Influence of Training on Direct and Indirect Measures for the Evaluation of Multimodal Systems

Julia Seebode[1], *Stefan Schaffer*[1], *Ina Wechsung*[2] *and Florian Metze*[3]

[1] Research training group prometei, Berlin Institute of Technology, Germany
[2] Deutsche Telekom Laboratories, Berlin Institute of Technology, Germany
[3] School of Computer Science, Carnegie Mellon University, Pittsburgh, USA

julia.seebode@zmms.tu-berlin.de, stefan.schaffer@zmms.tu-berlin.de,
ina.wechsung@telekom.de, fmetze@cs.cmu.edu

Abstract

Finding suitable evaluation methods is an indispensable task during the development of new user interfaces, as no standardized approach has so far been established, especially for multimodal interfaces. In the current study, we used several data sources (direct and indirect measurements) to evaluate a multimodal version of an information system, tested on trained and untrained users. We investigated the extent to which the different types of data showed concordance concerning the perceived quality of the system, in order to derive clues as to the suitability of the respective evaluation methods. The aim was to examine, if widely used methods not originally developed for multimodal interfaces are appropriate under these conditions, and to derive new evaluation paradigms.

Index Terms: multimodal interfaces, usability evaluation methods

1. Introduction

Concerning multimodal user interfaces usability evaluation is still an object of active research. While several evaluation methods for unimodal applications have been standardized [1, 2], equivalent methods for multimodal systems so far have not been discovered. Further on it is still not clear, if standardized methods can be adapted for multimodal systems. Our earlier work showed that methods designed for unimodal interfaces should not be used as the only data source when evaluating multimodal systems [3].

Getting information about the perceived quality of a system is a difficult task, because one is not able to quantify a user's opinion about a system automatically. Thus the usability of a system is often measured through user ratings, collected by means of questionnaires. The parameters, obtained directly from the users, are called direct data.

However, questionnaires enclose many sources of unreliability and inaccuracy decreasing the quality of direct measures [4]. The ratings can be affected by individual differences between the users [5], whereas differences regarding the memory are likely to have a strong influence [6]. Walker [7] exemplified in his work about the PARADISE methodology that expert ratings could not be explained by a model built upon novice ratings. Further on Sturm [8] observed different ratings and diverse interaction patterns between trained and untrained users.

In addition to direct data technical evaluation is conducted to gather so called indirect data (e.g. dialogue duration, number of trials, modality usage). The system related parameters can be extracted from log files comprising information about the interaction with the system (e.g. timestamps, system state). As both kinds of measures try to quantify the usability of a system, high relations between them can be expected.

For example a high correlation between pragmatic quality and task duration could be found in one of our previous studies [9]. Similar results were reported by Sauro and Kindlund [10], who found positive correlation between satisfaction (direct data) and time, errors and task completion (indirect data).

However, several studies reported opposing findings: user ratings of multimodal interfaces were not affected by increased intuitivity and efficiency [11], no correlation between task duration and user judgements was found [12] and user's experience of the interaction (direct data) and indirect data differed considerably from each other or showed even negative correlations [13].

The mentioned studies indicate that both kinds of data should be explored in the evaluation of multimodal interfaces to obtain reliable results. Therefore the aim of the present paper is to analyze the relation between indirect and direct data and to investigate if the results are consistent to our earlier findings. In addition we try to explore the influence of training on direct and indirect measures and their relation.

2. Method

2.1. Participants and material

Thirty-one German-speaking individuals (16 male, 17 female) between the age of 22 and 39 (M = 27.06) took part in the study of a multimodal information system. This system was originally accessible as a system with touch screen only, and was extended with speech input. Fourteen of the participants were employees at Deutsche Telekom Laboratories and had attended a previous evaluation study with an earlier system version [14]. Thus they were familiar with the system and are considered as experts. The other seventeen participants were students paid for participating in this study. They are considered as novices as they that had no prior experience with the system.

The system we tested is a wall mounted room information and management system operable via a graphical user interface with touch screen, speech input and a combination of both. Cameras are used to detect faces in front of the screen and

activate speech recognition only if a person is allocated to the system. The output is always given via the graphical interface.

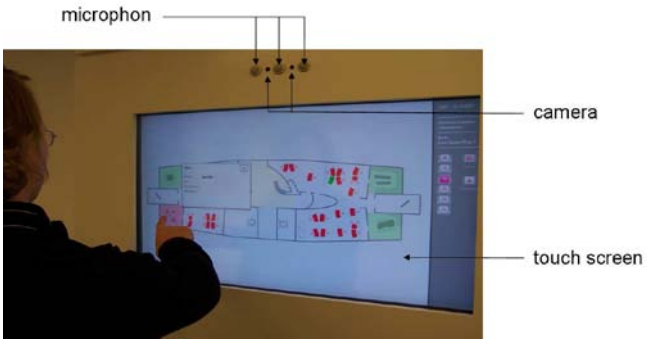

Figure 1: *User interacting with the system*

The AttrakDiff questionnaire [15] was used to collect user ratings about the system. This questionnaire contains four subscales that measure different aspects of perceived quality. The scale of pragmatic quality is used to measure terms of effort and efficiency and matches the classic concept of usability. In contrast to that, the two scales of hedonic qualities are not related with efficiency but with terms of joy. They distinguish between different aspects of hedonic qualities. The subscale of hedonic quality – stimulation was designed to measure whether a system can stimulate users with novel concepts. To measure how far users can identify with a system, the scale of hedonic quality – identity was developed. The scale of attractiveness is seen as a global scale that measures general satisfaction and is influenced by the other three subscales.

To collect indirect data, performance measures such as dialogue duration, number of trials and modality usage were recorded via log files.

2.2. Procedure

The participants had to perform six different tasks with the system, which are shown in Table 1. They were instructed to perform every interaction step for every single task with the modality they preferred.

Afterwards participants filled out the AttrakDiff questionnaire to rate this version of the system. The subscales for the AttrakDiff questionnaire were calculated according to the instructions in [15].

Beyond that audio, video and log data were recorded during the sessions. Dialogue duration was assessed via log files summed over all tasks and is seen as a measure of efficiency. To analyze which modality was used more by the participants, log data of the test was annotated. For every single interaction step of every task, the modality used to perform the step was logged. This way, the percentages of modality usage per task and over all tasks have been computed. Furthermore the number of trials has been counted for every task and was averaged over all participants. This data is also seen as a measure of efficiency.

Table 1. *Description of tasks and required interaction steps.*

Task	Description	Minimum # of interaction steps required	
		Speech	Touch
T1	Show main screen	1	1
T2	Show 18. floor	1	1
T3	Search for employee	3	6
T4	Search for room	2	-*
T5	Show event screen	1	1
T6	Show room for event	1	1

*If the room was accidentally booked at the time of the test, the task was solvable with two clicks. Otherwise a search overall rooms was necessary.

3. Results

3.1. Direct data - Questionnaire

Ratings on AttrakDiff questionnaire showed no differences between trained and untrained users: On the scale measuring pragmatic qualities the system got lowest ratings (M = 0.58, SD = 1.01). Best ratings were observed for the scale of hedonic quality – stimulation (M = 1.11, SD = 0.68). The detailed results for both groups are given in Table 2.

Table 2. *Ratings on AttrakDiff subscales.*

Scale	Group	Mean	SD	T(29)	p
Pragmatic Quality	Expert	0.57	1.12	.023	.982
	Novice	0.58	0.94		
Hedonic Quality - Stimulation	Expert	1.11	0.82	.012	.991
	Novice	1.11	0.58		
Hedonic Quality - Identity	Expert	0.79	0.65	.262	.795
	Novice	0.73	0.51		
Attractive-ness	Expert	0.96	0.79	.298	.767
	Novice	1.04	0.75		

3.2. Indirect data - Performance measures

3.2.1. Dialogue duration

Expert participants needed less time to solve the tasks (t(29)=1.867, p=.036). Means and standard deviation of the dialogue duration for trained and untrained users and for all participants in total are given in Table 3.

Table 3. *Overall dialogue duration.*

	Mean	SD
Expert	4:07 min	2:44 min
Novice	5:47 min	2:13 min
Total	5:02 min	2:33 min

3.2.2. Modality Usage

Speech as input modality (47.8%) was used less frequently than touch (52.2%) over all tasks. Detailed analysis showed that modality usage was strongly determined by task characteristics as can be seen in table 4. Especially Task T4 can be solved much more efficient (in terms of fewer interaction steps) by speech than via touch input and is therefore performed with speech input most frequently. For this task there is also a significant difference between experts and novice users. For all other tasks and overall tasks the results show, a numerical higher speech usage for the expert users.

Table 4. *Modality usage by tasks.*

Task	Group	Speech usage	T(29)	p
T1	Expert	71.4 %	0.503	.324
	Novice	62.8 %		
T2	Expert	52.4 %	0.584	.282
	Novice	42.4 %		
T3	Expert	40.0 %	0.269	.395
	Novice	36.9 %		
T4	Expert	74.0 %	2.648	.007
	Novice	51.1 %		
T5	Expert	47.6 %	0.935	.179
	Novice	33.4 %		
T6	Expert	41.1 %	0.895	.189
	Novice	27.6 %		
Total	Expert	54.4 %	1.106	.139
	Novice	42,3 %		

3.2.3. Number of trials

Over all tasks participants needed 3.55 trials on average. The six tasks differ in the level of difficulty and complexity as there were most trials needed to perform task T3 (7.74) and least trials to perform task T1 (1.37). Table 5 shows the number of trials for every task. Again for task T4 and additionally for task T6 expert users needed significantly fewer trials.

Table 5. *Number of trials by tasks.*

Task	Group	# of trials	T(29)	p
T1	Expert	1.43	0.279	.291
	Novice	1.29		
T2	Expert	2.00	0.110	.414
	Novice	2.12		
T3	Expert	8.21	0.326	.260
	Novice	7.35		
T4	Expert	4.64	0.987	.029
	Novice	6.82		
T5	Expert	1.64	0.065	.495
	Novice	1.65		
T6	Expert	2.21	0.953	.034
	Novice	3.18		
Total	Expert	3.36	0.478	.174
	Novice	3.74		

3.3. Correlation of direct and indirect measures

The results show some significant negative correlations between direct and indirect measures especially for the scale of pragmatic quality.

Users in general rated the pragmatic quality of the system to be worse when more time (Pearson's r= -.557, p = .001) and more trials (Pearson's r= -.402, p = .013) were needed to perform the tasks. The system's pragmatic quality was also rated highly significant lower by users who used speech as input modality more frequently (Pearson's r= -.516, p = .001).

Some differences between trained and untrained users were observed, as can be seen in table 6. For novices a negative correlation between the dialogue duration and the scale of attractiveness could be detected that was not seen for expert users.

Table 6. *Correlation (Pearson's r) of direct and indirect measures (*p<.01; **p<.05).*

Scale	Group	Dialogue duration	# of trials	Speech usage
Pragmatic Quality	Expert	- .627**	- .692**	- .626**
	Novice	- .548*	.020	- .422*
Attractive-ness	Expert	- .367	- .390	- .340
	Novice	- .556*	.180	- .067

4. Discussion

In this study we found some differences in the interaction patterns of trained and untrained users, as experts needed less time and fewer trials to solve the tasks and used speech more frequently as input modality than novice users did. According to [5, 6, 8], we also expected these two user groups to rate the system differently in the questionnaire. Since trained users could remember the interaction with the prior system versions, they performed better (in terms of efficiency) and could have rated the quality of the system to be higher. This expectation could not be proven as the judgments were similar for both

trained and untrained users. A possible explanation for this is that the experts had different expectations about the system and another level of motivation as they were not paid for participation. In addition the AttrakDiff as a semantic differential is a very uncommon form of questionnaire for many novice users whereas experts are more familiar with this kind of questionnaire. Thereby trained and untrained users might have had different response styles that were cleared by the different interaction patterns.

Another indication for this is the fact that correlations of direct and indirect measures show differences between expert and novice users. As in previous studies [3, 14], the AttrakDiff scale that measures pragmatic qualities showed a high agreement with the indirect measure of dialogue duration for trained and untrained users. In Addition this scale showed a high concordance with the usage of speech as input modality. Especially for trained users a high agreement of pragmatic qualities and the number of trials needed has been observed. By contrast the scale of attractiveness showed a concordance with dialogue duration only for the group of untrained users.

5. Conclusion

We see that trained users seem to rely on different measures when they answer a questionnaire to rate a system's quality. For untrained users especially data related to time has a big influence on the perceived quality. As this group of users is not familiar with the interaction they accept a higher number of trials to perform their tasks to get used to it.

In conclusion it can be said that the scale of pragmatic quality measures the construct it was developed for and is thus a reliable direct measure in terms of efficiency. But we have to keep in mind different groups of users with different interaction patterns that have an influence on their ratings. Especially for measures of hedonic qualities new approaches need to be found as there are no appropriate methods by now. In further studies for instance physiological techniques could be used to record additional indirect data to find suitable measures for hedonic qualities [16].

6. References

[1] Kirakowski, J. & Corbett, M.: SUMI: The Software Usability Measurement Inventory, In: British Journal of Educational Technology, 24(3), pp. 210–212, 1993.

[2] Hone, K. and Graham, R. (2001) Subjective assessment of speech-system interface usability, In: Proceedings of Eurospeech 2001, Volume 3, 2001.

[3] Naumann, A. & Wechsung, I.: Developing Usability Methods for Multimodal Systems: The Use of Subjective and Objective Measures. In: Proceedings of the International Workshop on Meaningful Measures: Valid Useful User Experience Measurement (VUUM), pp. 8-12, 2008.

[4] Annett, J.: Subjective Rating Scales: Science or Art? Ergonomics, 45, pp. 966-987, 2002.

[5] Krosnick, J. A.: Survey research. Annual Review of Psychology, 50, pp. 537-567, 1999.

[6] Engelbrecht, K. P., Möller, S., Schleicher, R. & Wechsung, I.: Analysis of PARADISE Models for Individual Users of a Spoken Dialog System, In: Elektronische Sprachsignal-verarbeitung. Tagungsband der 19. Konferenz, DE-Frankfurt a. M., A. Lacroix (ed.), TUDpress, DE-Dresden, 2008.

[7] Walker, M., Kamm, C. & Litman, D.: Towards developing general models of usability with PARADISE. In Natural Language Engineering, 6, pp. 363-377, Cambridge University Press, 2000.

[8] Sturm, J. A.: On the Usability of Multimodal Interaction for Mobile Access to Information Services, PhD Thesis, 2005.

[9] Wechsung I. & Naumann, A.: Established Usability Evaluation Methods for Multimodal Systems: A Comparison of Standardized Usability Questionnaires, In: Proceedings of PIT 08, pp 276-284, Heidelberg: Springer, in press.

[10] Sauro, J. & Kindlund, E.: A method to standardize usability metrics into a single score. In: Proceedings of CHI 2005, pp 401 – 409, ACM Press, 2005.

[11] Krämer, N. C. & Nitschke, J.: Ausgabemodalitäten im Vergleich: Verändern sie das Eingabeverhalten der Benutzer? [Output modalities in comparison: Do they change user's input behaviour?] In R. Marzi, V. Karavezyris, H.-H. Erbe & K.-P. Timpe (Hrsg.), Bedienen und Verstehen. 4. Berliner Werkstatt Mensch-Maschine-Systeme. Düsseldorf: VDI-Verlag, 2002.

[12] Möller, S.: Messung und Vorhersage der Effizienz bei der Interaktion mit Sprachdialogdiensten [Measuring and predicting efficiency for the interaction with speech dialogue systems], In S. Langer & W. Scholl (Eds.), Fortschritte der Akustik – DAGA 2006. DEGA, Berlin, 2006.

[13] Hornbæk, K. & Law, E. L.: Meta-analysis of correlations among usability measures, In: Proceedings of CHI 2007, pp. 617 – 626, ACM Press, 2007.

[14] Metze, F., Wechsung, I., Schaffer, S., Seebode, J. & Möller, S.: Reliable Evaluation of Multimodal Dialogue Systems. In: Proceedings of International Conference on Human-Computer Interaction, 2009 (to apear).

[15] Hassenzahl, M., Burmester, M., & Koller, F.: AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität [A questionnaire for measuring perceived hedonic and pragmatic quality]. In: J. Ziegler & G. Szwillus (Hrsg.), Mensch & Computer 2003. Interaktion in Bewegung . Stuttgart, Leipzig: B.G. Teubner, 2003.

[16] Mandryk, R. L., Inkpen, K. M., & Calvert, T. W.: Using psychophysiological techniques to measure user experience with entertainment technologies, In: Behaviour & Information Technology 2006, volume 25, pp. 141-158, 2006.