

Interference with task performance using the think-aloud technique in Usability Testing

Anton Stasche

University of Osnabrueck,
Faculty of Human Sciences,
Cognitive Science Programme

Abstract

The widespread use of the think-aloud technique by Usability specialists, researchers and system developers has made the need for standardized procedures evident in recent years. Despite the lack of consistency in using any form of the method, think-aloud represents the most frequently applied data gathering method in lab testing. The current Bachelor's thesis addresses one debated problem of the concurrent think-aloud technique, interference with user task performance. Standardized instructions and guidelines for the think-aloud procedure were used in an experiment evaluating the differences in completion times of four interaction tasks in an online e-learning platform. Assumptions were made concerning the effect of the respective cognitive processes designed to vary across the tasks on the task completion times. No significant differences in overall completion time were found between the group asked to simultaneously verbalize and a control group. Users in the think-aloud condition turned out to have an advantage in the most complex task involving long-term memory. An argument is made to exercise care when using speed as a performance measure in think-aloud sessions. Interaction effects between the cognitive processes required for verbalization and those required in the tasks need also be considered.

Table of contents

Overview	4
Introduction	5
Historical perspective.....	5
The model of Ericcson & Simon	6
Think-aloud in Usability research.....	7
The User Action Framework	9
Interference of think-aloud	11
Relevance of the current Bachelor’s thesis	12
Method	15
Participants/Users	15
Equipment	16
Procedure	22
Data analysis	23
Results	23
Task performance.....	23
<i>Task completion time</i>	23
Discussion and Conclusion	26
Acknowledgments	30
Appendices	33
Appendix A.....	33
Appendix B.....	39
Appendix C.....	43
Appendix D.....	49

Overview

Software and system engineers have increasingly recognized the importance of Usability Testing (UT) in the past two decades. The testing of the user's interaction with the system for evaluative purposes can now draw upon a variety of techniques based on work from applied HCI and Psychology. Apart from using techniques like direct observation, key-stroke and mouse-movement logging and video recording, interviews and questionnaires as well as expert evaluation such as walkthrough and inspection techniques are some of the choices that are applied. Furthermore, the field of UT has identified the need to combine techniques for proper system evaluation (Nielsen, Clemmensen, & Yssing, 2002).

Among the different techniques in UT, one that is most widely used is thinking-aloud (Boren & Ramey, 2000). In one variant, concurrent think-aloud, users are asked to continuously verbalise everything that comes to mind while interacting with a given system such as a user interface. As opposed to retrospective think-aloud, where users are asked to report on their thoughts after the test, it is characteristic for concurrent think-aloud that users keep verbalising simultaneously while working with the system. This has been subject to investigation inquiring into possible disruption or distortion of cognitive processing of the user completing the task.

A great hindrance to assess possible interference of concurrent think-aloud with user performance has been the lack of standardized methodological guidelines and a thorough theoretical basis. Moreover, the articles reporting about the application of the technique in UT settings frequently miss out on positioning the technique amongst other available tools inside a sound framework uniting the different steps in the Usability Engineering cycle. This also entails using standardized descriptions of the Usability issues that can be isolated using concurrent think-aloud (Andre, Hartson, Belz, & McCreary, 2001).

The current Bachelor's thesis is an attempt to re-address the question about interference of the technique with user's performance in UT task scenarios. The following two sections will specify the theoretical elements and procedure guidelines that formed the theory behind using concurrent think-aloud. The results of the experiment that was conducted in a UT setting will be reported in the fourth section of the thesis. These will be described using terms and concepts of the User Action Framework by Andre et al. (2001) to demonstrate how the technique can be embedded into a sound framework in Usability Engineering.

Introduction

Historical perspective

The origins of think-aloud lie in the field of psychology where the practice of introspection was used by scientists such as W. James and W. Wundt (discussed i.e. in Vermersch, 1999). Using introspection, highly trained subjects or the experimenters themselves reported directly on their thought processes to test psychological hypotheses. Problems due to little consistent application and difficulties replicating the results led to the denial of introspection as a research technique with the advent of Behaviourism. Until the growth of the field of Cognitive Psychology in the 1960's, few studies addressed think-aloud (overview in Ericsson & Simon, 1980). A broad review of the studies involving verbalisation in the following period was written by Nisbett & Wilson (1977) who criticized the use of results of participant's verbalisations as data. They concluded that the subjects cannot have direct access to the cognitive processes and hence are not able to report correctly about them.

The model of Ericcson & Simon

Ericsson and Simon's (1980) work addressed these concerns in their attempt to create an underlying model for the varying forms of think-aloud that were being used by psychologists. Their model was based on a theory that regards human cognition as information processing with different types of memory stores (short term memory, long term memory). Depending on the time when the verbalisation is elicited the authors named the different types of think-aloud “concurrent” and “retrospective”. For the kind of verbal output they proposed three levels with varying degrees of reliability. Important for gaining reliable verbal data from participants was if information in short term memory (STM) could be verbalised during the tasks without much further processing and in how far verbalisation interferes with the participant’s cognitive processes. The content of STM would be comprised of currently heeded information of the task or information from long-term memory retrieved through preliminary recognition or association processes.

Level 1 verbalisations describe information in STM that can be reported directly without conversion into verbal code and without changing ongoing cognitive processing, e.g. numbers that can be verbalised in the same form as they are processed in STM. Level 2 verbalisations are characterized by the prior encoding of the information in STM, e.g. when pictures or abstract concepts need to be encoded into words. It was assumed that with Level 2 verbalisation the only effect would be a delay in task completion. Level 3 verbalisations are described as involving more cognitive processing apart from what is required for the task completion or conversion, that is, when the information is filtered or constrained prior to the verbalisation.

According to Ericcson and Simon only Level 1 & 2 verbalisations are to be considered reliable data when data collection is performed accurately. These provide data that can be

considered unbiased, though without relating to the subjective content of the verbalisation which is never to be considered hard data under their model.

Think-aloud in Usability research

The model and guideline developed by Ericsson and Simon (1980) are the most frequently quoted by authors using think-aloud in the recent years. Despite the fact that it was originally developed to deal with research problems in Cognitive Psychology, system designers and usability practitioners have recognized the utility of the technique for evaluation and applied it according to their customized needs (Bowers & Snyder, 1990; Lewis, 1982). This led Boren and Ramey (2000) to assess the applicability of the model in the field of usability research. Stating that usability research may to a certain degree be interested in user's cognitive processing, the authors stress that "the primary concern is to support the development of usable systems by identifying system deficiencies" (Boren & Ramey, 2000). This relates to the differences in practice between the highly artificial and controlled experiments in Psychology and the often incomplete systems and interfaces under study in UT settings.

One aspect of Ericsson and Simon's model is the rejection of any communication between the participant and the experimenter apart from the initial instruction to think aloud. Hence, the need for at least a minimum of communication between the user and the usability practitioner in UT settings frequently led to customized application of the technique (Jørgensen, 1989). The resulting lack of consistency in how usability researchers cite Ericsson and Simon's model but use the technique differently is Boren and Ramey's major motivation to reunite theory and practice.

In order to create guidelines for standardized procedures that are based on Ericsson and Simon's model but suitable for UT settings, the authors augmented their proposed methodological framework with elements from Speech Communication Theory. The inevitable communication in UT sessions motivated them to create guidelines for the interaction between the user and the usability practitioner. Their major suggestion is to accept the existence of communicative roles and hence to create an "asymmetrical speaker/listener relationship" (Boren & Ramey, 2000).

Boren and Ramey propose modifications of Ericsson and Simon's model in four areas. First, usability practitioners should work towards clarifying the role of the user as the "expert" and main speaker, the practitioner as the learner and listener and the system or interface as the object of study. Additionally, putting the user at ease prior to the test session is crucial. Secondly, the test procedure and data collection using think-aloud should be modified to guarantee unbiased, undisturbed and continuous verbalisation while at the same time accounting for the asymmetrical communicative situation. This can be accomplished by using reminders and acknowledgements that are as unobtrusive as possible. Thirdly, rules are provided how to intervene in typical test situations such as system malfunctions, breakdowns etc., that can arise when working with potentially incomplete systems. Such instances create issues that are not considered in the model of Ericsson and Simon who also deny any type of active probing to elicit specific responses by the users.

Taking the methodological framework of Boren and Ramey as the underlying theory, it should be possible to obtain reliable results from verbal data collection during task completion. Generally, standardized guidelines for using think-aloud in UT settings can be derived in order to make replication of results possible. An unpublished diploma thesis by Hoemske (2005) provides

detailed descriptions of instructions and procedures for the think-aloud technique when used in the usability lab for the evaluation of an online portal. The method section of the current study describes how these materials were used in an experimental paradigm emulating a usability test to ensure a standardized application of concurrent think-aloud in order to address possible interference with user task performance.

The User Action Framework

A further factor that would lead to a more consistent application of think-aloud in UT and hence strengthen its reliability is to position and define the technique in a unified conceptual framework for usability activities and problem reports. The User Action Framework (UAF, Andre et al., 2001) developed by HCI researchers at Virginia Tech provides a unifying structure in which tools for usability inspection, design, classification and reporting of usability problems can be incorporated whilst drawing upon equivalent concepts and terminology. The core principle of this work is the *Interaction Cycle* which was inspired by Norman's (1986) theory of action model that describes seven stages in the interaction of humans with a system or in terms of cognitive and physical user actions.

Using the UAF for defining the think-aloud technique as part of traditional lab testing would include the characterisation of the user's behaviour that can be extracted from their verbalisations. In the planning stage, users articulate their intentions and goals as well as the necessary tasks to complete the

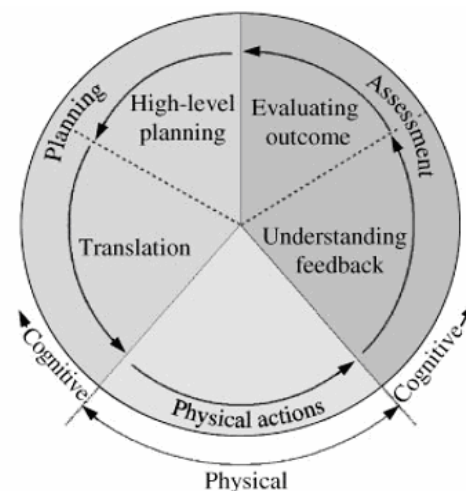


Figure 1. The parts of the Interaction Cycle : (adapted from Andre et al., 2001)

test scenario. These are consecutively translated into the physical actions which consist of perceiving and manipulating system parts or functions. Subsequently, the users assess the results of these actions based on the feedback of the system and often vigorously increase verbalisation when the outcome does not match their intentions or goals. A consistent analysis of the content of the data obtained from think-aloud sessions would pave the way for a more efficient discovery of usability problems. Likewise, it would increase the possibility to compare the results of studies using think-aloud in order to assess the utility of the technique in UT settings. Think-aloud as a usability data collection technique hence requires what Janni Nielsen and his colleagues (2002) stated in their review of the use of think-aloud in Usability Testing:

We suggest that the use of think-aloud has as a prerequisite explicit descriptions of design, test procedure and framework of analysis. (2002, p.1)

In the current study however, the application of UAF terminology is limited to the classification of the activities in the different task scenario descriptions. Further analysis of the verbal data obtained in the experiment may be subject to further research. The major motivation for specifying explicit norms and guidelines in addition to an underlying model was to increase the reliability of the technique and the validity of the obtained verbal data in UT settings. Having provided for this, the emphasis in the user test was to investigate possible interference of the users' verbalisation activities with their processing of the tasks.

Interference of think-aloud

The debate about potential disruptions in cognitive processing while verbalising has seen the technique under scrutiny for its application in experimental paradigms of Cognitive Psychology. In their seminal work “Verbal reports as data”, Ericsson and Simon (1980) provide a review of empirical studies involving concurrent think-aloud dealing mostly with Level 1 & 2 verbalisations. In accordance with their model, the authors acknowledge the lack of performance differences in tasks requiring verbalisation of Level 1 compared to control groups working silently. It is at most a difference in performance speed that could be observed in the results of studies featuring Level 2 verbalisations.

In a similar fashion, Knoblich and Rhenius (1995) report the absence of interference with performance in a wide range of experimental results from Cognitive Psychology including paradigms working with RAVEN matrices and navigation through hypertext documents. However, most of these results did not consider speed as the main indicator of participant performance. Measures of solution quality as well as strategy preference were selected as the essential data for performance analysis. Consequently, results from paradigms involving optimizing tasks (Müller, 1993, reported in Knoblich & Rhenius, 1995) that showed better performance with concurrent think-aloud in terms of a decreased error rate also entailed an increased time to completion. This was interpreted by the authors as the cost for a more analytical processing by the participants.

Deffner (1989) concludes in his review that a general increase in task completion time could be observed for tasks involving information in verbal code (Level 1) and non-verbal code (Level 2) in most of the included studies where groups of participants using think-aloud were compared to control groups working silently. As the author points out however, these studies

were selected because of their focus on defining the task material according to the dimension of verbal encoding and not e.g. the particular cognitive processes involved.

Studies incorporating speed in their performance measures rarely focus entirely on this factor. It is indeed risky to exclusively rely on task completion times when trying to assess the validity of think-aloud as a data collection technique for purposely created problem material in experimental paradigms of Cognitive Psychology. In the domain of HCI and Usability Testing, however, different settings and requirements for the useful application of the technique calls for a renewed assessment of potential interference with performance. Apart from speed measures, specification of the cognitive processes involved in e.g. evaluating a user interface should shed light on the factors that have to be taken into account upon designing and executing a user test.

Relevance of the current Bachelor's thesis

Concurrent think-aloud as defined by Ericsson and Simon (1980) has also been subject to inquiry in the field of HCI and Usability Testing reflecting the widespread use of the technique. Compared to the paradigms using think-aloud in Cognitive Psychology, typical settings in Usability Testing require a data collection method to exhibit an intrinsic robustness that is usually not as crucial in the highly controlled experimental situations of Psychology. In a UT setting, complex systems such as software in development can lead to a large number of unexpected situations which are difficult to control when designing a test and its task scenarios. As Lewis (1982) points out, application of think-aloud in user interface evaluations can entail limitations concerning the type of data obtainable. Measures of task completion time as well as the user's behaviour in the test situations are likely to be influenced by simultaneous verbalisation. Similarly, the accumulation of data (verbal protocols, time and steps to

completion, behavioural measures) is difficult to evaluate with standard statistical methods.

Reflecting this, Hartson, Andre, and Williges (2001) indicate the lack of consistent reporting on statistical analyses in review studies comparing different software usability evaluation techniques. As the authors point out, the usual focus of these meta-studies is the number and severity of usability problems found by the methods as well as cost effectiveness.

Nonetheless, an investigation into the criteria of validity, reliability and thoroughness of think-aloud as an evaluation technique in lab testing requires as a starting point a basis of empirical evidence in terms of “hard data” reflecting interference with performance of the participating users.

A good example of such empirical work in the applied HCI domain is a paradigm emulating a realistic UT setting in an article by Bowers and Snyder (1990). Not only did the authors obtain numerical data of task performance such as time and steps to completion but also included an analysis of the verbal protocols in terms of number and types of verbalisations. Despite the fact that the study compared two different types of think-aloud, concurrent and video-cued retrospective, no exact information is provided about how the participants were instructed in the two conditions. Furthermore, no arguments are made for the choice of variation in the different tasks (difficulty and monitor size) in terms of the required cognitive processing.

The current Bachelor’s thesis is an attempt to provide a similar basis of statistical data in order to allow a more complete assessment of the concurrent think-aloud technique in UT settings. Together with detailed standardized instructions and guidelines drawing on suggestions given in Boren and Ramey (2000) and prepared for application by Hoemske (2005), task scenarios were created for the evaluation of different functions of an online portal in a realistic UT environment. Task completion times were obtained in four tasks designed to involve

different cognitive operations (see Method below) for both groups (think-aloud/silent work).

Operationalisations were defined as follows:

Dependant variable: DV, Task completion time

Independent variables: IV1, Experimental condition (think-aloud/silent work);

IV2, task type (type of cognitive processing involved);

IV3, task complexity (in terms of effort and expected time to completion) was implicitly assigned for all tasks.

The operational definitions were chosen to check for possible interaction effects between the factor of experimental condition and the cognitive processes required in typical usability test scenarios. Limiting the measurement to one dependent variable, all assumptions concerning effects are focussing on the differences in task completion time. Having specified these factors, the main hypothesis was derived from results in the literature reviewed above that indicate an advantage for verbalising users in terms of a more systematic proceeding. In the current study however, this advantage was expected to be reflected in the task completion time. The tasks involving only processing in STM and basic attention were assumed to entail no effect. Working on the more complex tasks requiring planning and retrieval of information from LTM was expected to benefit from simultaneous verbalisation. On the other hand, since there was no effort made to control the level of verbalisations, a general tendency for slower processing times in the less complex tasks was expected. This was derived from evidence in the literature suggesting a delay in task completion times due to the act of simultaneous verbalisation.

The main hypothesis to be tested had its focus on a potential interaction effect between the experimental condition (IV1) and the respective task type defined by IV2 and IV3 which would be observable in task completion time (DV):

[h0] Verbalisation facilitates task processing in terms of completion time for more complex tasks involving planning and retrieval of information from long-term memory.

The statistical results are intended to be a preliminary indication of the validity of concurrent think-aloud as a data collection technique in usability tests that involve tasks with characteristics similar to those presented here. Further data includes the verbal protocols acquired in one experimental condition and results from an AttrakDiff (www.attrakdiff.org) post-test questionnaire in both think-aloud and control conditions. Although examination of these results is not included in the current thesis, recommendations are made for the effective application of the technique in usability evaluation procedures in the usability laboratory.

Method

Participants/Users

Twenty participants (15 male, 5 female) were recruited via advertisement e-mails on the Cognitive Science mailing list at the University of Osnabrueck. Only native German speakers were accepted and the students had to be acquainted with the online portal used in the experiment to a minimum degree. All participants had to sign a consent form (see Appendix D) prior to the test to allow the videotaping of the session for backup reasons and received sweets as a small compensation after the session. Participants were assigned randomly to the two

conditions with one group receiving instructions to verbalise and the other to silently work through the tasks.

Equipment

The experiments were conducted in the premises of the Usability laboratory of the University of Osnabrueck. The PC workstation was situated in the left corner of the test room which was separated from the observer room by a one-way mirror. The experimenter was sitting on a chair in a 100° angle to the left of the participant during the test session in both conditions. The test room provided video recording equipment in the form of two tracking cameras with fixed angles to record the participant's facial expressions and their hand movements. Together with the converted images of the 19"-TFT workstation screen, the video data were conveyed to the observer room's video server. The participant's voice was recorded by an unobtrusive ambient microphone. The video and audio feeds were recorded both on the video server and on backup miniDV tapes. Task completion time was measured using the recorded video footage that was augmented with a small overlay display of the run time.

As the experimental vehicle in the experiment, the e-learning platform of the University of Osnabrueck, a customized variant of the open-source content management system Stud.IP (www.studip.de) was chosen. The system had been in use at the university for three years at the time of the experiment and is a tool for regulating all academic activities during the terms. Functions include subscription to lectures and seminars, access to the respective materials by the students and organization of course lists by lecturers. At present, employing the portal is compulsory for both students and teaching personal, hence all participants were assumed to have at least some experience with the major functions and navigation.

Task scenarios

Each participant in both groups received the same set of task scenarios consisting of four tasks designed to vary in complexity and required cognitive processing. A between-subjects design was chosen to compare the performance between the participants belonging to the verbalising group and the control. In the task scenarios, concepts adapted from the UAF (Andre, 2000) are used and the respective activities and potential sources of usability issues are described according to their location in the Interaction Cycle.

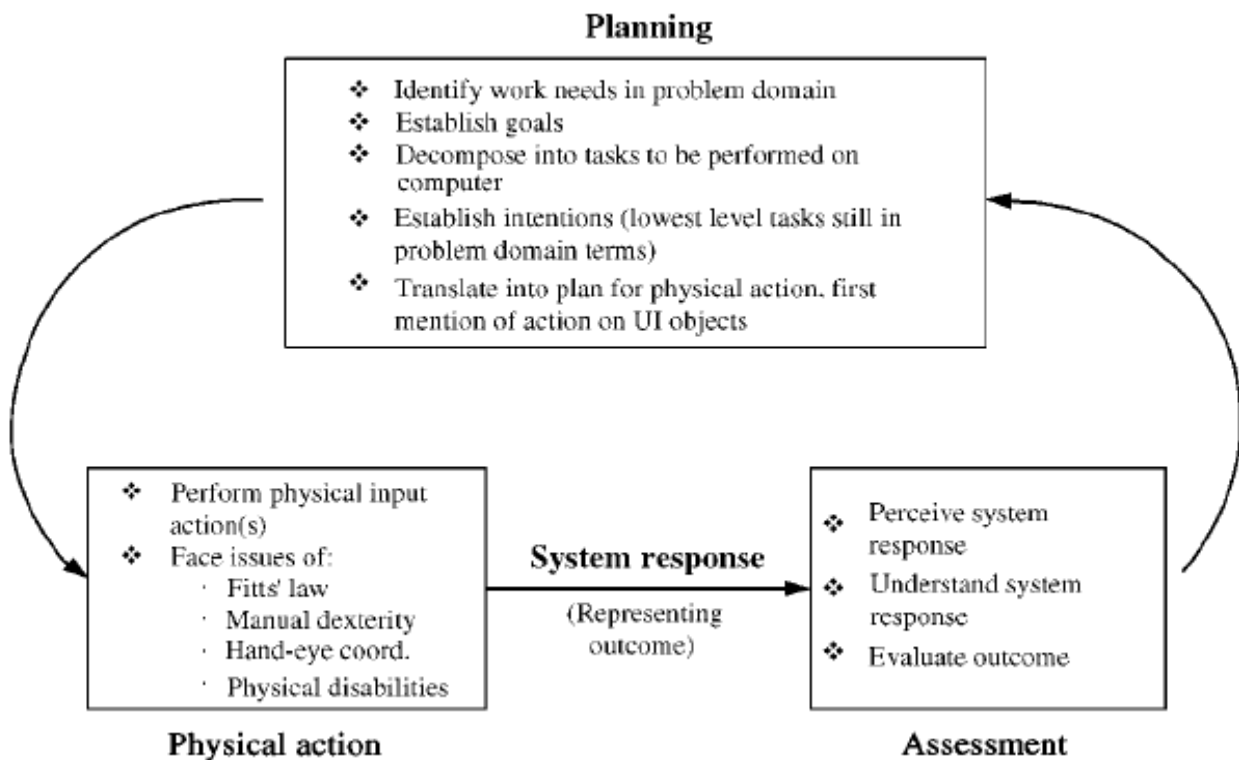


Figure 2. User task processing steps through the Interaction Cycle
(adapted from Andre et al., 2001)

The task list as given to the participants can be found together with a translation in Appendix B. The four tasks always appeared in the same order, starting with the least difficult (task 1, STM) which asked for the browsing for specific information in a course forum. The following tasks were intended to require more effort in terms of the number of instructions and steps in the scenarios. Task 2 (Att) included downloading a file, task 3 (LTM) the deliberate access of information previously encountered (in task 1) and task 4 (Plan) a choice to be made concerning the order of two subgoals to complete the scenario. All scenarios dealt with functions to be performed in the same seminar entry in the course list. The general understanding, awareness and expectations by the user concerning the system model of the portal is assumed to have been more or less homogeneous among the participants and any related issues reflected in the tasks are not further considered. More so, specific concerns are explained in terms of their occurrence in the Interaction Cycle process flow.

Task 1. STM.

The first scenario required only few steps across two levels in the course forum. A list of publications had to be located among the entries of the forum. No further processing was intended to be involved apart from short-term memory. Since there were no physical manipulations required besides navigation, potential issues respective stages in the Interaction Cycle were mainly planning and translation:

Planning: The concept of a forum is considered an issue in terms of the user's understanding of the terminology and possible functions. The instructions in the task

included the steps and locations of the relevant information. The decomposition of goals thus only required the translation of the steps into system tasks.

Translation: The existence and presentation of perceived (Norman, 1999) or cognitive affordances (Hartson, 2003), is judged as an issue regarding the objects and their possible functions (here: opening and browsing through forum entries).

Physical Actions: The manipulation of forum objects and the back button via mouse clicks is considered an issue.

Assessment: The presentation of the desired publication list as part of an entry in the forum is regarded an issue of information display. Feedback of navigation is also a potential source of usability problems.

Task 2. Att..

The second task prepared the user for the main goal, the retrieval of a certain publication in the file folder, but asked for a side-step before actually downloading the file. This detour was expected to cause a temporary shift in attention concerning the main goal. However, all processing was assumed to be taken place in short-term memory.

Planning: The decomposition of the main goal into system tasks called for restructuring of the hierarchy of these subtasks after receiving the instruction of detouring from the downloading task.

Translation: The terminology and object alignment (buttons, folder icons) in the file section of the seminar is considered as a major source of potential usability problems (presentation of cognitive affordances). Also, the return from the detour to the main task in terms of the levels in the system is important.

Physical Actions: Three possibilities to download the file via mouse-clicks are viewed as possible usability issues in the completion of the main task. The detection of the path to return to the file section is also a matter of manipulation of interface objects.

Assessment: Information display of the subgoals (especially the detour) is once more expected to entail potential problems of usability.

Task 3. LTM.

The third and most complex task had the users navigate through multiple levels in the forum and add text to two entries. The instructions asked for the augmenting of one entry which had to be recognized or recalled from the first task. This retrieval of information from long-term memory was included to trigger additional processing in short-term memory other than procedural information during the completion of the scenario.

Planning: The explicit listing of subtasks in the scenario instructions left the most effort on the interrelating of the different levels in the forum by the users.

Translation: Major instances of usability problems are expected to arise from the labelling and presentation of the forum buttons because a large part of the scenario required adding text or comments to forum entries.

Physical Actions: The opening of forum entries concerns the appropriate clicks. Further issues are attributed to the typing required in the subgoals.

Assessment: Results of the augmenting of text or comments by the users (“save”) is expected to imply usability problems in terms of feedback by the system.

Task 4. Plan.

The fourth and final task dealt with functions concerning system settings. A general goal was introduced in the instructions and two possibilities to achieve the respective subgoals were outlined. The user was then asked to deliberately choose the order for completing the subtasks.

Planning: Since the respective functionality (“visibility”) of the portal had been made accessible only recently in the system, potential usability issues were expected to arise from the inability of the users to plan steps through the system. No explicit instructions to achieve the subgoals (settings of the system) were given while only the locations of the specific functions to change the settings were named.

Translation: Similarly, the cognitive affordances to change the state of the system setting (visibility) are expected to represent usability issues. The respective objects to be

manipulated (a button/a drop-down box) are also thought to represent potential causes for usability problems.

Physical Actions: The required clicks on the objects in the two subgoals were to be performed in a different manner. The function of the drop-down box is presumed to represent an issue.

Assessment: Major issues were the display of the relevant information to deduce the current state of the system setting. Similarly, upon altering the setting, feedback provided by the system (change of text, button) is expected to raise usability concerns.

Procedure

In the beginning of each session the participant was welcomed in the greeting area of the lab and given a short outlook on the procedure. The user was shown briefly into the operator room to demonstrate the video recording equipment in order to reduce possible apprehensions of being videotaped. After signing the consent form, the user was asked to take a seat in front of the workstation and read the first page of the task manual. In the case of the think-aloud condition, it contained a description of the technique together with an example (see Appendix A). Concurrent verbalisation was practiced by letting the participant replace the batteries of a common cassette recording device which was placed on the left side of the workstation screen. The participant then read the next page with a description of the task procedure. This page was given to the control group (silent work) immediately after the consent form.

The participant was then asked to log into the portal while the experimenter returned shortly into the observer room to start the recording. Upon return, the session was started with the participant processing the first task. Total processing time for each task was additionally recorded by the experimenter using a PocketPC device. A full description of the actual procedures for a session can be found in Appendix C together with communication guidelines for the experimenter.

Data analysis

The main hypothesis was tested using 2x4 split-plot design with experimental condition (IV1) as the between-subjects factor, task type (IV2) as the within-subjects factor and task completion time (DV) as the dependent variable. The major interest lay on the expected interaction effect in the different tasks resulting from the experimental condition (think-aloud vs. silent). Apart from that, another analysis concerned the experimental condition as the between-subjects factor in measuring the average task completion time for all tasks.

Results

Task performance

Quantitative results were analysed using an ANOVA in order to search for significant interaction effects between the experimental condition (IV1) and the task type (IV1).

Task completion time

Time to completion was measured in all tasks of the two conditions to obtain mean values. All participants needed an average of 91.86 seconds for the first task (standard deviation 72.03 seconds) and 1 minute 38 seconds for the second task (standard deviation 62.39 seconds).

The third and longest task saw participants require considerably longer with an average time of 5 minutes 79 seconds and a standard deviation of 2 minutes 98 seconds. In the last task, all participants needed an average of 1 minute 93 seconds to complete it (standard deviation 1 minute 66 seconds).

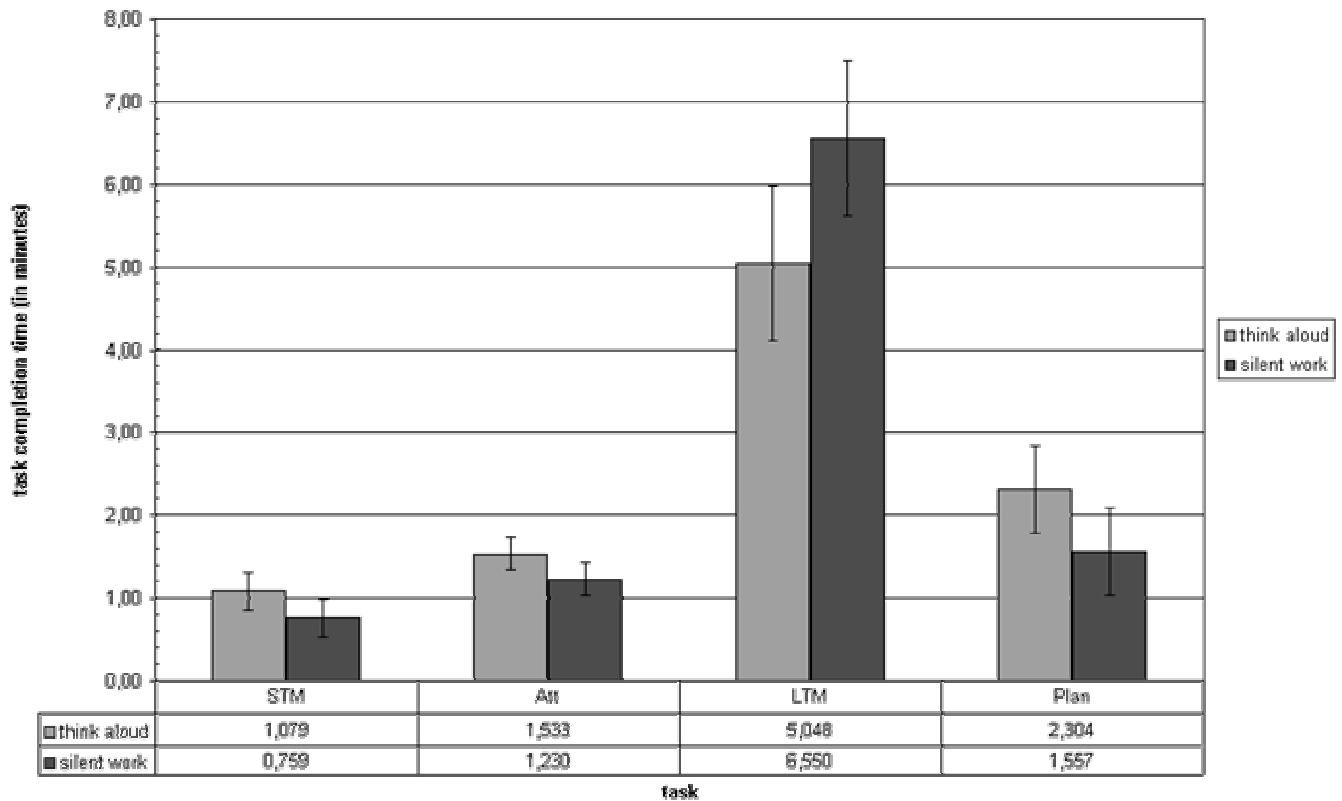


Figure 3. average task completion times and *SEM*-indicator for the four tasks in both conditions.

An applied 2x4 ANOVA (condition x task) with repeated measurement on the second factor (see Figure 3) only revealed a significant effect for the task type. Considering that the four tasks were deliberately designed to differ in amount of time required for completion (complexity, IV3), a significant difference in completion times for the within-subject factor is not surprising ($F[1.716,30.882] = 36.384, p = .000$, Greenhouse-Geisser corrected).

Concerning the between-subjects factor, differences in average task completion time (DV) across the two experimental conditions (IV1) did not turn out to entail a significant effect ($F[1,18] = .005, p = .943$). However, due to the large variance stemming from the factor of task type (differences in task completion time), the error variance for the factor of experimental condition leads to skewed results. A choice of more homogeneous task types with less variance in the respective task completion times may have yielded significance for the factor of condition.

Results concerning the main assumption (interaction effect) that would have allowed strong argumentation against relying on task completion time as a usability measure under think-aloud conditions did not emerge. An interaction effect between IV1 (experimental condition) and IV2 (type of cognitive processing)/IV3 (complexity) was not revealed. No significant differences in task completion time due to the factor of condition (think-aloud vs. silent work) in each of the tasks were discovered ($F[1.716,30.882] = 1.827, p = .182$) by the ANOVA. Hence, the absence of a statistically significant interaction effect does not support the main hypothesis which expected such an effect to appear due to advantages of the think-aloud group in tasks 3 & 4.

A more detailed analysis of the tasks and applied cognitive processes may however allow a few conclusions concerning possible interaction effects of IV1 and IV2: tasks requiring short attention span and limited processing requirements in STM exhibit no large differences in completion time between the think-aloud and silent condition. Mean task completion times for the first task (STM, 1.079min vs. 0.759min) and the second task (Att, 1.533min vs. 1.23min) do not lead to any clues apart from the slightly better performance by the silent working participants. This supports the additional assumption that a general delay in task completion time for verbalising participants was to be expected. Task 3, involving retrieval of facts from long-term memory via recognition or association, reveals an advantage for participants using think-

aloud (LTM, 5.048min vs. 6.550). In other words, a possible effect of the accessing of information in long-term memory cannot be excluded from a discussion of the results and may help corroborating the main hypothesis. On the other hand, users in the silent work condition were somewhat faster than average in the task requiring making a choice (Plan, 2.304min vs. 1.557min) which outlines the limits of confirming the main assumption.

Taking into account the small number of participants, the given data of the ANOVA over all tasks do not reflect a large observed power (0.327) for finding a rather small interaction effect size of 0.092. Focussing on task 3, the quite substantial effect ($d = 0.51$) motivated a T-test to investigate if increasing the sample size would deliver a statistically significant effect of the independent variable (IV1, experimental condition). Finding a significant effect in this task with a similar effect size and a probability of 0.80 (alpha error = 0.05) yet turned out to require at least 98 participants for each of the two conditions, resulting in a total of 98×2 participants necessary for the entire sample.

Considering the small sample size of just 20 participants, the high variance in the data could yet simply have emerged only for this particular group of participants and effects would thus be bound to disappear with a larger number of users. Nevertheless, the current selection of participants reveals that the group using think-aloud were 1.5 minutes faster in average for task 3. This is by itself capable of supporting at least preliminary assumptions.

Discussion and Conclusion

The present study investigated potential interference of simultaneous verbalisation with task performance of users working on four tasks in a realistic UT setting. Verbalisation was elicited through instructions described in the attached guidelines for the think-aloud technique

based on a model of human information processing developed by Ericsson and Simon (1980) and methodological work by Boren and Ramey (2000) and Hoemske (2005). Evaluating user task performance was performed by recording and reporting the aspect of task completion time using representative statistical tools as well as terms and concepts from Andre et al.'s User Action Framework.

Within the data acquired with the current experimental paradigm, no statistically significant effects due to the factor of experimental condition could be isolated. Task completion time yielded no significant differences for the main between-subjects factor (IV1). Furthermore, a statistically significant interaction effect between the condition (IV1) and the task type (IV2) in terms of task completion time (DV) could not be observed. Consequently, the type of cognitive processing as operationalised here does not hint to any significant effect on the task completion time. A post-hoc power analysis reveals that an increase in participant population size is not predicted to lead to the eventual discovery of a significant size of the interaction effect.

Considering the predictions made by Ericsson and Simon, evidence for a lack of differences in task completion time for Level 1 verbalisations could not be obtained since the type of verbalisation was not controlled for in the paradigm. Hence, type 1 and type 2 verbalisations were likely to occur unsystematically and type 3 verbalisations at least in task 3 (LTM). This restricts the analysis of potential differences in processing times to the context of the type of cognitive processing primarily required in the tasks (IV2). This factor reveals only a slight increase in completion time when using think-aloud for tasks relying on short-term memory and basic attention. It allows the assumption that the act of simultaneous verbalisation does not strongly interfere with the serial processing of steps in the tasks. In concordance with the results from other studies and confirming one assumption stated above, only a minor increase

in processing time can be observed. Hence, although no statistically significant effect could be found for task 1 and task 2, a general delay in task processing could be observed for the think-aloud group.

Concerning the main assumption that the participants using think-aloud would benefit from a more systematic proceeding in task 3 (more complex with retrieval from LTM) and task 4 (requiring planning and choice), the results from the data analysis can at least justify a weak hypothesis concerning the advantage of the think-aloud group. The faster processing of these participants in task 3 (effect size $d = 0.51$) could be a hint to an influence of the verbalisation act on the required cognitive processing. Since no advantage of the think-aloud group could be observed in task 4, it may be feasible to conduct follow-up investigations concentrating on the characteristics of task 3 to rule out chance for the current result. Yet, interpreting the results from an a-priori power analysis shows that at least 98 participants would be needed to find a medium size effect with a probability of 0.8 that would prove statistically significant for this particular task. Due to the limited scope and time of the current Bachelor's thesis, the acquired data has thus to be regarded taking into account the small number of participants in the study.

The results of task 3 (LTM) may further hint to a speculation introduced by Knoblich and Rhenius (1995) who observed a closer orientation of the verbalising participants on directly accessible information and less pre-processing. Task 3 as the most complex task in terms of steps to completion could have led the participants in the think-aloud condition to concentrate closer on the actual task instructions and thus save processing capacity in short-term memory because they refrained from thinking ahead. This in turn may have permitted a faster retrieval and inclusion of facts from long-term memory and consequently faster average task completion times

for this task. This could be incorporated in follow-up experiments including an operationalisation of processing strategy in similar UT-oriented task scenarios.

After interpreting the current results, it can be recommended that usability specialists would be well advised to carefully create scenarios when using the think-aloud technique in user tests. Although the results for task completion time in this study do not provide statistically significant clues for interference between the technique and users task performance, user tests dealing with tasks that involve a strong recognition or association component might turn out to produce skewed results for task completion time when using concurrent think-aloud. In these cases, avoiding the use of concurrent think-aloud and applying other forms (video confrontation) instead is more likely to produce valid results.

The proceeding in the current Bachelor's thesis provides an example of how a reassessment of think-aloud as a data-collection method in Usability Testing settings can be performed. Moreover, it explicates the elements to allow reproducible results when using think-aloud in studies in the UT field. It emphasizes complete elaboration of the instructions used to elicit verbalisations by the user/participant, a framework for classifying the usability problems and its aspects in the different stages of user interaction and finally a thorough reporting of quantitative and qualitative results using the proper statistical tools. In the field of Usability Testing, a more consistent implementation like in the current study would lead to comparable results that allowed a comprehensive assessment of different usability evaluation methods such as lab testing using the think-aloud technique.

Acknowledgments

This study was conducted in the Usability Laboratory of the University of Osnabrueck. I am grateful to Frank Ollermann for introduction to the equipment and technical supervision. I thank also PD Dr.Kai-Christoph Hamborg for competent supervision and inspiration. Both are based at the Department of Work- and Organisational Psychology and gladly accepted me, a Cognitive Science-based student, for this undergraduate research.

References

- Andre, T. S. (2000). *Determining the effectiveness of the usability problem inspector: A theory-based model and tool for finding usability problems*. PhD thesis, Blacksburg, VA: Virginia Polytechnic Institute and State University.
- Andre, T. S., Hartson, H. R., Belz, S. M., & McCreary, F. A. (2001). The user action framework: A reliable Foundation for usability engineering support tools. *International Journal of Human-Computer Studies*, 54, 107-36.
- Boren, M. T., & Ramey, J. (2000). Thinking Aloud: Reconciling Theory and Practice. *IEEE Transactions on Professional Communication*, 43(3), 261-78.
- Bowers, V. A., & Snyder, H. L. (1990). Concurrent vs Retrospective Verbal Protocol for Comparing Window Usability. *Proceedings of the Human Factors Society 34th Annual Meeting*, (1270-1274).
- Deffner, G. (1989). Interaktion zwischen Lautem Denken, Bearbeitungsstrategien und Aufgabenmerkmalen? Eine experimentelle Pruefung des Modells von Ericsson und Simon [Interaction of thinking aloud, solution strategies and task characteristics? An experimental test of the Ericsson and Simon model]. *Sprache & Kognition*, 8, 98-111.
- Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological review*, 87(3), 215-51.
- Hartson, H. R. (2003). Cognitive, physical, sensory, and functional affordances in interaction design. *Behaviour & Information Technology*, 22(5), 315-338.

- Hartson, H. R., Andre, T. S., & Williges, R. C. (2001). Criteria for Evaluating usability evaluation Methods. *International Journal of Human-Computer Interaction*, 13(4), 373-410.
- Hoemske, T. (2005). *Qualitätssicherung in der Softwareevaluation*. diploma thesis, University of Osnabrueck, Osnabrueck, Germany.
- Jørgensen, A. H. (1989). Using the Thinking-Aloud Method in System Development. In G. Salvendy & M. J. Smith (Eds.), *Designing and Using Human-Computer Interfaces and Knowledge Based Systems* (pp. 743-750). Amsterdam: Elsevier.
- Knoblich, G., & Rhenius, D. (1995). Zur Reaktivität Lauten Denkens beim komplexen Problemlösen [The Reactivity of Thinking Aloud During Complex Problem Solving]. *Zeitschrift für Experimentelle Psychologie*, XLII(3), 419-54.
- Lewis, C. (1982). Using the "thinking aloud" method in cognitive interface design. *IBM Research Reports RC 9265* (40713). Yorktown Heights: New York: IBM Thomas J. Watson Research Center.
- Nielsen, J., Clemmensen, T., & Yssing, C. (2002). *Getting access to what goes on in people's heads? - Reflections on the think-aloud technique*. Paper presented at NordiCHI 2002, Aarhus, Denmark.
- Nisbett, R. E., & DeCamp Wilson, T. (1977). Telling more than we can know: Verbal reports on mental Processes. *Psychological Review*, 84(3), 231-59.
- Norman, D. A. (1986). Cognitive Engineering. In D. A. Norman & S. W. Draper (Eds.), *User Centered Design: New Perspectives on Human-Computer Interaction* (pp. 31-61). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Norman, D. A. (1999). Affordance, conventions, and design. *Interactions*, 6(3), 38-42.
- Vermersch, P. (1999). Introspection as practice. *Journal of Consciousness Studies*, 6(2-3), 17-42.

Appendices

Appendix A

Experiment materials

Aufgaben Stud.IP

Nutzertest – Lautes Denken

Durchführung:

Anton Stasche, Cognitive Science Bachelor-Program,
Im Rahmen einer Bachelor-Forschungsarbeit und zur
eventuellen Auswertung durch Mitarbeiter des
Fachbereichs Arbeits- und Organisationspsychologie
Universität Osnabrück

Bearbeitung der Aufgaben – Lautes Denken

Während des Versuchs wollen wir mit Ihnen mit der „Methode des Lauten Denkens“ arbeiten. Wir wollen dabei folgendermaßen vorgehen:

Bitte sagen Sie während der gesamten Bearbeitungszeit laut was Sie tun. Sagen Sie laut was Ihnen durch den Kopf geht während der Arbeit mit dem System. Dies erscheint vielleicht ungewöhnlich, es ist aber notwendig, damit wir die Gründe von Problemen bei der Nutzung von Stud.IP besser verstehen können.

Ich werde Ihnen ein kurzes Beispiel geben, damit Sie sich das Vorgehen leichter vorstellen können. Stellen Sie sich vor, Sie bereiten sich morgens einen Kaffee zu und denken dabei laut:

Beispiel:

„Also, ich nehme die Schachtel mit den Filtertüten aus dem Regal und nehme mir eine heraus. Nun öffne ich das Teil, wo die Filtertüten reinkommen und setze die Filtertüte ein. Ich nehme mir die Kanne. Wie viel Tassen Kaffee mache ich heute wohl?: Also, ich brauche für mich mindestens 2 Tassen zum Frühstück und dann noch etwa 2 für später. 4 Tassen brauche ich also. So, nun die Kanne unter den Wasserhahn halten... den Wasserhahn aufdrehen und beobachten ob der Wasserstand schon die Linie für ´4 Tassen´ erreicht hat. Ärgerlich, jetzt habe ich etwas zu viel eingefüllt. Ich gieße ein wenig aus und kontrolliere den Wasserstand. Ja, jetzt ist der Wasserstand in etwa bei ´4 Tassen´...“ usw.

Es geht also darum, dass Sie das, was Ihnen durch den Kopf geht äußern während Sie die einzelnen Aufgaben bearbeiten.

Das Laute Denken während der Aufgabenbearbeitung kommt Ihnen wahrscheinlich ungewohnt vor. Bevor wir mit den eigentlichen Aufgaben beginnen, werden wir deshalb eine kurze Aufgabe zur Übung durchführen.

Sagen Sie nun bitte dem Untersuchungsleiter, dass Sie bis zu dieser Stelle gelesen haben!

Aufgaben:

Während der Bearbeitung der Aufgaben sind Sie unser Experte, denn nur durch ihre Interaktion mit dem Softwaresystem können wir Schwächen des Systems aufdecken. Der Versuchsleiter sorgt nur dafür, dass die technischen Voraussetzungen zur Verfügung stehen. Durch ihre Aufgabenbearbeitung lernen wir von Ihnen, wo die Probleme des Systems liegen!

Die Aufgaben in dieser Aufgabenbeschreibung sind als einzelne Punkte mit 1,2,3 usw. bezeichnet. Lesen Sie bitte zunächst nur den Anfangstext und die erste Aufgabe durch und bearbeiten dann die erste Aufgabe. Danach lesen sie sich die zweite Aufgabe durch und bearbeiten sie dann und so weiter bis zur letzten Aufgabe.

Lassen Sie sich während der Bearbeitung ausreichend Zeit für die einzelnen Aufgaben. Sollten Sie einmal an einer Stelle nicht mehr weiterkommen, so probieren Sie einfach so lange weiter bis der Untersuchungsleiter Sie darauf hinweist, dass Sie zum nächsten Punkt übergehen können. Es ist wichtig, dass Sie versuchen die Aufgaben selbständig bis zum Schluss durchzuführen. Wenden Sie sich bitte nur in äußersten Notfällen an den Versuchsleiter. Eventuelle Probleme, die während der Bearbeitung auftreten sollten werden vom Versuchsleiter protokolliert.

Wenn Sie außer der Aufgabenbearbeitung noch Fragen oder Probleme haben, können Sie sich natürlich vor oder nach der Bearbeitung an den Untersuchungsleiter wenden. Zum Abschluss des Versuchs wird Ihnen noch ein kurzer Fragebogen präsentiert, den Sie in Ruhe ausfüllen können.

***Die Aufgaben benötigen eine kleine Vorbereitung.
Sagen Sie bitte dem Untersuchungsleiter Bescheid, dass Sie bis zu dieser
Stelle gelesen haben!***

Experiment materials – English translation

tasks Stud.IP

user test – think aloud

experiment supervision:

Anton Stasche, Cognitive Science Bachelor-Program,

in the context of a Bachelor-study and for

potential further evaluation by members of the

department of Work- and Organizational Psychology,

University of Osnabrueck

Working on the tasks – think aloud

During the course of the experiment we would like to employ the “think aloud” technique. The procedures for it are as follows:

Please say out loud everything you do as you’re working on the tasks. Articulate everything that comes to your mind during your interaction with the system. This may appear a bit bizarre at first but is essential for allowing us to investigate into possible problems that arise during the use of stud.IP.

I will describe shortly an example to help you appreciate the procedure. Imagine you’re preparing a cup of coffee in your coffee machine in the morning while thinking aloud:

Example:

“Right, I’m taking the box of filter paper from the shelf and take one out. Now I open the part where the paper has to be inserted and put it in. Now I grab the jug. How many cups do I want to do today? Well, I know I definitely need at least two for breakfast and maybe two for later. So four cups. OK, I put the jug under the tap and open the tap. Now I have to watch out that the water doesn’t go over the line that indicates four cups. Darn it! Now I’ve put too much in! I pour a bit into the sink and check the level again. OK, now it’s at the right level for four cups...” etc.

You might have realized that it’s basically about verbalizing all the things that go through your head while you’re working on the tasks.

Thinking aloud while completing the tasks might come a bit odd at first. So in order to get used to it a little more, we will do a short practice exercise now.

Please tell the supervisor that you have finished reading up to this point.

tasks:

During the working on the tasks you are our expert since your interaction with the software is vital to expose weaknesses of the system. The supervisor only takes care that the technical requirements are up and running. Through your work on the tasks we learn where problems of the system are to be found!

The tasks in the scenario descriptions are denoted as individual sections with the numbers 1,2,3 and 4. Please read the introduction on the sheet and the instructions for the first task before proceeding to the other tasks. After completion of the first task you can read the instructions to the second and work on it etc.

Allow yourself enough time to work on each task. Should you feel stuck at any point, try to continue working until the supervisor tells you to advance to the next task. It is important that you complete the tasks by yourself! Please consult the supervisor only in emergencies. Potential problems that might occur are written down in the supervisor's log.

Should you have any questions or problems apart from the task procedures, feel free to contact the supervisor at any time before or after each task. In the end of the user test you will be asked to complete a short questionnaire.

The tasks require a few preparations.

Please tell the supervisor that you have finished reading up to this point.

Appendix B

Task list

Einleitung: Sie sind als Student/-in an der Universität Osnabrück eingeschrieben. Sie möchten sich eine der von Ihnen besuchten Veranstaltungen im Veranstaltungsordner des Kursmanagementsystem Stud.IP ansehen, um dort verschiedene Aktionen auszuführen. Zu Beginn der Aufgabenbearbeitung loggen Sie sich bitte als *user#* mit Passwort *user#* im System ein und wählen „Meine Veranstaltungen“.

Zur Erinnerung: Bitte sagen Sie während der gesamten Bearbeitung (d.h. ab dem Zeitpunkt ab dem Sie mit Stud.IP arbeiten) laut was Ihnen durch den Kopf geht.

Aufgaben

1. Sie sind im Seminar „Physiology of Emotions“ des Kursleiters *Anton Stasche* an Publikationen zum Forschungsbereich interessiert.
Im Kursforum wollen Sie deshalb nach einer Liste schauen, die der Kursleiter gepostet hat.
Werfen Sie hierzu einen kurzen Blick auf die Titel der wissenschaftlichen Arbeiten des Forumthemas „Liste der Publikationen“ durch.
Danach kehren Sie zu „Meine Veranstaltungen“ zurück.
2. Sie möchten sich nun eine Datei mit einer wissenschaftlichen Publikation aus dem gleichen Seminar herunterladen.
Im Dateienordner suchen Sie nach „Clugnet.pdf“.
Bevor Sie die Datei speichern, werfen Sie aber noch einen kurzen Blick auf die weiteren Veranstaltungen des Kursleiters indem Sie auf dessen Namen rechts neben dem Dateinamen klicken.
Zurück im Dateienordner speichern Sie die Datei im Ordner „Eigene Dateien“ auf ihrem PC.
Kehren Sie anschließend zu „Meine Veranstaltungen“ zurück.

3. Als nächstes wollen Sie im Forum des gleichen Seminars auf neuere Ergebnisse des Forschungsbereiches hinweisen.
Unter „Allgemeine Diskussionen“ erstellen Sie hierfür einen Antwortbeitrag mit dem Namen „Neueste Forschungsergebnisse“ und verzichten zunächst auf Ergänzung des Textinhalts.
Speichern Sie und gehen Sie anschließend zum Forumsthema aus der ersten Aufgabe und zitieren den Beitrag.
Fügen Sie den Text „Nachfolgeuntersuchung erfolgt“ unter das Zitat hinzu und speichern Sie.
Anschließend kehren Sie zum anfangs erstellten Antwortbeitrag zurück und ergänzen Ihren Beitrag mit der Zeile „LeDoux et. Al., JNPHYS, 2004“.
Speichern Sie und kehren zum Schluss zu „Meine Veranstaltungen“ zurück.

4. Abschließend möchten Sie noch die Sichtbarkeit Ihres Status als User im gleichen Seminar verändern.
Eine Möglichkeit die generelle Sichtbarkeit im System zu verringern ist unter den „Homepage“ Einstellungen im oberen Symbolmenü zu finden.
Im dortigen „myStud.IP“ Ordner können Sie Ihre Sichtbarkeit direkt verändern.
Eine weitere Einstellung lässt sich in der Teilnehmerliste des Seminars tätigen.
Gehen Sie nach eigenem Ermessen zu beiden Einstellungen und verändern Sie Ihren Status auf „unsichtbar“.
Kehren Sie anschließend zu „Meine Veranstaltungen“ zurück.

5. Loggen Sie sich zum Schluss aus Stud.IP aus.

Vielen Dank für Ihre Hilfe bei der Evaluierung!

Füllen Sie nun in Ruhe den Online-Fragebogen aus. Sie finden diesen in einem weiteren Browser-Fenster unten rechts.

English translation

Introduction: You are enrolled as a regular student at the University of Osnabrueck. In order to view information about one of your seminars you would like to gain access to the respective folder in the academic content management portal stud.IP. At the beginning of the procedure, please login as *user#* with password *user#* and choose "My courses".

Reminder: Please say out loud everything that crosses your mind during the time you're working on the tasks (from the time you start interacting with stud.IP).

Tasks:

1. You're enrolled in the seminar „Physiology of Emotions“ of the lecturer *Anton Stasche* and interested in publications available in the research field. In the course forum you therefore like to look for a list that the lecturer posted. For this you take a quick look at the titles of the scientific papers in the forum entry "list of publications". After that, return to "My courses".
2. Next you would like to download a file containing a publication of the same seminar. In the file folder you are therefore looking for a file called "Clugnet.pdf". Before saving the file though, you want to read up on further courses offered by the lecturer by clicking on his name to the right of the file name. Back in the file folder, save the file in your PC's "my documents". Return to "My courses".

3. After that, you would like to refer to the latest findings in the research field.
In the course forum you therefore create a new reply posting for the entry "General discussions" with the title "latest findings" and leave the text field empty for now.
Now save the posting and return to the forum entry from the first task and quote the posting. Add the text "follow-up research carried out" beneath the quote and save.
Next, you return to the previously created reply posting and augment the text field with the lines "LeDoux et. Al., JNPHYS, 2004".
Now save the posting and return to "My courses".

4. Finally, you would like to change the visibility of your user status in the same seminar.
One possibility to alter the general visibility of your status in the system can be found in the "homepage" settings in the upper symbol menu. Under the „myStud.IP" tab you can directly change your visibility in the system.
Another possible change of this setting can be done on the participant list of the seminar.
Check both possibilities in order of your choice and change your user status to "invisible"
Return to "My courses" afterwards.

5. In the end, please log out of stud.IP

Thanks a lot for your help with the evaluation!

Now you can take your time filling out the questionnaire. You can find it in another browser windows below.

Appendix C

Procedure guidelines – think aloud

Verfahrenrichtlinie VI 2 – Lautes Denken

„Bitte setzen Sie sich hier an den PC-Arbeitsplatz.“

„Ich gebe Ihnen nun eine Beschreibung der Methode mit der wir arbeiten wollen.“

VI legt der Vp das Blatt „Bearbeitung der Aufgaben - Lautes Denken“ vor.

Nach dem Lesen der Methodeninformation:

Einleiten der Probephase Lautes Denken:

„Wir wollen nun zur Übung eine Probeaufgabe mit dem Lautes Denken durchführen.“

Kurze Erklärung der Probeaufgabe mit dem Kassettenrekorder.

Durchführen der Probedurchlaufs.

Feedback des VI:

„Vielen Dank für diesen Probedurchlauf mit der Methode des Lautes Denkens. Es hat schon sehr gut geklappt mit den Äußerungen zum Handeln. Wichtig bei der Methode ist, kontinuierlich alles zu äußern, was ihnen durch den Kopf geht, während sie die Aufgabe bearbeiten. Versuchen Sie dies beizubehalten, während Sie die Versuchsaufgaben bearbeiten.“

„Wir wollen nun mit den eigentlichen Versuchsaufgaben beginnen.“

VI übergibt schriftliche Versuchsaufgaben an Vp.

Während des Versuchsdurchlaufes:

VI setzt sich auf vorbereiteten Platz

Beginn der Aufgabenbearbeitung durch Vp

Zeitnahme zur Eingrenzung der Aufgabendauer; Anhalten bei Systemabsturz und Problemen der Vp außerhalb der Aufgabenbearbeitung

Kommunikation anhand Leitfaden auf der nächsten Seite

Leitfaden zur Interaktion mit Vp:

Konstanter Gebrauch von Aufmerksamkeitsmarker:

Wenn eine Äußerung mit einer langgezogenen fragenden Intonation beendet wird, öffnet dies einen „Slot“, in den der VI ein leises „Mh mmm“ einfügt, der dann die Sprecherschaft wieder an die Vp zurückspielt.

Vp unterbricht das Laute Denken:

Unterbricht die Vp das LD, so äußert der VI leise ein fragendes „Mh hmm?“, wenn keine Reaktion erfolgt, fragt der VI „Und nun...?“.

Folgt keine Reaktion, fragt der VI ein zweites Mal „Und nun...?“

Folgt keine Reaktion, sagt der VI: „Bitte sagen Sie weiter laut, was Ihnen durch den Kopf geht.“

Vp sagt, dass sie mit der Aufgabenbearbeitung nicht weiterkommt

VI: „Einfach weiter probieren!“

Bei weiteren Fragen zusätzlich: „Lassen Sie sich nicht entmutigen. Versuchen Sie es einfach weiter! Dadurch helfen Sie Schwächen der Software aufzudecken.“

Vp fragt etwas, was in der Aufgabenliste enthalten ist

VI: „Vielleicht kann ihnen die Aufgabenbeschreibung eine Hilfe sein.“

Vp fragt etwas, was in der Aufgabenliste nicht enthalten ist

VI: „Versuchen Sie die Aufgabe mit den Informationen der Aufgabenbeschreibung zu bearbeiten. Wenn es nicht gleich klappt ist das nur ein Hinweis, dass vielleicht ein Problem im System vorhanden ist.“

Bei weiteren Fragen zusätzlich: „Lassen Sie sich nicht entmutigen. Versuchen Sie es einfach weiter! Dadurch helfen Sie Schwächen der Software aufzudecken.“

Vp fragt etwas, was auf ein falsches Verständnis der Aufgabenbearbeitung hindeutet

VI: „Vielleicht kann Ihnen die Aufgabenbeschreibung eine Hilfe sein. Versuchen Sie die Aufgabe mit den Hinweisen aus der Aufgabenbeschreibung zu bearbeiten. Wenn es nicht gleich klappt ist das nur ein Hinweis, dass vielleicht ein Problem im System vorhanden ist.“

Vp spricht zu VI über aufgabenferne Inhalte

VI: „Schauen Sie sich noch einmal die Aufgabenbeschreibung an. Versuchen Sie weiterhin, alles zu äußern, was Ihnen zur Aufgabenbearbeitung durch den Kopf geht.“

Systemabsturz

VI: „Scheinbar ist es zu einem Absturz des Systems gekommen. Dies liegt am Softwaresystem oder an einem Problem bei der Nutzung. Das kann uns bei der Weiterentwicklung des Systems sehr hilfreich sein. Ich werde kurz das System wieder in die Ausgangsposition zur Bearbeitung der Aufgabe bringen. Bitte warten Sie einen Augenblick. *[Pause zur Systemwiederherstellung]*
Nun können Sie die Aufgabenbearbeitung fortsetzen, in dem sie mit der Aufgabe neu beginnen.“

Nach der Versuchsdurchführung:

VI: „Vielen Dank für Ihre Mithilfe bei dieser Evaluationsuntersuchung. Zum Abschluss füllen Sie bitte noch den kurzen Online-Fragebogen aus, der sich in einem weitere Browser-Fenster befindet.“

Procedure guidelines – think aloud. English translation

Procedure guidelines – think aloud

“Please take a seat at the PC-workstation”

“I will now hand out a description of the technique we’re working with.”

The supervisor places the sheet "Working on the tasks – think aloud" in front of the participant.

After reading the description of the application of the technique:

Introduction to the practice task for think aloud:

"Now you can practice thinking aloud in a short task."

Brief explanation of the practice task with the cassette recording device.

Carrying out of the practice task by the user.

Feedback of the supervisor:

“Thanks for this practice run with the think-aloud technique. It already worked out really well with the verbalizing of all you were doing. It’s important for this technique to continuously say out loud what comes to your mind while you’re working on the task. Try to apply this also when you’re working on the tasks with the system.”

“Now we can begin with the actual tasks of the user test”

Supervisor hands over the task list to the user.

During the test:

Supervisor is sitting on the assigned seat

Start of the work on the tasks by the user.

Recording of run time to determine task completion time; Stop upon system crash or problems of the user that are other than task related.

Communication according to guidelines on the following pages.

Guidelines for the interaction with the user.

Constant use of acknowledgement tokens:

If an utterance is ended with a stretched, interrogative intonation, it opens a "slot" to be filled with a short "Mh mmm" by the supervisor in order to pass speakership back to the user.

User stops verbalizing:

Should the participant pause thinking aloud, the supervisor will utter a short questioning "Mh mmm?". In case of no reaction, a short "and now?" is directed at the user.

Should there be no reaction, a second "and now?" is uttered.

In case of no reaction, the supervisor advises shortly by "Please continue to say out loud what crosses your mind."

The user claims to be stuck and unable to continue.

Supervisor: „Just keep on trying.“

Upon further questions: "Don't let yourself get discouraged. Just keep on trying! You're helping us to uncover flaws of the software."

The user asks something that is answered in the task list.

Supervisor: "Maybe the task descriptions can help."

The user asks a question about the task that is not answered in the task list.

Supervisor: "Try to complete the tasks with the help of the task description. If something's not working it might just be a hint that there's a problem with the system."

Upon further questions: "Don't let yourself get discouraged. Just keep on trying! You're helping us to uncover flaws of the software."

User asks something that suggests an unexpected interpretation of the task.

Supervisor: "Try to complete the tasks with the help of the task description. If something's not working it might just be a hint that there's a problem with the system."

User talks to supervisor about task-unrelated topics.

Supervisor: "Have a look at the task descriptions. Keep on trying to say out loud everything that crosses your mind."

System breakdown.

Supervisor: "Apparently there's been breakdown of the system. That's due to the system or the use of its functions. This information can be very helpful for further development of the system. I will quickly reset the system back to the start of the task. Please hold on for a moment. *[Break for system recovery]*
Now you can continue with your work by starting over with the task."

After the user test:

Supervisor: "Thank you for your help with this evaluation. To wrap things up, we'd like you to fill out a short questionnaire that you can find in another browser window below."

Appendix D

Einverständniserklärung

Consent form

Einverständniserklärung zur Videoaufzeichnung

Wir möchten diese Evaluation auf Video aufzeichnen. Zwar wird diese Sitzung protokolliert, aber manchmal entgehen dem Protokollanten einige Details, so dass wir bei der Auswertung das Video zu Hilfe nehmen. Die Videos werden nur zu Auswertungszwecken und Forschungszwecken weiterverwendet. Selbstverständlich werden Ihre Daten dabei vertraulich behandelt.

Ich verstehe, dass während dieser Untersuchung Videoaufnahmen gemacht werden. Ich gestatte dem Versuchsleiter sowie den Mitarbeitern und Mitarbeiterinnen des Fachgebiets Arbeits- und Organisationspsychologie (FB 08) der Universität Osnabrück, die Aufnahmen zu den oben genannten Zwecken zu benutzen.

Ort/Datum: Osnabrück, _____

Name: _____

e-mail: _____

Unterschrift: _____

Eidesstattliche Erklärung

**Eidesstattliche Erklärung
zur Bachelorarbeit**

Name: **Stasche**

Vorname: **Anton**

Ich versichere, die Bachelorarbeit selbständig und lediglich unter Benutzung der angegebenen Quellen und Hilfsmittel verfasst zu haben.

Ich erkläre weiterhin, dass die vorliegende Arbeit noch nicht im Rahmen eines anderen Prüfungsverfahrens eingereicht wurde.

Osnabrück, den _____

(Unterschrift)