Mobile Web Usability Evaluation - Combining the Modified Think Aloud Method with the Testing of Emotional, Cognitive and Conative Aspects of the Usage of a Web Application

Franka Moritz
Hasso-Plattner-Institut für Softwaresystemtechnik
Universität Potsdam
Potsdam, Germany
franka.moritz@hpi.uni-potsdam.de

Christoph Meinel
Hasso-Plattner-Institut für Softwaresystemtechnik
Universität Potsdam
Potsdam, Germany
christoph.meinel@hpi.uni-potsdam.de

***Keywords*-Usability, Think Aloud Method, Field Study, Emotion testing**

***Abstract*—This paper proposes a usability testing method that alters a given usability testing method to make it less costly and time consuming for the investigator. The usage of user-centred methods is motivated and a combination of two user-centred methods suggested. Furthermore this method is combined with other techniques to additionally detect the state of satisfaction within the user. User based factors like emotions, opinions, cognitive and conative aspects are therefore considered. A method for the joint evaluation of all data gathered is proposed.**

I. Introduction

Nowadays user friendliness and usability are considered very important for all kinds of web interfaces. In E-Commerce for example it was shown that an improvement of the usability of a website resulted in a significant growth of the number of visitors and even more the number of sales [1]. Three main issues still exist with the current methods of usability testing. First of all the usability testing with user-centred methods still requires a large amount of equipment and is very time-consuming. Second, these methods do certainly reveal most of the usability problems, those parts of the web-sites which are not clearly designed or not easy to use. But they do not address the part of the definition of usability that was labelled *satisfaction* in the ISO-standard 9241. Last, no standard methods for the evaluation of the data gained from usability tests have been determined. This paper therefore suggests a combination of different methods in order to test the usability more cost efficient and with mobile equipment. Additionally other methods that are combined with it aim at testing the satisfaction that is an important part of the usability definition in the ISO-standard. A technique to evaluate all these methods together is also proposed.

II. Factors for the Usability Of A Web-Site

The user based factors that concern the evaluation of web applications are the basis for the research method proposed. A literature research for a first study with this method [2] showed the status quo of current research on the field for the testing of 3D-web-applications. The important factors that were identified are generalizable though, as all of them are connected to the definition of usability. The attribute *usability* is the center of the method suggested in this paper. Looking at the definition of usability one can discover the other factors from it.

1) Usability: The attribute usability first of all includes the concept of user-friendliness. Furthermore usefulness and utility are part of the meaning. A comprehensive definition of the term usability can be found in the ISO-standard 9241. According to the standard usability is a measure that shows if a user can use a product in a specific context to reach certain goals effectively, efficiently and with satisfaction. Effectiveness means the accuracy and completeness wherewith a user can reach a goal. Efficiency is the psychological and physical demand, time, material and monetary costs that arise to achieve the goal. And satisfaction arises if efficiency and effectiveness are fulfilled and the expectations are exceeded.

2) Cognitive Aspects: In order to be able to reach the desired goal in a web-application efficiently and effectively the user has to be able to easily understand the structure and content of the application. Humans understand the world and also the world wide web by building conceptual models of all things in their brain and simulating their functionality in the mind. The process of understanding is called cognition. As this process is the key to a successful usage of a web-interface, it will be addressed individually in the method suggested.

3) Emotional Factors: Emotions are important for a usability test, because they are connected with motivational aspects. Comfortable emotions are looked for and displeasing emotions are avoided. Furthermore emotions can influence cognitive processes and the other way around. Therefore, if the emotions that evolve during the usage of an application are known, one may draw conclusions on the quality of the application from them. [3] Positive and negative emotions will be considered in the usability test.

Positive emotions lead to an improved perceived usability. Triggers for positive emotions can be for example the stimulus of new features or the interactivity of applications [4, 5]. The cause of negative emotions are often technical problems, because these are a trigger for frustrations [6].

Monitoring the emotions of users can therefore give helpful insight into the users to see which parts of the web-application needs to be altered to reach the desired effect. It is also the way to fulfil the last part of the definition of good usability: to reach user satisfaction and exceed the expectations of the users.

4) Conative Aspects: Conative aspects depict the attitude which is evoked by the application or feature tested. For example a study that looked at the buying behaviour of consumers showed that very stimulating experiences on e-commerce websites will lead to a more rapid purchase whereas rather pleasing but not over-stimulating experiences will lead to more exploration and foraging on the site [7].

III. Research Methods

In order to address all before mentioned user based factors, a combination of several research method needs to be applied. In this paragraph the different methods will be explained and it will be shown how they can be combined.

A. Usability evaluation

Usability evaluation is the systematic investigation of the practicability of an item. It is used as a conception and decision guidance or as comparison of alternatives as it checks the object of investigation, gives an evaluation about it and shows possibilities for improvement. [8]

1) Overview of usability evaluation methods: There are two main branches of usability evaluation methods, expert oriented, also called analytical, and user oriented, as well named empirical, methods. With expert oriented methods the experts simulate with the help of their experience how a user would use a certain object. Techniques like heuristics, check lists, cognitive walk-through and usability audits with design guidelines belong to these methods. They are oriented at general usability guidelines. Their advantage is the comparably low effort. The major disadvantage is that the experts simulation is never an exact replica of a users' behaviour and an insight into details might not be possible. With user oriented methods the users are the evaluators of a system. Product tests in a laboratory (for example with the think-aloud- or plus-minus-method), field studies, focus groups and interviews are some techniques for this method. Those are very time and resource intense, but deliver tangible data that reflects real opinions, problems and proposals by users. These methods will be explained shortly in order to make a decision on the most reasonable method.

Group discussions are mostly utilized for new products in order to find out prejudices and opinions of users by discussing the products in a group. During user tests in a laboratory the probands work on different previously defined, usually realistic, tasks. They will be observed by the researcher and their actions will be recorded. With the plus-minus-technique the probands afterwards evaluate different aspects of the product with a positive or negative marking. The think-aloud-method involves that the probands talk their thoughts out loudly and therewith explain their actions. In the co-discovery-method two candidates work together on the tasks and their interactions are recorded. Fieldwork includes the observation of the user in his familiar surrounding, for example in order to find mistakes or detect new requirements. Surveys, like on-line-panel or on-screen-survey, involve questionnaires. They are chosen to generate quantitative data and are mostly used to analyse trends. [8, 9]

2) Selection of a usability evaluation method: Analytical usability evaluation methods are not useful for the goal defined, as an analysis of emotional, conative and cognitive aspects requires insight into the users' thoughts, which cannot be given by an expert [10]. Direct reactions of users and subjective data like emotions and opinions can be better analysed with empirical usability evaluation methods. The three main requirements on the method are:

1) As probands are not considered to have experience with the web-application they are testing, an exploration of the system must be possible within the evaluation.
2) Not only quantitative measures should emerge as result, but subjective measures like emotions and thoughts should be collected.
3) Captured data should be repeatedly retrievable so that the evaluator can precisely extract useful information.

Discussions and online surveys are not very useful for the purpose, because with both methods the evaluator cannot monitor the exploration of the web-application by the user. Therefore useful insights into subjective user experiences are not possible. Monitoring the user is considered one of the most valuable techniques in usability evaluation as unadulterated information will reach the evaluator [8, 9]. The think-aloud method is the technique where most comprehensive observation of the user is possible, as the researcher not only monitors the user, but the user also analyses himself. A more intense insight into problems as well as opinions and thoughts is possible that way [11, 8]. Also the usability researcher Nielsen considers this method the most valuable of all methods [12]. But field studies are considered to be more objective and more complex content can be evaluated with them [9]. This is why a combination of think-aloud-method and field study is most suitable. The next section will explain in detail how the two methods are combined.

3) Think-Aloud-Test Combined with Field Study: For combining the two methods think-aloud and field study, the user is visited in his familiar surrounding by the investigator. He is given a familiar working equipment, but is then asked to work through given tasks on a laptop and to think aloud

during the test. A recording of speech and image of the test person takes place via camera and microphone that are integrated in the laptop. This is better than the laboratory set-up, because the proband does not feel observed, but behaves more natural. In a first study with this set-up it could be noticed that most participants forget about the camera at all and behave quite natural [2]. The actions that the proband performs on the screen will be recorded with the help of a screen-capturing-tool.

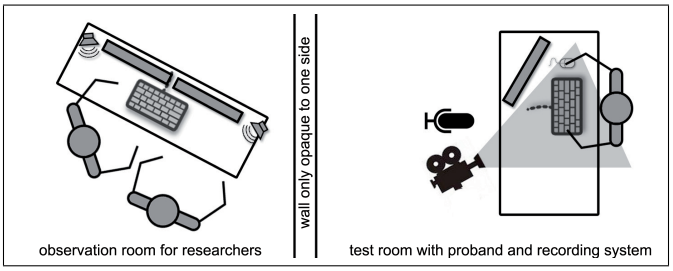

Figure 1. Test arrangement for the usability test with a standard usability laboratory

In the original set-up in a research laboratory, which can be seen in figure 1, the proband would sit on his own in one room with cameras and microphone recording his speech and actions. The investigators follows the exploration of the product in a second room by looking through the semi-permeable wall or following the screen and camera recordings on the monitor. No interaction between investigator and test persons is planned here.

The new approach rather follows the field study: the investigator sits next to the test person and an interaction is possible as can be seen in figure 2. The investigator might ask questions and the atmosphere is more relaxed than in a laboratory environment. This promotes a realistic behaviour of the proband.

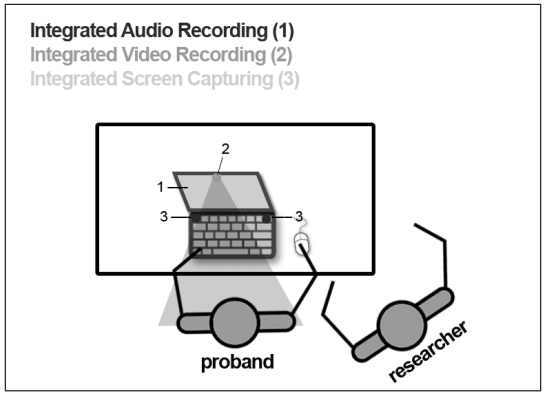

Figure 2. Test arrangement for the usability test with a mobile usability testing system

A possible hard- and software-set-up for the mobile usability testing environment is proposed in table I. This set-up was used in a first study [2] and found to be working stable. Another option is the usage of a mobile lecture recording system like tele-TASK [13]. This system records the screen of the presentation laptop, which is the laptop that the test person uses to fulfil the tasks in this case. The speech of the proband is recorded via wireless microphone as well as a video of the test person via external camera. The advantage is that the load of the recording is not on the test laptop so that also resource intense projects can be tested without loss of performance. Furthermore all recordings will automatically be synchronized and merged into one SMIL-file whereas the investigator has to synchronize the videos in the first set-up himself. The disadvantage is that a separate camera is used, which might make the test person feel more observed.

Attribute	Value
Laptop Model	Apple Macbook
Screen	13”, 1280 x 800, 16,7 mill. colours
Processor	2 GHz Intel Core 2 Duo
Memory	2 GB 1067 MHz DDR 3
Camera	integrated Camera
Recording Software for Camera and Sound	iMovie ’09
Recording Software for Screen Capturing	Snapz Pro X 2.1.5

Table I
SYSTEM ATTRIBUTES OF THE MOBILE USABILITY TESTING SYSTEMS

B. *Emotions as factor for the study*

As emotional aspects should be researched in this study, different techniques to do so will be compared in this paragraph. Emotions can be gathered and examined by three different methods.

The first is the measuring of physiological data like blood pressure and heartbeat. That method is not useful for this study as the measuring instruments might hinder the test user when executing the tasks and researchers without medical training would not be able to evaluate this data. [3]

The second is the self-disclosure of the proband. It can be achieved via a rating-scale, questionnaires or standardised lists of adjectives. Most of the self-disclosure techniques derive from clinical psychology and would therefore not support the goal of this study to find out mood and emotional tendencies that emerge due to the usage of a web-application. The mental state scale of Abele-Brehm and Brehm [14], an adjective list connected with a rating-scale, is more suitable since the attributes are chosen for psychologically healthy people. It is also a confirmed method [15]. According to Abele-Brehm and Brehm the mental state is a mixture of eight different aspects. It can be visualized in a circular model (figure 3). The rating scales are composed of eight sub-scales which represent the eight aspects. In the original test five items were selected for each of the sub-scales. Therefore it takes about eight minutes to fill out the

whole test. It is advisable to test the mental state after all tasks of the usability test in order to gain a more detailed insight into how the specific tasks affect the user. That is why it might be necessary to shorten the test dependent on the total time available.

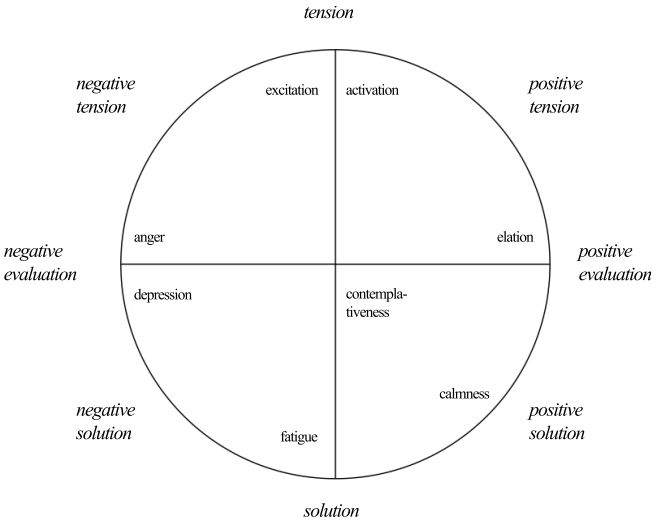

Figure 3. Circular Model of the Mental State [14]

A last method is the monitoring of the proband where a second person observes mimic and gesture. The face of the test person is recorded as video and the investigator might analyse extreme expressions. Usually test user and investigator evaluate the recording together to get a deeper insight into thoughts and emotions. This is not a good option for the combination with the think-aloud usability study, because the test itself already takes a long time and the evaluation of the whole video would needs equally long time. Therefore the investigator should mark interesting time spans during the test and only discuss those with the proband afterwards if necessary. A full evaluation will be conducted by the investigator alone. [3]

The next section explains how the analysis of emotions can be combined with the usability test.

C. Combination of methods

In order to receive a comprehensive overall picture of the situation of test persons and consider all before mentioned aspects a combination of several methods is useful [6, 12]. Therefore questionnaires will be used as supplement for the usability test, which is a proven method [12]. A pre-test questionnaire is used to capture demographical data as well as attitude and expectations towards the product or field tested, like for example online-shopping and 3D-applications in the internet. Post-task and post-test questionnaires are used to ask the test persons to evaluate the different parts of the web-site they have seen and how they experienced and understood it. An evaluation of opinions and the cognitive and conative aspects can be addressed with the think-aloud method as well as in the questionnaires. The mental state scale is then combined with the post-task tests. An interview might be conducted after the test as extension of the monitoring of the proband and his evaluations.

IV. Evaluation of the data

A. Evaluation of the data gained from the think-aloud method

The data gained with the think-aloud-method is the core part of the evaluation and will therefore be evaluated first. As for most methods in usability testing there is also no standardised and well documented way to analyse and evaluate data generated with the think-aloud-test. Because the goal of this combined method is not only to find out general usability problems, but to reason about the sensible adoption of new technologies and functions, an appropriate evaluation technique needs to be found. Since the think-aloud-method belongs to the qualitative research methods [12], evaluation techniques from that area are applied.

The first step is the transcription of actions and expressions of the probands from the video and audio recording to a tabular list. As the think-aloud-test generates large amounts of data, the investigator already filters the information and only lists the relevant ones. Next follows the coding of the data, either according to the grounded theory method of Corbin/Strauss [16] or the data analysis according to Chi [17]. Coding is used to reduce the amount of data, extract the most important actions and declarations and deduce general ideas from them [18].

It starts by allocating keywords (concepts) to phrases from the transcript. The coding entity needs to be determined beforehand. It is usually a sentence or several sentences that are connected as regards content. It is possible to work with a set of keywords defined beforehand or to work with an extensible set of keywords where keywords can be added as required. This phase is called open coding in grounded theory and phase of developing or choosing a coding scheme according to Chi. Next the keywords are grouped to categories and the categories are named. One option for the categories is the utilization of the user-based factors described earlier. Furthermore the connections between the keywords and categories are evaluated. This phase is similar to the selective coding in grounded theory and the search of patterns according to Chi [19].

The final phase includes detecting coherences between the categories and interpreting the results. The last phase can involve some interpretation whereas the first two steps should be conducted without valuation of the data. Mind-maps, representation of sets and cognitive maps can be used to visualize the results [20, 18, 19, 21].

B. Evaluation of the data gained from questionnaires

Questionnaires are a quantitative measuring instrument. But when they are combined with the think-aloud test usually not a sufficient number of items can be gathered. The results of the questionnaires are therefore no valid quantitative results, but should rather be used to support qualitative findings from the usability test and to find out about general tendencies. It is still helpful to visualize the data gained from the questionnaire in charts. If however a large enough number of probands can be tested, statistically significant results can be extracted from the questionnaires that can directly be integrated into the combined evaluation which will be explained in the next section.

C. Combined evaluation of all data

The data gained from the different methods can be combined when interpreting their results. The categories and concepts chosen when coding the data from the think-aloud-test should be used to structure the results according to different aspects. When interpreting the data in each concept, the starting point is the data from the usability test. Afterwards the findings from the mimic and gesture observation and the results of the questionnaires and mental state scales can be included into the analysis.

When all data is interpreted it is necessary to summarize the results in a form that allows an easy comparison of different approaches or an evaluation of the different aspects that were looked at in the test. A good overview is given, if the different categories are ranked and the results displayed in a table. If enough test persons were examined, the results from the mental state scale and the questionnaires can directly be used for that ranking. The findings from qualitative data have to be ranked by the researcher though, as no statistical values can serve as measure for it. To ensure the reliability of this ranking, the concept of inter-coder reliability needs to be considered. This means that a second person should process the ranking independent from the main researcher. The reliability of the ranking can be calculated with different ratio calculations like Scott's π or Cohen's κ [22]. Both calculate the correlation between two persons who independently allocate objects into predefined categories. The equations calculate the actual correlation by removing the random correlation.

V. Examples of the usage

This evaluation method was first used to test the potential of 3D-web-applications for electronic commerce web sites. Three different product categories were identified and one sample product per category tested in a user study. In the experiment three to four different product-presentations were evaluated for each sample product. Finally a mapping of suitable product presentations for each product category could be identified. E-Commerce is a suitable target field for this method, because all identified user-related factors that can be evaluated with this method are relevant for e-commerce as was shown in the study. [2]

A future project will be the testing of different functionalities, like varied search options or social web features, within a tele-teaching platform. These web-platforms first of all have to have a very good usability to support the users' learning process easily without that further learning how to use the platform is necessary. Also it is interesting to know which features make the learning more interesting or fun for students as it was proven that emotions do have an influence on the learning progress [23].

VI. Discussion and Conclusions

In this paper a user testing method for the combined evaluation of the usability and utility, cognitve, conative as well as emotional aspects was proposed. It was shown that the method was successfully utilized in a pilot test to identify suitable product presentations for different categories in e-commerce web-sites. [2] During this study some issues concerning the validity and reliability of the method were detected.

The think-aloud-method is an approved scientific method from cognitive psychology. Therefore its results are basically considered valid and reliable [24]. The constraints arise from the combination of qualitative with quantitative research methods. The think-aloud-method is a technique from qualitative research whereas the others, the questionnaire and the emotion test, are rather quantitative methods. The first approach usually only handles a small number of test persons, because the testing is very time-consuming. But the quantitative methods rather need large numbers of items for a valid and reliable result. Most validity and reliability would be gained through a larger number of participants in the test, because then the results can be regarded as statistically significant. The actual number depends on the goal and design of the test. If it is not possible to test more probands, all data gained from quantitative methods, like the emotions test and the questionnaires, may only be considered as a general tendency and not as valid statistical analysis.

Problematic constraints for the validity of the emotions test are interior factors of the probands like fatigue and interest in the product that is tested. An example would be the testing of product presentations for clothing in e-commerce web-sites. If the test person is not interested in that particular style of clothes offered in the test shops, the attentiveness might not stay fully with the test all the time and less positive emotions can be the result. The solution to that issue is again a larger number of probands and a variation of the sequence of the test tasks. These options will achieve a more valid result. Pre-test-questionnaires can also help in filtering test persons with similar interests and relation towards the product tested.

Another problematic issue is the decision to shorten the mental state scale in order to reduce the time the test

persons need. It is always a trade-off between the preciseness of the data and the time the test takes together with the exhaustion of the proband. This is the case because it was shown that the usage of slightly different items may result in totally different answers [25]. Therefore the utilization of several items for one aspect will result in more valid results. But a first usability test using a within-subjects test-design with this combination of methods [2] showed during observation of the probands that it is necessary to minimize the time spent on questionnaires and rating scales, because the mental state of the test persons will be affected negatively otherwise. To use an in-between-subjects test-design and more probands rather than an within-subjects test-design with less probands will reduce the problems, as the mental state scale will only have to be filled out at the beginning and the end.

Considering all these constraints the goal of a study has to be clearly defined beforehand. If merely a design-decision between several alternatives is required, a study with less test persons, a reduced mental state scale and rather qualitative results might be sufficient. But if significant statics are required for example for research purpose, more probands have to be tested and the mental state scale should be used in its original form.

Despite the constraints of this combined method, the first study conducted with it has proven that considering all different aspects of the users' perception is beneficial for a usability test. Because a more diverse point of view on the whole web-site is taken into consideration, the outcome of the study is more accurate. The researcher will furthermore get a deeper insight into the positive facets as well as the problems of an application and will therefore be able to precisely identify future points for improvements.

REFERENCES

[1] S. Sulzmaier, *E-Usability*, M. Beier and V. von Gizycki, Eds. Springer-Verlag Berlin Heidelberg 2002, 2002.

[2] F. Moritz, "Potentials of 3D-Web-Applications in E-Commerce - Study about the Impact of 3D-Product-Presentations," in *9th IEEE/ACIS International Conference on Computer and Information Science*. Yamagata, Japan: IEEE Computer Society, 2010.

[3] R. Mangold, "'Digitale Emotionen' - Wo bleiben die Gefühle bei Medialen Informationsangeboten?" *Hallische Medienarbeiten*, vol. 14, 2001.

[4] M. Hassenzahl, A. Beu, and M. Burmester, "Engineering joy," *IEEE Software*, vol. 18, pp. 70 – 76, 2001.

[5] H. Li, T. Daugherty, and F. Biocca, "Impact of 3-d advertising on product knowledge, brand attitude, and purchase intentions: The mediating role of presence," *Journal of Advertising*, vol. 31, 3, pp. 43 – 58, 2002.

[6] T. K. Hoppmann, "Examining the 'point of frustration' - the think-aloud method applied to online search tasks," *Quality and Quantity*, vol. 43, Number 2, pp. 211–224, 2009.

[7] S. Menon and B. Kahn, "Cross-category effects of induced arousal and pleasure on the internet shopping experience," *Journal of Retailing*, vol. 78, pp. 31–40, 2002.

[8] W. Schweibenz and F. Thissen, *Qualität im Web : Benutzerfreundliche Webseiten durch Usability Evaluation*, Springer-Verlag, Ed., Berlin Heidelberg New York, 2003.

[9] S. Stoessel, *Methoden des Testings im Usability Engineering*, M. Beier and V. von Gizycki, Eds. Springer-Verlag Berlin Heidelberg 2002, 2002.

[10] J. Wardag, *Design und Usability : Gegenüberstellung zweier Gestaltungsansätze im Webdesign*. VDM Verlag Dr. Müller, 2006.

[11] G. Gediga and K.-C. Hamborg, "Evaluation in der Software-Ergonomie: Methoden und Modelle im Software-Entwicklungsprozess," *Zeitschrift für Psychologie*, vol. 210(1), pp. 40 – 57, 2002.

[12] F. Sarodnick and H. Brau, *Methoden der Usability Evaluation*, E. Bamber, G. Mohr, and M. Rummel, Eds. Verlag Hans Huber, 2006.

[13] V. Schillings and C. Meinel, "Tele-TASK – tele-teaching anywhere solution kit," in *Proceedings of ACM SIGUCCS*, Providence, USA, 2002.

[14] A. Abele-Brehm and W. Brehm, "Zur Konzeptualisierung und Messung von Befindlichkeit. die Entwicklung der "Befindlichkeitsskalen" (BFS)," *Diagnostica*, pp. 209–228, 1986.

[15] N. Cours, "Wahrnehmungspsychologische Evaluation eines Dreidimensionalen Visualisierungssystems," Ph.D. dissertation, Universität Kassel, 2004.

[16] A. Corbin and J. Strauss, *J. Basics of Qualitative Research - Techniques and Procedures for Developing Grounded Theory*. Sage Publications,, 1998.

[17] M. T. H. Chi, "Quantifying qualitative analyses of verbal data: A practical guide," *The Journal of the Learning Sciences*, vol. 6(3), pp. 271–315, 1997.

[18] L. M. Given, Ed., *The SAGE Encyclopedia of QUALITATIVE RESEARCH METHODS*. SAGE Publications, Inc., 2008.

[19] M. Norgaard and K. Hornbaek, "What do usability evaluators do in practice? an explorative study of think-aloud testing," in *ACM Conference on Designing Interactive Systems*. ACM Press, 2006, pp. 209–218.

[20] T. Hynek, *User Experience Research - treibende Kraft der Designstrategie*, M. Beier and V. von Gizycki, Eds. Springer-Verlag Berlin Heidelberg 2002, 2002.

[21] S. Kvale, U. Flick, U. Gerhardt, P. Wiedemann, and D. Bühler-Niederberger, *Überprüfung und Verallgemeinerung*. Uwe Flick, 1995.

[22] C. Kuckein, "Datenschutz im Gesundheitswesen - Explorative Studie zur Sorgfalt von Ärzten im Umgang mit Patientendaten," diploma thesis, Technische Universität München, 2009.

[23] B. Kort, R. Reilly, and R. Picard, "An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion," in *International Conference on Advanced Learning Technologies*. IEEE Computer Society, 2001, pp. 43–48.

[24] E. Krahmer and N. Ummelen, "Thinking about thinking aloud: A comparison of two verbal protocols for usability testing," *IEEE Transactions on Professional Communication*, vol. 47, pp. 105–117, 2004.

[25] M. Gläser-Zikuda, *Emotionen und Lernstrategien in der Schule*. Beltz, 2001.