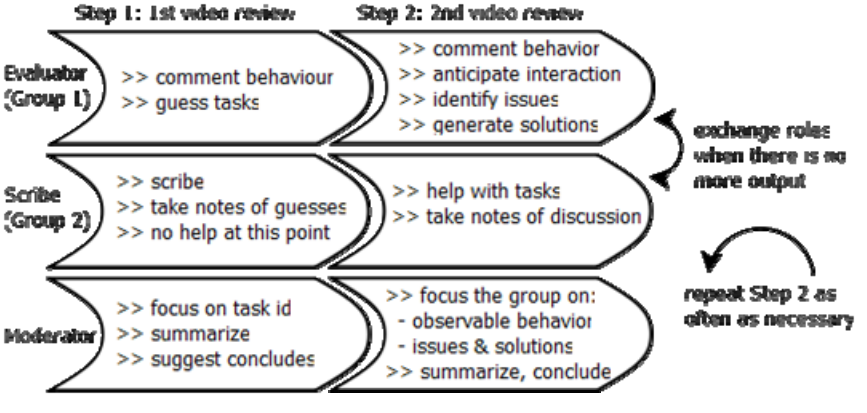

Fig. 2. Roles and actions in the discussion phase of SET

In the first step, the evaluators review the video and try to guess the tasks the user was performing. In this step, the sound of the video is muted, which avoids the evaluators to know the user's tasks in advance. Meanwhile the scribes take notes of the guesses made by the evaluators. Step 1 is finished when all tasks have been reviewed and reasonable guesses of task descriptions have been made.

In the second step the video is reviewed again, but this time with audio. Thereby the evaluators focus on understanding the user's cognitive processes involved in performing the particular tasks (by commenting on behavior and anticipation of interaction) and identifying issues as well as generating proper solutions. The audio from the thinking aloud test enables the teams to better understand the subject's expectations, intentions and emotions. Therefor the evaluators might first use problem oriented statements like The user does … because he thinks / perceives / feels… In order to generate solutions these statements can be reversed, e.g. In order to prevent that the user … thinks / perceives / feels etc. … or In order to support the user … thinking / perceiving / feeling etc. a possible solution would be… and so on. The moderator guides the discussion toward concrete solutions for identified issues. Meanwhile the scribes take notes of the discussion and may ask questions for clarification on certain statements. Step 2 can be repeated if necessary, however when there is no more output the groups should swap their roles.

3.4 Analysis Phase

In the final step of the analysis phase each group uses their notes and the record of the discussion phase to refine the results. Ideally the record is transcribed and then consolidated and clarified. Issues without solutions or multiple solutions can be investigated further, in order to generate an optimum solution. However the final report should contain the alternatives as well as the optimum solution, beside the found issues. The report should aim at suggesting the developers where and how to improve the system. Graphical mockups and flow graph of optimized interaction processes are appropriate means to communicate this. Ideally developers are part of the evaluation team or take part as silent observers in the discussion phase.

3.5 Experimental Design and Development

As already mentioned there were two trials in two consecutive years. They were used to analyze certain usability aspects. The first trial was about consistency of the interface while the second trial was concerned with visual complexity. In both cases the evaluators were primed by a three week research on the topics which led to the generation of specific heuristics for the FHE.

According to experiences in the first trial, some changes were done to optimize the technique in the second trial. In the first trial a performance test (PT) was used to generate the input material. This kind of material made observing harder, as a PT usually tries to simulate a real life setting with focus on efficiency, thus it's more difficult to find out about the users intentions and inner states. In the second trial the PT was replaced by a thinking aloud test (TA). Although replayed without audio the resulting material allowed the observers much better insight, as the subjects acted slower and expressed their emotions more clearly. Another intention of using TA was

to analyze the performance of the observers. This can be done by comparing the statements of the subject with the comments of the observers, who are unaware of what the subject is talking since the video playback is muted. This way the observers are forced to more active behavior.

Another change in the material was the utilization of an average user for every task instead of determining an average user from the whole trial. This change provided the advantage to observe much more realistic issues. A disadvantage however is that the observers have to cope the constant change of users, which might somehow be disruptive to the process of getting more familiar with the users expressions.

4 Discussion and Further Works

This paper presents ongoing research on a novel usability method called Shadow Expert Technique (SET). It was developed, tested and evolved as part of an HCI master lecture held by Prof. Holzinger at Graz University of Technology. SET was applied at the university wide LMS with the focus on improving certain usability aspects of the system. There were two trials. Between the two trial slight changes were made in order to improve SET with respect to an increase in quality and quantity of the outcome. The changes as described in the previous chapter 3.3 proved to be reasonable and applicable.

In the last trial we discovered a problem that can occur during the discussion phase. At a point where the groups have changed the roles there might be an intersection of issues. Although both groups have different tasks to analyze, there might be a common base for several issues. If this base has already been identified before switching the roles it might be hard for the 'new' evaluators to come up with genuine own solutions. Our students later proposed to do the second step of the discussion each in absence of the 'scribes'. As the whole discussion is recorded anyway this should not affect the overall performance of the SET. This will be tested in the next trial.

Another slight change in the next trial might be to instruct the moderator to focus the evaluators also on more actively mimicking the subject's behavior and expressions in order to create a more intensive immersion. It might also be interesting to use psychophysiological measures and visual self-assessment in both, the TA (for subjects) and the discussion phase (for the evaluators) in order to compare the subjects emotions with the artificial elicited ones of evaluators. Questions that might be continuative for further evolving of SET concern minimum and maximum numbers of group size, more effective priming and techniques for deeper immersion.

At the current point we conclude that the Shadow Expert Technique in the current described state is an effective (not yet efficient) method for detecting big flaws and tiny issues in interaction processes. It can be used focusing on a specific topic (e.g. consistency, visual complexity) or in a general way, whereby latter is not proved yet. SET provides a deeper understanding of the end user's behavior and usually generates in the process diverse alternative applicable solutions.

Acknowledgements. Headings 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: Markus Fassold, Claus Bürbaumer, Marco Garcia, Daniela Mellacher, Thomas Gebhard, Christian Partl, Patrick Plaschzug, Dieter Ladenhauf, Muhammad Salman Khan, Marco Lautischer, Georg Michael Lexer, Pulkit Chouhan and Alois Bauer.

References

1. Bevan, N.: Measuring Usability as Quality of Use. Software Quality Journal 4(2), 115–130 (1995)
2. Desmet, P.: Measuring emotion: development and application of an instrument to measure emotional responses to products. In: Blythe, M.A., Overbeeke, K., Monk, A.F., Wright, P.C. (eds.) Funology, pp. 111–123. Kluwer Academic Publishers, Norwell (2005)
3. Ekman, P., Friesen, W.V.: Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Palo Alto (1978)
4. Frijda, N.H.: The emotions. Cambridge University Press, Cambridge (1986)
5. Holzinger, A., Stickel, C., Fassold, M., Ebner, M.: Seeing the System through the End Users Eyes: Shadow Expert Technique for Evaluating the Consistency of a Learning Management System. In: Holzinger, A., Miesenberger, K. (eds.) USAB 2009. LNCS, vol. 5889, pp. 178–192. Springer, Heidelberg (2009)
6. Holzinger, A.: Usability engineering methods for software developers. Comm. ACM 48, 71–74 (2005)
7. 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)
8. 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)
9. 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 4(3-4), 447–462 (2002)
10. Lang, P.J.: Behavioral treatment and biobehavioral assessment: Computer applications. In: Sidowski, J., Johnson, J., Williams, T. (eds.) Technology in Mental Health Care Delivery Systems, pp. 119–137. Ablex, Norwood (1980)
11. Leonard, D., Swap, L.: When Sparks Fly: Igniting Creativity in Groups. Harvard Business School Press, Boston (1999)
12. Lewis, C.H.: Using the "Thinking Aloud" Method. In: Cognitive Interface Design. Tech. rep., IBM RC-9265 (1982)
13. Nielsen, J., Molich, R.: Heuristic evaluation of user interfaces. In: CHI 1990, pp. 249–256. ACM, New York (1990)
14. Nielsen, J.: Finding usability problems through heuristic evaluation. In: CHI 1992, pp. 373–380 (1992)
15. Ortony, A., Turner, T.J.: What's basic about basic emotions? Psychological Review 97, 315–331 (1990)
16. Rhee, C., et al.: Web interface consistency in e-learning. Online Information Review 30(1), 53–69 (2006)
17. Rizzolatti, G., Fogassi, L., Gallese, V.: Mirrors in the Mind. Scientific American 295(5), 30–37 (2006)

18. Rizzolatti, F.G.L., Gallese, V., Fogassi, L.: Premotor cortex and the recognition of motor actions. Cognitive Brain Research 3, 131–141 (1996)
19. Stickel, C., Ebner, M., Steinbach-Nordmann, S., Searle, G., Holzinger, A.: Emotion Detection: Application of the Valence Arousal Space for Rapid Biological Usability Testing to enhance Universal Access. In: Stephanidis, C. (ed.) UAHCI 2009. LNCS, vol. 5614, pp. 615–624. Springer, Heidelberg (2009)
20. Virzi, R.A.: Refining the test phase of usability evaluation: how many subjects is enough?'. Human Factors 34(4), 457–468 (1992)