T-Labs Series in Telecommunication Services

Ina Wechsung

An Evaluation Framework for Multimodal Interaction

Determining Quality Aspects and Modality Choice

Springer

T-Labs Series in Telecommunication Services

For further volumes:
http://www.springer.com/series/10013

Ina Wechsung

An Evaluation Framework for Multimodal Interaction

Determining Quality Aspects and Modality Choice

Springer

Ina Wechsung
Quality and Usability Lab
TU Berlin
Berlin
Germany

Berlin, Technische Universität, Diss., 2013

ISSN 2192-2810 ISSN 2192-2829 (electronic)
ISBN 978-3-319-03809-4 ISBN 978-3-319-03810-0 (eBook)
DOI 10.1007/978-3-319-03810-0
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013955882

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Zusammenfassung

Zwar sind multimodale Systeme heutzutage weitverbreitet, doch sind nur wenige der etablierten Evaluationsmethoden auf diese Systemen zugeschnitten. Ein weitestgehend akzeptierter Evaluationsstandard ist nicht verfügbar. Zur Entwicklung eines einheitlichen Evaluationsansatzes multimodaler Systeme wird in der vorliegende Arbeit zunächst eine Taxonomie von Qualitätsaspekten multimodaler Interaktion vorgestellt. Obwohl viele der dargestellten Qualitätsaspekte mit etablierten Methoden erfassbar sind, ist zu beachten, dass die Mehrzahl dieser Methoden zur Evaluation grafischer, unimodaler Schnittstellen entwickelt wurde. Eine im Rahmen dieser Arbeit durchgeführte empirische Studie zeigte, dass diese Methoden nur begrenzt zur Evaluation multimodaler Schnittstellen geeignet sind. Basierend auf der eingangs vorgestellten Taxonomie wurde dementsprechend ein neues, auf multimodale Systeme zugeschnittenes, Evaluationsinstrument, der MMQQ Fragebogen, entwickelt und validiert. Parallel dazu, wurden die theoretischen Annahmen der Taxonomie mittels konfirmatorischer Kausalmodelle empirisch überprüft.

Im Anschluss wurde der Zusammenhang zwischen Qualitätswahrnehmungen einzelner Modalitäten und der Bewertung des Gesamtsystems untersucht. Vorrangiges Ziel war die Vorhersage der Qualitätsbewertungen des multimodalen Systems basierend auf den Bewertungen der einzelnen, unimodalen Komponenten. Es zeigte sich, dass insbesondere für Globalskalen, welche sowohl instrumentelle als auch nicht-instrumentelle Qualitätsaspekte erfassen, gute Schätzungen der wahrgenommenen Qualität erzielt werden können. Als wesentliche Einflussgröße für solche Vorhersagen stellten sich in diesem Rahmen die Nutzungsraten der einzelnen Modalitäten dar. Höhere Nutzungsraten einer Modalität gehen mit einem stärkeren Einfluss dieser Modalität auf die wahrgenommene Qualität des Gesamtsystems einher.

Entsprechend wurden im nächsten Schritt die Faktoren untersucht die diese Nutzungsraten beeinflussen. In vier empirischen Studien wurden Effizienz, situative Anforderungen verbunden mit der Allokation kognitiver Ressourcen, sowie individuelle Nutzercharakteristika als relevante Einflussfaktoren identifiziert.

Danach wurde der Frage nachgegangen, inwieweit sich Modalitätennutzung durch Qualitätsbewertungen und Interaktionsparameter vorhersagen lässt. Gemäß der Taxonomie von Qualitätsaspekten multimodaler Systeme sollte dies möglich sein, da davon ausgegangen wird, dass Qualitätsbewertungen und Interaktionsparameter ebenfalls von für die Modalitätennutzung relevanten Faktoren beeinflusst werden. Diese Annahmen wurden größtenteils bestätigt: Es zeigte sich ein direkter Effekt der wahrgenommen nicht-instrumentellen Qualität auf Modalitätennutzung.

Darüber hinaus zeigte sich mittels der vorhergesagten Nutzungsraten eine signifikante Verbesserung der Vorhersagen der Gesamtqualität aus den Beurteilungen der einzelnen Modalitäten, im Vergleich zu Baselines unter Annahme einer 50/50- bzw. 60/40-Verteilung der Modalitätennutzung.

Die wissenschaftlichen Beiträge diese Arbeit lassen sich wie folgt zusammenfassen: (1) eine ausführliche und empirisch validierte Taxonomie von Qualitätsaspekten multimodaler Systeme; (2) einen validierten, auf die Evaluation multimodaler Systeme zugeschnittenen Fragebogen; (3) die Untersuchung des Zusammenhang zwischen Qualitätswahrnehmungen einzelner Modalitäten und der Bewertung des Gesamtsystems; (4) die Identifikation und empirische Überprüfung von Faktoren welche die Modalitätennutzung beeinflussen; und (5) Modelle, die den Zusammenhang zwischen der wahrgenommenen Qualität einer Modalität und der tatsächliche Modalitätennutzung abbilden.

Summary

Although multimodal systems have entered the mass market, evaluation methods specifically tailored to multimodal systems are rather rare and a widely accepted standard method is not available. As a first step towards a unified evaluation approach, a unified framework, a taxonomy, of quality aspects of multimodal interaction is presented. For the assessment of the experience-related aspects, the vast majority of available methods have been developed for unimodal (predominantly GUI-based) systems. The most widespread evaluation method regarding experience-related factors are questionnaires, consequently four well-known and popular questionnaires, which were initially developed for unimodal systems, were investigated concerning their appropriateness for the evaluation of multimodal systems. The results of this study indicated, that these questionnaire are only partly applicable to multimodal systems.

Therefore, a new questionnaire, the MMQQ, which is specifically tailored to multimodal systems, was developed. The theoretical ground of the MMQQ is the taxonomy of quality aspects of multimodal systems suggested by Möller and colleagues (2009). In parallel to the questionnaire development, Möller's taxonomy was empirically validated and altered accordingly. This validation was achieved with the employment of confirmatory modelling approaches in addition to the exploratory approaches, which are usually employed in questionnaire development.

Next it was investigated how quality ratings of single modalities relate to the global evaluation of multimodal systems. The main intention was to examine if the quality perceptions of a multimodal systems can be predicted based on the quality perceptions of its constituent modalities. It was shown, that especially for overall scales, measuring both, pragmatic and hedonic qualities, a rough estimation of the multimodal system is possible (based on the quality ratings of the single modalities). Moreover, modality usage rates were observed to be central for such predictions. The more frequently a modality was used the higher was the modality's influence on the quality perceptions of the multimodal system.

Due to this observed importance of modality usage rates, the factors, which influence those rates, were addressed in four empirical studies. Efficiency, situational demands related to the allocation of cognitive resources, and user characteristics were identified as factors, which are relevant for modality choice.

Finally it was investigated if modality selection is predictable based on quality ratings and interaction parameters. According to the taxonomy presented previously, this should be possible: In the taxonomy it is assumed that all the factors identified as relevant for modality selection also influence quality ratings and interaction parameters. The results are partly in line with this assumption: Modality choice was found to be directly influenced by perceptions of a system's hedonic quality. While the influence of pragmatic qualities was not as prominent, it could further be shown that interaction parameters influence a system's perceived pragmatic qualities. Moreover, predictions of the quality of multimodal systems based on the ratings of its individual modalities are more accurate, if the predicted modality usage rates are used as weights in the regression equation, compared to baselines assuming 50/50 or 60/40 usage distributions.

In summary, this thesis presents (1) an exhaustive and empirically validated taxonomy of quality aspects of multimodal interaction as well as respective measurement methods, (2) a validated questionnaire specifically tailored to the evaluation of multimodal systems and covering most of the taxonomy's quality aspects, (3) insights on how the quality perceptions of multimodal systems relate to the quality perceptions of its individual components, (4) a set of empirically tested factors which influence modality choice, and (5) models regarding the relationship of the perceived quality of a modality and the actual usage of a modality.

Acknowledgements

I would like to thank everybody who supported me doing my research and writing this thesis.

Special thanks go to my principal supervisor Prof. Sebastian Möller, who has been a continuous influence on me and my work throughout the last years, on both an academic and a personal level. This thesis would not have been possible without his support and patience. Thank you for that!I also greatly appreciate that Prof. Markku Turunen agreed to be the co-examiner of my thesis.

A large part of the thesis was only possible due to the kind-heartedness of Prof. Christian Rietz, who travelled from Bonn to Berlin at his own expense to explain structural equation modelling to me. I am sure, that only few people would have done this.While the English in this thesis is probably far from perfect, it would have been much worse without the help of Dr. Stephen Wilson, who helped me to transform my “German English” into “English English”. You made writing everything up so much easier for me and due to your sense of humour also much more fun.

I would like to thank Dr. Robert Schleicher for listening to my endless questions and complaints about life, the universe and everything and suggesting the proper solutions often in the form of home truths, or if I was lucky, in the form of seriously funny one-liners.Dr.-Ing. Christine Kühnel patiently tolerated my many endless telephone conferences and the messiness of my desk (which sometimes spread to hers). She never lost her patience with me, even when I called her in the middle of the night and asked her to explain the difference between power spectral density and energy spectral density. I am very happy that we were colleagues once and are still friends now.

Although I had the great opportunity to visit many wonderful conferences, which were held at amazing locations all over the world, one of the most memorable conference trips was to Duisburg together with Dr.-Ing. Julia Niemann. With you, it is true: a joy shared is a joy doubled, and a problem shared is a problem halved.

Matthias Schulz selflessly helped in conducting one of the more complex user studies. I would not have been able to do this without his support.Julia Seebode and Stefan Schaffer allowed me to participate actively in their research, and were even so kind to adapt their test designs in order to gather the data I needed.Dr. Benjamin Weiss was so nice to use my questionnaire in one of his studies, which made it possible for me get a final validation data set. Florian Hinterleitner in particular (andChristine, Stephen, Julia, Klaus and Patrick as well) helped me so much with the preparations for my “last minute” exam in Speech Communication - you did an excellent job. Dr. Rahul Swaminathan, Lydia Kraus, Maija Poikela and Niklas Kirschnik helped me with fixing my defence talk. I still have to laugh when I remember that “it turns out that hand plus face equals handface”.

Many studies of this thesis were conducted within Telekom projects; I am grateful to Ralf Kirchherr, the project manager of most of the projects, for trusting in my choices regarding the design of the studies and the analysis of the results. Another colleague and friend from the “Telekom universe” is Kathrin Jepsen, although most of the work we did within the AUR&I project is not directly related to the research conducted within this thesis; the collaboration with you was a deeply satisfying experience. I am still a fan of quantitative research, but you reminded me how valuable qualitative data is, in order to have “good” products.

Hannah Bohle, Patrick Ehrenbrink, Artjom Sajatz and Katja Lauermann were great student workers – thank you for all your help in carrying out the experiments and the many hours you spent entering the data. It must have been so boring.

Thank you also to Irene Huber-Achter for her support regarding the dreadful organizational stuff and to our hotline guys, especially to Jan Binder, for fixing my computer and resetting my password numerous times.

Matthias Geier and Blazej Lewico made my life more pleasant as they (like me) never got tired of pizza for lunch. Other colleagues, who are not as crazy for pizza, but were equally pleasant company during the last years, are Dr.-Ing. Jens Ahrens, Jan-Niklas Antons,Marcus Berlin, Audubon Dougherty, Dr.-Ing. Marie-Neige Garcia,Stefan Hillmann,Tobias Hirsch,Ulrike Kehrberg, Dr. Hamed Ketabdar, Volker Presse, Tim Polzehl,Dr. Matthias Rath, Ulrike Stiefelhagen, and Dr.-Ing. Marcel Wältermann.

Even though, nowadays a long physical distance is between me and my long-time companions, Ulrike Dreyheller and Katrin Weck, the long distance calls with them put things into perspective.

Mama und Papa, Ihr seid die allerallerallerbesten Guten ;-). Ihr wart immer, wirklich immerfür mich da (und seid es hoffentlich auch weiterhin).Die Gewissheit Euch hinter mir zu haben und Eurer Vertrauen in mich, haben mir vieles leichter und fast alles erst möglich gemacht.Ihr habt mir nicht nur den ideellen Wert von Bildung vermittelt, sondern habt, als ob es selbstverständlich wäre (und das ist es nicht),auch den materiellen Preis dafür gezahlt.Ich werde Euch, Opa und Oma (mal 2), und Tommi und Eva nie genug für Eure bedingungslose Liebe und Unterstützung danken können.

And thank you to Dr.-Ing. Klaus Engelbrecht for reading the previous versions of this thesis, and for improving it with your many helpful comments (and commas). Apart from this, you motivated me and were there for me, even when I was sure I will never be able to finish. The best thing in T-Labs, and there have been many good things, was meeting you.

TABLE OF CONTENTS

1 Introduction

Multimodal systems have come a long way since Bolt presented his "Put-that-there-demonstrator" in the 1980s (Bolt, 1980). For a long time multimodal interfaces were of interest only to academics and industrial researchers, with the majority of commercial interactive systems allowing input only via the keyboard or through direct manipulation of devices such as a mouse, and offering only graphical output. Nevertheless, for at least the past 15 years additional interaction modalities such as speech have been part of commercial products, (for example, Microsoft's Speech Application Programming Interface was already included in Windows 95 (Shi & Maier, 1996)). However, although such additional interaction possibilities were available, the majority of users stuck to using the keyboard and mouse. Indeed, researchers at the time believed that the perceived shortcomings of speech-based interaction would be difficult to overcome. In 1998, Ben Shneiderman, a pioneer in human-computer interaction research and recipient of the ACM SIGCHI Lifetime Achievement Award, phrased his opinion on speech technology as follows: “Speech is the bicycle of user-interface design: It is great fun to use and has an important role, but it can carry only a light load. Sober advocates know that it will be tough to replace the automobile, graphical user interfaces.” (Shneiderman, 1998).

Despite being available as an interaction medium for many years, it is only relatively recently that speech technology has become widely popular among end-users (Geller, 2012). With reference to Shneiderman’s metaphor, it could be said that current state-of-the-art bikes seem to be able to carry a higher load and/or the automobile's capacities are getting smaller. One example in this development was the introduction of Apple`s voice-based program Siri in 2011. While Siri has certainly had its critics, and has been seen by some as a relative failure, and not as one of Apple’s major successes, a survey by market research and consulting company Park Associate found satisfaction levels among Siri users to be generally high, (70% of the 482 iPhone 4S owners surveyed reported they were either very satisfied or satisfied with the system) and levels of dissatisfaction to be relatively low, (only 9% of respondents in the same survey said they were dissatisfied), (Barrett & Jiang, 2012). This represents a large increase in satisfaction levels, when compared to a similar study of German end-users carried out by Peissner and colleagues (2006) who interviewed a representative sample of 1034 participants by phone. Of these, 420 had previously used a speech-based application with 32% reporting they were very satisfied or rather satisfied (*sehr zufrieden, eher zufrieden*), while 26% said they were very dissatisfied (*sehr unzufrieden*). Clearly, in the six years between both studies mentioned there has been a marked improvement in the speech-based interaction systems which has led to a rise in general levels of satisfaction with such systems (or at least with a subset of them) with a corresponding fall in general levels of dissatisfaction. The exact reasons for

Ina Wechsung, *An Evaluation Framework for Multimodal Interaction*,
T-Labs Series in Telecommunication Services,
DOI: 10.1007/978-3-319-03810-0_1, © Springer International Publishing Switzerland 2014

these increases and decreases remain to be established, and there is some disagreement in the field as to what they may be.

In a recent article, Geller (2012) interviewed experts from academia and industry in order to identify the possible factors, which have led to Siri,'s increased likeability compared to earlier systems. According to Roger K. Moore, editor in chief of the journal of Computer, Speech & Language and former president of the International Speech Communication Association, "the field of research hasn`t changed dramatically. What's new is that Siri's brought several complementary technologies together. Our business has been going for many years. Only now, with Siri, everybody knows about it." (cited after Geller, 2012). Moore implies that the success of Siri is not necessarily due to specific technological improvements by Apple, but rather that the power of the Apple brand allowed Siri to bring innovative speech technologies developed by researchers over the last number of decades to a much wider market.

As Alan W. Black, associate professor in the Language Technologies Institute at Carnegie Mellon University, points out, Siri would not have been possible 10 or 15 years ago due to the lack of available high-quality data necessary to train robust recognition systems. According to Black one of the main purposes of Google`s toll-free telephone-based 411-GOOG informational service "was finding out how ordinary people asked questions" (cited after Geller, 2012) with a view to gathering vast quantities of natural language for training purposes. In the past, collecting such training data was both labour- and resource-intensive. In Geller's article, the procedure of data collection, as it used to be, is described by Dan Faulkner, vice president of product and strategy for the Enterprise Business Unit at Nuance Communications. Faulkner is quoted as follows: "We'd pay people to come into the office and give them scripts. We'd give them a mobile phone, put them in a cab, tell them to call a number, then record their speech." (cited after Geller, 2012).

It is obvious that with such time-consuming and expensive procedures, databases used to be smaller. The enormous databases, which are now available, in conjunction with the widespread adoption of stochastic techniques in the field, have led to massive improvements in speech recognition technology (Hearst, 2011). Therefore, it is likely that recent gains in user satisfaction levels with respect to speech interaction systems are due to a combination of greater uptake of such systems by users as a result of marketing drives from Apple, as well as genuine technological advancements which have led to more reliable and robust systems.

In addition, according to Hearst (2011), another trend, which fostered the popularity of speech and weakened the dominance of the classical mouse/keyboard input has been the touch screen. Today, many mobile devices are both small in size and equipped with touch screens, a combination which may be rather unfavourable for long text inputs (Hearst, 2011). In such cases, speech input may be preferred as a faster and perhaps even more effective alternative. Therefore, to return to Shneiderman's analogy above, the automobile is actually becoming less convenient.

However, it should be emphasised that a successful interaction is not automatically a satisfying interaction. An issue mentioned by Alan Black: “One standard measure of spoken dialogue systems is task completion. Did the user successfully get the weather? But it’s clear that that’s not the only goal. You can have an interaction that’s successful and takes little time, but is unpleasant. So satisfaction is another goal. [Siri] doesn’t just answer questions. It has a character. It wants to name you, to know who you are. You can tell it to call you ‘Master’ or ‘Darth Vader’ or whatever, but it wants to call you that. It makes things a little more personal, and that’s important.” (cited after Geller, 2012). As per Geller’s (2012), both Alan Black and Dan Faulkner believe that Siri’s popularity is partly due to its characteristics beyond its task-based functionality. At this point, the questions arises as to what additional aspects should be evaluated if the standard measure of task success is not enough.

Moreover, with current multimodal devices users have both a bicycle (speech) and an automobile (touch). The work presented in this thesis predominantly investigates such systems that offer both speech and touch as input options, a choice that is mainly due to the fact that the studies carried out were conducted in a joint academia-industry setting , and the combination of speech and touch is currently a standard for modern (mobile) devices.

While the data above indicates that users like to have a bicycle (speech technology), it does not necessarily mean that they actually use it. They may just use it under specific circumstances, such as in a driving scenario, where touch input is impractical, or in other scenarios where speech is perceived as simply more fun.

This thesis addresses questions, which arise from the statements above:

- What aspects of system interaction should be evaluated, if standard measures like task success are not sufficient?
- Do standard evaluation methods cover these additional aspects and are they suitable for multimodal interfaces? If not, how should a new instrument be designed?
- How does the quality of the individual modalities relate to the quality of the system as a whole?
- Why do users select one modality over another? What are the factors, which determine modality selection?
- Is modality selection predictable?

This thesis is organised as follows. Chapter 2 presents concepts and theories that are central to the understanding of multimodal interaction. In addition, an overview of evaluation methods is given.

Chapter 3 answers the above question regarding what factors should be evaluated in order to determine a multimodal system’s quality. A taxonomy of quality aspects of such systems is presented.

Chapter 4 deals with the question of how and with which instrument, an evaluation can be carried out in accordance with the presented taxonomy. Standardized questionnaires are investigated concerning their suitability for this purpose. As a result of this, a new questionnaire, the MMQQ, is developed. In parallel, the taxonomy of quality aspects is empirically validated.

Questionnaires like the MMQQ are typically used in summative, global evaluations of interactive systems. With such global evaluations assessing the whole system, relatively little knowledge may be gained on how different modalities relate to each other, an issue, which will be tackled in Chapter 5. Here three studies are presented, which focus on the impact of the quality of the individual modalities on the quality of the multimodal system as a whole. One major finding is that more frequently used modalities have a higher influence on the quality perceptions of the multimodal system.

The factors, which lead to different usage frequency, are identified in Chapter 6, where the central research question examined is what determines modality selection strategies.

Finally in Chapter 7, it is investigated whether modality selection can be predicted based purely on quality ratings and interaction parameters.

Chapter 8 summarizes and discusses the findings of the previous chapters and closes the thesis with an outlook on future work. Please note that the book is partly based on material already published.

In the following, all publications used in this thesis are listed:

- Wechsung, I., K.-P. Engelbrecht, C. Kühnel, S. Möller & B. Weiss (2012): Measuring the Quality of Service and Quality of Experience of Multimodal Human-Machine Interaction Journal on Multimodal User Interfaces. In*: Journal on Multimodal User Interfaces 6* (1), 73–85.
- Wechsung, I., Engelbrecht, K.-P. and Möller, S. (2012). Using Quality Ratings to Predict Modality Choice in Multimodal Systems. *Proceedings of the 13th Annual Conference of the ISCA (Interspeech 2012)*. International Speech Communication Association (ISCA), 1-4.
- Wechsung, I., Schleicher, R. (2012). Modelling Modality Choice Using Task Parameters and Perceived Quality. *Proceedings of the ITG Conference on Speech Communication*, IEEE, 1-4.
- Wechsung, I., Schleicher, R. and Möller, S. (2011). How Context Determines Perceived Quality and Modality Choice. Secondary Task Paradigm Applied to the Evaluation of Multimodal Interfaces. *Proceedings IWSDS2011 Workshop on Paralinguistic Information and its Integration in Spoken Dialogue Systems*. Springer, 327-342.
- Wechsung, I., Schulz, M., Engelbrecht, K.-P., Niemann, J. and Möller, S. (2011). All Users Are (Not) Equal - The Influence of User Characteristics on Perceived Quality, Modality Choice and Performance. *Proceedings IWSDS2011 Workshop on Paralinguistic Information and its Integration in Spoken Dialogue Systems*. Springer, 175-188.

- Wechsung, I., Schaffer, S., Schleicher, R., Naumann, A. and Möller, S. (2010). The Influence of Expertise and Efficiency on Modality Selection Strategies and Perceived Mental Effort. *Proceedings of the 11th Annual Conference of the ISCA (Interspeech 2010).* International Speech Communication Association (ISCA), 1930-1933.
- Wechsung, I., Engelbrecht, K.-P., Naumann, A., Möller, S., Schaffer, S. and Schleicher, R. (2010). Investigating Modality Selection Strategies. *Proceedings of* the *IEEE workshop on spoken language technology SLT 2010*
- Wechsung, I., Engelbrecht, K.-P., Schaffer, S., Seebode, J., Metze, F. and Möller, S. (2009). Usability Evaluation of Multimodal Interfaces: Is the whole the sum of its parts?. *Proceedings of the 13th International Conference on Human-Computer Interaction (HCI International 2009), Part 2.* Springer, 113–119.
- Wechsung, I., Engelbrecht, K.-P., Naumann, A., Schaffer, S., Seebode, J., Metze, F. and Möller, S. (2009). Predicting the Quality of Multimodal Systems Based on Judgements of Single Modalities. *Proceedings of the 10th Annual Conference of the ISCA (Interspeech 2009).* International Speech Communication Association (ISCA), 1827-1830.
- Naumann, A., Wechsung, I. and Hurtienne, J. (2010). Multimodal Interaction: A Suitable Strategy for Including Older Users?. *Interacting with Computers*, 465-474.
- Wechsung, I., Naumann, A. and Hurtienne, J. (2009). Multimodale Interaktion: Intuitiv, robust, bevorzugt und altersgerecht? [Multimodal Interaction: Intuitive, Preferred and Senior-Friendly?] . *Proceedings of Mensch & Computer 2009. 9. fachübergreifende Konferenz für interaktive und koooperative Medien - Grenzenlos frei 2009.* Oldenbourg, 213-222.
- Möller, S., Engelbrecht, K.-P., Kühnel, C., Wechsung, I. and Weiss, B. (2009). A Taxonomy of Quality of Service and Quality of Experience of Multimodal Human-Machine Interaction. *Proceedings of the First International Workshop on Quality of Multimedia Experience (QoMEX'09)*
- Wechsung, I., Naumann, A. and Möller, S. (2008). Multimodale Anwendungen: Einflüsse auf die Wahl der Modalität [Multimodal Applications - Factors Influencing Modality Choice]. *Proceedings of Mensch & Computer 2008: 8. fachübergreifende Konferenz für interaktive und koooperative Medien - Viel Mehr Interaktion.* Oldenbourg Wissenschaftsverlag, 437-440.
- Wechsung, I. and Naumann, A. (2008). Evaluation Methods for Multimodal Systems: A Comparison of Standardized Usability Questionnaires. *Proceedings of Perception in Multimodal Dialogue Systems, 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems, PIT 2008.* Springer, 276-284.
- Naumann, A. and Wechsung, I. (2008). Developing Usability Methods for Multimodal Systems: The Use of Subjective and Objective Measures. *Proceedings of the*

International Workshop on Meaningful Measures: Valid Useful User Experience Measurement (VUUM) 2008. Institute of Research in Informatics of Toulouse (IRIT), 8–12.

2 What Are Multimodal Systems? Why Do They Need Evaluation? - Theoretical Background

The following section introduces relevant concepts and definitions. After that, the cognitive foundations of multimodal interaction are briefly described. Next current evaluation methods are introduced and reviewed with respect to their appropriateness regarding multimodal systems.

2.1 Modality, Media and Multimodal Systems - Definitions and Terminology

This sub-section will discuss several conceptualisations of the term *modality.* On this basis a definition of multimodal systems is adopted. This definition is used throughout the remainder of this thesis.

Several definitions of modality can be found in the relevant literature. These can be broadly categorized into three general groups: Physiological/human-centred definitions, technology/system-centred definitions, and definitions incorporating both views.

The term modality has its origins in physiology and is, according to Charwat (1992), defined as

> "perception via one of the three perceptual channels. You can distinguish the three modalities: visual, auditive, and tactile (physiology of senses)."

Thus, the three senses sight, hearing, and touch, correspond to the three perceptual channels. Thereby the terms visual and auditive refer to the perception and the sensory modalities; the terms optical and acoustical refer to physical (and not physiological) parameters (Schomaker et al., 1995). According to Charwat's (1992) definition only three different modalities respectively three different human senses can be distinguished. Although the aforementioned senses are nowadays those with the highest relevance for human-computer-interaction (HCI), at least three more senses (smell, vestibular, taste) are defined in physiology.

Another definition of modality is offered by Bernsen (2008):

> "Modality is a way of representing information in some physical medium. Thus, a modality is defined by its physical medium and its particular "way" of representation."

With this definition, Bernsen (2008) moves away from the physiological understanding of modality: Modalities according to Bernsen (2008) refer to *ways of information representation* rather than to the human sense. Thus, he broadens the term modality: Humans use many ways to represent information and these different ways of information representation may refer to the same sensory modality (e.g. images and text are different ways of representation but both refer to vision). A multimodal system in Bernsen's sense is a system employing at least two different modalities ("ways of information representation") for input and/or output. In contrast, Bernsen defines a

Ina Wechsung, *An Evaluation Framework for Multimodal Interaction*,
T-Labs Series in Telecommunication Services,
DOI: 10.1007/978-3-319-03810-0_2, © Springer International Publishing Switzerland 2014

unimodal system as a system using the same modality for input and output. This conceptualization of modalities and multimodal systems is insofar rather unconventional, as devices offering a traditional graphical user interface (GUI) only, are multimodal since they offer texts and graphic as output and haptics as input. An example of a unimodal system in Bernsen's sense is a system offering only spoken language as input and output.

Another system-oriented conceptualisation of modality is presented by Nigay and Coutaz (1995). They too expand the strict physiological view but unlike Bernsen, they focus on the *interaction technique* rather than on the way of information representation. They posit a modality as the combination of an *interaction language (L)* and a *physical input or output device (d)*, which can be formalized as a tupel <*d, L*>. Examples for interaction languages, as proposed by Nigay and Coutaz (1995), are direct manipulation, gestures, or pseudo natural speech. Thus interaction modalities on a smart-phone could be <*touchscreen, gestures*> or <*microphone, speech*>.

Yet another definition is given by Oviatt (2002) defining a multimodal system as systems processing

> "two or more combined user input modes— such as speech, pen, touch, manual gestures, gaze, and head and body movements— in a coordinated manner with multimedia system output. [….] Such interfaces eventually will interpret continuous input from a large number of different visual, auditory, and tactile input modes […]".

Thus, in Oviatt's definition modalities refer to *input modes*, in contrast to the above mentioned definitions which explicitly refer to the term modality. However, as no guidance is given on what defines an input mode, this definition is rather unhelpful but lends itself as a good example for a strict technology-driven definition of multimodality, which is not expanding the strict physiological view but neglecting it. According to Baber (2001), only the combination of the technical system-oriented view, (which focuses on interaction techniques, input device and output devices) and the user-oriented view (which focuses on human perception), will be useful to investigate multimodal human machine interaction. Typically, a multimodal system employs different interaction techniques and a user needs to have different sensory modalities to interact with such a systems. However, most of the definitions, as those presented exemplarily above, focus either on one or the other perspective.

Möller et al. (2009) take both views into account, stating that

> "multimodal dialogue systems are systems which enable human-machine interaction through a number of media, making use of different sensory channels."

The understanding of the term *media* in the scientific community is, in contrast to the term modality, mostly uniform. Media is associated with the physical realization respectively presentation of information via input and output devices (cf e.g. Bernsen, 1997; Gibbon, Moore, & Winski, 1998; Hovy & Arens, 1990; Jokinen & Raike, 2003; Sturm, 2005).

Möller et al. (2009) elaborate there definition further as follows:

> "These channels may be used sequentially or in parallel, and they may provide complementary or redundant information to the user."

With this part of the definition, Möller et al. (2009) refer to Nigay and Coutaz (1993) who developed a design space for multimodal systems along the three dimensions (1) *level of abstraction*, (2) *usage of modalities* and (3) *fusion*. Level of abstraction represents the technical level, on which information of the different input and output devices is processed. Speech input may be processed as a signal, a sequence of phonemes or as parsed sentences bearing meaning. Usage of modalities means the temporal availability of the modalities: While some systems allow for parallel usage, other systems only offer sequential interaction. The third dimension, fusion, describes if and how the information of the different modalities is combined. Based on the design space Nigay and Coutaz (1993) identified four different types of multimodal systems:

- **Exclusive interactions:** The system offers different modalities but usage is only sequential (one modality at one time). Fusion is absent.
- **Alternate interactions:** The system offers different modalities. Like for exclusive interactions, the modalities can only be used sequential but they can be related to each other. Fusion, the combination of the possible input data, is implemented.
- **Simultaneous interactions:** The modalities can be used in parallel (simultaneously). Fusion is absent.
- **Synergistic interactions:** The modalities can be used in parallel and the information can be related to each other.

In the following the term *multimodal system* is used as defined by Möller et al. (2009). The systems used as material in the presented studies offered just sequential input and rudimentary fusion modules, as a result for the most part only exclusive interactions were possible.

2.2 Cognitive Foundations of Multimodal Interaction and Assumed Advantages of Multimodal Systems

By providing multiple communication channels, multimodal systems are assumed to support human information processing by using different cognitive resources. This assumption is largely based on cognitive theories postulating multiple, modality-specific processing resources (e.g.; Baddeley & Hitch, 1974; Baddely, 2003; Paivio, 1986; Wickens, 1984, 2002).

According to the working memory theory proposed by Baddeley, different types of information refer to different cognitive resources. Baddeley's working memory model includes three components: the *central executive*, the *visual-spatial sketchpad* and the *phonological loop*. Hence, the short-term storage of visual-spatial information (e.g. colors and shape) is the visual-spatial sketchpad, whereas the phonological loop is the short term storage for auditory-verbal data. It has to be noted, that visual information (e.g. written text) might be recoded into verbal information by sub-vocal articulation and consequently be stored in the phonological loop. The central executive is the controlling unit, monitoring and where required adjusting thinking processes and actions. Later, a fourth component, the *episodic buffer*, was added. The *episodic buffer* is a multimodal component and represents an interface to the long-term memory.

Also, regarding the long term memory, the coding of information into mental representations is assumed to be modality specific to a large extent. *Dual coding theory* by Paivio (1986) postulates two largely independent cognitive systems: the *imaginal, non-verbal systems* and the *verbal systems*. As Baddeley's phonological loop, the verbal system processes verbal information whereas the imaginal system, analogue to the visual-spatial sketchpad, processes visual-spatial information. According to Paivio (1986), findings of neuro-psychology support these assumptions. It was shown, that dependent on the type of information (verbal vs. spatial) different brain areas are active. Still, these systems are connected. This explains why multimodal presentations, e.g. verbal-auditory paired with visual information, can be superior to unimodal presentations. The dual coding leads to higher recognition and recall performance.

A third modality specific theory is the *Multiple Resource Theory* (*MRT*) by Wickens (2002, cf. Section 3.2.1). This theory proposes three different *processing stages*, two *different response codes*, two different *perceptual modalities* and two different *input codes*. Moreover, two *visual channels* are suggested. In contrast to the theories mentioned above, Wickens does not only differentiate between the information processing modalities but also between sensory modalities. For each of the different stages, modalities and response codes, different cognitive resources are assumed. MRT predicts that tasks accessing the same resources are very difficult to be performed in parallel. Or the other way around: Timesharing, the splitting of attention between two tasks, is easier when the necessary information is presented via two modalities instead of one modality.

In summary the redundant information presentation and the splitting of information to several channels reduces the overall cognitive load experienced by the user. With lower cognitive load, errors are less likely and the interaction gets more robust (Qvarfort, 2004).

Additionally to a more robust interaction, multimodality may also enhance a system's flexibility, its naturalness, and its efficiency (Hedicke, 2000; Höllerer, 2002; Oviatt, 1999; Qvarfort, 2004). With a higher degree of freedom the user is free to choose his/her preferred interaction modality with regards to situation, task and context. This higher flexibility is assumed to increase the systems inclusiveness (Jokinen & Raike, 2003) and efficiency; however, regarding the latter also the opposite is reported (Sturm, 2005). The hypothesis regarding naturalness stems from the observation of human-human communication being multimodal, e.g. verbal human communication has a visual part through lip-reading, mimicry or gestures (Schomaker et al., 1995). Due to the possibilities to use richer natural languages and new flexible ways of interaction, multimodality has the potential to realize the s*ystem-as-an-agent* metaphor proposed by Jokinen (2009). Jokinen describes such agents as interaction partners mediating between the user and the application, rather than as a tool that is used to perform certain tasks. Consequently, the gulf between user and system can be minimized by adapting the system to the user's natural characteristics (Norman, 1986).

While empirical findings support the above assumptions - multimodal systems have indeed been shown to be more natural, more efficient, more reliable and more robust (e.g. Oviatt, 1996; Oviatt et al., 2000, Cohen, McGee, & Clow, 2000; Burke et al., 2006) - it has to be noted that these benefits are not an inherent property of all multimodal systems. Oviatt (1999) points out that all these advantages are mediated through the design of the interface and the usage context; multimodality was shown to be especially benefical in situation with high workload and high task complexity (Oviatt, Coulston & Lunsford, 2004). Moreover, considering the above presented MRT (Wickens, 2000) a multimodal system can also be inferior compared to a unimodal system, e.g. if additional speech input is necessary in verbally demanding situations like in air traffic control or call centres. In addition, a higher cognitive load due to more degrees of freedom may occur (Schomaker et al., 1995). Furthermore, the different modalities may interfere with each other (Schomaker et al., 1995): When presenting identical information via two modalities (e.g. reading and listening to the same text simultaneously) a synchronization problem can arise (Schnotz, Bannert, & Seufert, 2002). Additionally, if different modalities refer to the same cognitive resources, task performance may even decrease (Oviatt, 1996). Thus, it is not surprising that it has been shown, that making a system multimodal by just adding a further modality to a unimodal system may not necessarily lead to an improvement (Oviatt, 1999). Consequently, evaluation of the interface is an indispensable issue.

2.3 Quality and Usability Evaluation Methods

The following section will give an overview of established and well-known evaluation methods. Advantages and disadvantages regarding their appropriateness and suitability for multimodal systems will be reviewed. Please note, that in the literature all the described methods are often labelled as *usability* evaluation methods rather than as *quality* evaluation methods. While those constructs are partly overlapping, they are not identical. Hence, before introducing the evaluation methods, the relationship between quality and usability will be discussed.

2.3.1 Quality vs. Usability

Most of the described methods were developed to measure usability in the narrow sense as described by the International Organization for Standardization (ISO) with the standard 9241-11 (ISO 9241-11, 1998). Here usability is defined by the three factors efficiency, effectiveness, and satisfaction. While user satisfaction is mentioned in this standard, the focus of early usability evaluation was focused on the efficiency and effectiveness of the system. For instance in a meta-analysis by Hornbæk and Law (2007), user satisfaction was found to be the usability factor which was assessed least frequently. However, in this thesis the user-experienced satisfaction is considered an essential part of usability (cf. Section 3.3). Usability itself is one important, but not the only aspect of quality. Quality, as understood in this thesis, is the result of the user's appraisal of the perceived capabilities of the system to support the user's individual goals.

The first part of this conceptualisation is based on the definition of Jekosch (2000, as cited in Möller, 2005). It implies that quality is an inherently "subjective" concept; it is a result of the user's individual perceptual and judgemental processes. Please note that, Jekosch (2000) original definition suggests that users appraise the perceived entity in comparison with a desired entity. This part of the definition was not adopted as it implies a rather resource-extensive cognitive process involving mental comparison of the features of the perceived and the desired system. Findings from cognitive psychology imply that the brain is rather lazy and avoids resource-intensive processing: Judgements are often biased and are based on heuristics or on intuition (Kahneman, 2003). Moreover, the original definition indicates that the user knows how the desired system should be. This is also debatable: While users may know their goals, they may not know how exactly a systems needs to be designed in order to fulfil those goals. Thus, the definition by Jekosch may apply to the quality of speech and voice signals, the context were it has been developed for, but not for rather complex multimodal systems.

Hence, the second part is based on Hassenzahl's work (Hassenzahl & Roto, 2007). Here, quality is the related to the fulfilment individual goals. Those goals can be either *do-goals,* for example "making a phone call", or *be-goals,* "being related to somebody". Do-goals are derived from the higher-level be-goals (Hassenzahl & Roto, 2007). For example, missing somebody may lead to the desire to communicate this person. Making the phone call is than the do-goal, the feeling of being related to this person is the be-goal

2.3.2 Evaluation Methods

With usability being understood as one of many quality aspects, usability evaluations can be considered as a subgroup of quality evaluations.

According to Preece et al. (1994) evaluation methods can be distinguished using five different criteria. The first distinction refers to the question addressed with the evaluation and comprises four different categories. Evaluation studies may be conducted,

- to see if the system is good enough,
- to compare two alternative and see if one system is better than another one,
- to get the system closer to the real world,
- to see how well the system is working in the real world or,
- to see if the systems complies to certain standards.

Depending on the question addressed with the evaluation, it has to be decided in which stage of the development cycle the evaluation should take place. Here, *formative, process-oriented evaluation* can be distinguished from *summative, goal-oriented evaluation*. Formative evaluation can already take place in early development cycles without a prototype and aims to improve the system as part of the iterative design process. For summative evaluation, an advanced prototype is necessary, as summative evaluation is typically carried out to assess the quality of a late version of the system.

Another distinction criterion is the level of user involvement with *user-centred, empirical methods* on the one side, and *expert-centred, analytical-formal methods* one the other side. Especially formative evaluations are often conducted in early phases of the system development without users. Elements of the interface and their consequences are analysed and modelled by experts. Consequently, neither users nor a running prototype are necessary. These methods are often also labelled as *inspections* (Holzinger, 2005). User-centred, empirical methods are methods, which are observing and “measuring” user’s reactions towards the interface. Measurements collected in this manner, are assumed to represent the system’s quality. A prototype, with which the user can interact, is at least to a certain extent necessary (Sturm, 2005). Another term for such methods is *testing* (Holzinger, 2005).

The type of data collected can be *qualitative* and *quantitative data*. Quantitative data are numerical, abstract data. Abstract means that such numbers do not directly represents the meaning of the measured date (Witt, 2001). Typically, this kind of data is analysed using statistical methods. Examples are questionnaires or time measurements. Qualitative data cannot be quantified in numbers, and its analysis is usually interpretative as applying statistical methods is not possible. However, it is often possible to transform qualitative data into quantitative data. For instance, free text answers are usually qualitative data, but they can be converted into quantitative data by first analysing them with respect to the opinion stated in the text. Then the number of positive, negative and neutral answers can be counted, the counts are forming a quantitative data set which can be analysed statistically.

In the context of usability and quality evaluation, *direct* and *indirect* measurements can be distinguished (Seebode et al., 2009, Möller et al., 2009). Direct measurements are assessed directly from the user and are a direct representation of the quality as perceived by the user. Indirect measurements refers to interaction parameters or psychophysiological parameters, these kinds of measurements cannot be interpreted as direct assessments of perceived quality or perceived usability – in the best case such data is correlated with the quality perceptions but might as well be unrelated to the user's judgement (c.f. Hornbæk & Law, 2007; Naumann & Wechsung, 2008). Additional characteristics to categorize evaluation methods are according to Dix et al. (1993):

- **The style of the evaluation.** It refers to the setting, laboratory or field of the study. While lab studies offer a more controlled setting eliminating interfering variables, field studies lead to a higher naturalness.
- **The level of information**. It describes how abstract the gathered information is. Low level information is very specific, e.g. if the wording of a specific prompt is understandable. High level information is more general for instance if the system is usable.
- **The immediacy of the answer.** It describes whether the data is assessed during or after the interaction, the latter possibly being influenced by memory biases.
- **The intrusiveness of the answer.** This characteristic is directly related to immediacy, as asking questions during the interaction is rather intrusive and might affect the user`s behaviour.
- **The resources required.** Resources comprise factors like time, money, effort, equipment, and manpower.

In Table 2.1 (adapted from Dix et al., 1993) established methods are categorized, based partly on the categories by Dix et al. (1993) and on additional own categories like advantages and disadvantages and appropriateness for different modalities.

In the following, established methods are described and discussed regarding their suitability for the evaluation of multimodal systems. At first, expert-centred analytical-formal methods are presented, followed by user-centred empirical methods.

Table 2.1. Classification of different evaluation methods

		Cognitive Walkthrough	Heuristic Evaluation	Model-Based Evaluation	Experiment	Interviews, Questionnaires	Protocol Analysis, Thinking Aloud (TA)
Stage of development process		Throughout	Throughout	Design	Throughout	Throughout	After prototyping
User involvement		No	No	No	Yes	Yes	Yes
Style		Laboratory	Laboratory	Laboratory	Laboratory	Laboratory/field	Laboratory/field
Level of information		Low	High	Low	Low and high	High	Low and high
Data type		Qualitative	Qualitative	Quantitative	Qualitative /quantitative	Qualitative /quantitative	Qualitative
Is method obtrusive?		No	No	No	Yes	No	Yes
Amount of required resources	**Time**	Medium	Low	Medium	High	Low	High
	Material	Low	Low	Low	High	Low	Low to high
	Expertise	High	Medium	High	High	Low	Medium to high
Task-oriented?		Yes	No	Typically yes	Yes and no	Yes and no	Yes and no
Appropriateness for different modalities		Auditory, visual and haptic/ limited for multimodal systems	Auditory, visual and haptic/ limited for multimodal systems	Auditory, visual and haptic/ limited for multimodal systems	Auditory, visual and haptic and multimodal systems	Auditory, visual and haptic and multimodal	Auditory visual and haptic/ limited for multimodal systems TA is not useful for auditory modality
Main disadvantages		Only „assumed“ problems, high expertise necessary	Only „assumed“ problems, high expertise necessary, appropriate heuristics are often not available	High expertise	High expertise necessary	Development of questionnaires is very resource-consuming	TA is often difficult for users and not appropriate for speech input Working prototype is necessary
Main advantages		Low resources required	Low resources required	Low resources required	Data is generalizable	Easy to use	Data is generated by real users

Cognitive Walkthrough

The *Cognitive Walkthrough* is a task-based, expert-centred, analytical method (Holzinger, 2005) based on explorative learning and problem solving theory (Wharton et al., 1994). It takes into account that user often learn the interface by exploring it, instead of reading the manual.

Experts, usually designers or psychologist, analyse the functionalities of the interface based on a description of the system, a description of the tasks the end user will carry out, a list of actions necessary to perform the tasks, and a description of user and usage context (Wharton, et al. 1994). Critical information is recorded by the experts using a standardized protocol. The procedure itself involves the following five steps (Wharton et al., 1994):

- Definition of inputs for the walkthrough (e.g. identifying the users, defining the tasks to evaluate, describing the interface in detail)
- Calling in the experts
- Analysing the action sequences for each task
- Protocolling critical information
- Revising the interface

The biggest advantages of the Cognitive Walkthrough are, as for almost all formative-analytical methods, that end users as well as an implemented system are not necessary. Disadvantages are the quite low level of information, only the ease of learning is investigated (Wharton et al., 1994). Moreover, a Cognitive Walkthrough might be very time consuming for complex systems. As multimodal systems are usually more complex than unimodal systems, due to more degrees of freedom offered by multiple modalities, the Cognitive Walkthrough, in its classical form, is rather unattractive for such systems. Moreover, the Cognitive Walkthrough is strictly task-based and will only be able to evaluate the ease-of-use of an interface rather than its joy-of-use.

Heuristic Evaluation

The term "heuristic" is derived from the Greek "heureskein" and means "to find" or "to explore" something (Holzinger, 2005). *Heuristic Evaluation* is a method of the so-called *Discount Usability Engineering*, a resource conserving, pragmatic approach proposed by Nielsen, aiming to overcome the argumentation that usability evaluation is too expensive, too difficult and too time consuming.

In a Heuristic Evaluation, several experts check if the interface complies with certain usability principles (heuristics). To ensure an independent, unbiased judgement of every evaluator, they do not communicate to find an aggregated judgement until each of them investigated the interface on his/her own (Nielsen, 1994). Result of a Heuristic Evaluation is a list of usability problems and the respective explanations. Additionally, problems might be judged according to their frequency and pertinence. According to Nielsen, three to five experts will find 60-70 % of the problems, with no improvements for more than ten evaluators (Nielsen, 1994). However, this statement has repeatedly caused disputes; research provides support (Virzi, 1992) as well as contrary findings (Woolrych & Cockton ,1992; Spool & Schroeder, 2001).

The Heuristic Evaluation is a cheap and quick to apply method and can be conducted throughout the whole development cycle (Holzinger, 2005). However, to the authors' knowledge, established usability heuristics tailored to multimodal systems do not exist.

Review-Based Evaluation

For a *Review-Based Evaluation,* existing experimental findings and principles are employed to provide a judgment. Relevant literature has to be analysed in order to approve or disapprove the design of the interface (Dix et al., 1993). Hereby, the context of the respective studies has to be carefully considered. To prevent a confirmatory bias not only the similarities, but also the dissimilarities between the interface to be evaluated and the studies serving as a basis for the evaluation have to be taken into account (Gerhard, 2003).

Review-Based Evaluation is faster and more economical than conducting an own experiment. But wrong conclusions might be drawn if the selection of the considered studies is not done with the required prudence. Additionally, the vast majority of studies are addressing very specific problems, making it difficult to generalize the results to other interfaces and vice versa, a specific interface is difficult to evaluate with the results of another specific interface (Gerhard, 2003).

Model-Based Evaluation

For *Model-Based Evaluation*, on a very general level, two different approaches can be distinguished. The first approach has its origin in cognitive psychology and focuses on the cognitive process while interacting with an interface; the other approach is rooted in the engineering domains and is focusing on the prediction of user behaviour patterns. Within both approaches, user models are employed for the predictions.

Methods of the first approach are usually addressing low-level parameters like task execution time, memory processes or cognitive load (cf. Engelbrecht, Quade, Möller, 2009) and are largely bottom-up oriented. Starting point to define user models are theories and findings from cognitive psychology. Examples are the methods GOMS (Goals, Operator, Methods, Selection rules; Card, Newell & Moran, 1983), the Cognitive Complexity Theory (CCT) by Kieras and Polson (1985), or ACT-R (Adaptive Control of Thought–Rational) by Anderson and his group. (e.g. Anderson & Lebiere, 1998).

With the method GOMS, the interaction with a system is reduced to basic elements, which are *goals, methods, operators* and *selection rules*. Goals are descriptions of the goals or sub-goals of the user, what he/she intends while using the system. Operators are the actions offered by the system to accomplish these goals. Methods are well-learned sequences of sub-goals and operators suitable to achieve a goal (John & Kieras 1996). Selections rules apply if several methods are possible and reflect the user's personal preferences. These four basic elements describe the procedural knowledge necessary to perform the tasks. This knowledge is applied to the design, to check if the system provides methods for all user goals; furthermore execution times of well-trained, error-free expert users can be predicted (John & Kieras, 1996).

In case of multimodal systems, GOMS analyses can become quite extensive due to the complexity of such systems. As multimodal systems allow for parallel, serial or combined usage of different modalities, multiple methods for one goal are possible, because of this the definition of multiple selection rules is required. The EPIC framework by (Kieras & Meyer, 1997) is a more sophisticated architecture better suitable for predicting execution times for interactions with multimodal systems, however, EPIC is first and foremost a research system and thus not focused on being a tool for evaluation purposes (Kieras & Meyer, 1997).

As all these cognitive models are grounded well in theory, they provide useful insights in user behaviour. Although cognitive modelling is an active research field, so far it has not been received particularly well by usability practitioners and only rarely finds its way into non-academic evaluations (Engelbrecht, Quade, Möller, 2009; Kieras, 2003). Reasons are their often-high complexity (Kieras, 2003) and possibly the aforementioned low level of the information possible to gain with cognitive modelling.

Thus the engineering-based, statistically-driven approach attempts to provide more high level information, e.g. if the user is "satisfied" with the system, and therefore rather utilizes top-down strategies. Here, user models are usually defined based on real user data and are not necessarily linked to cognitive theories (cf. Engelbrecht, Quade, Möller, 2009). Most of these methods and algorithms were developed for spoken dialogue systems, with PARADISE (Paradigm for Dialogue System Evaluation; Walker et al., 1997) likely being the most widespread one. PARADISE uses linear regression to predict user satisfaction based on interaction parameters such as task success or task duration. Other approaches are the MeMo (Mental Models) workbench using a probabilistic model of user behaviour, which includes a rule engine derived from empirical data in order to predict user behaviour (Engelbrecht, 2012). But like cognitive models, the engineering models are rather of academic interest than a widespread usability evaluation method amongst practitioners. Even though Model-Based Evaluation, like all the expert-centred methods above, can be conducted in very early design stages and is cheaper than testing with real users. The main disadvantages might be the very high expertise necessary (Holzinger, 2005) and, regarding multimodal interaction, the lack of theories to use for Model-Based Evaluation, as these processes are not well understood so far. For instance, the factors determining why users choose one modality over another have just recently been identified (cf. Chapter 6).

A model-based approach explicitly tailored to multimodal system is the PROMISE framework by Beringer and colleagues (2002), an extension of Walker's PARADISE. However, studies applying PROMISE are very seldom, possibly because some of the parameters are relatively ill defined (e.g. the way of interaction), and it is not specified how they should be assessed. Just recently (Kühnel, 2012) proposed a well-defined set of interaction parameters for multimodal interaction yielding reasonable prediction performance (>50% accuracy) for user judgements.

Protocol Analyses and Thinking Aloud

For *Protocol Analyses*, user behaviour is captured using video, audio and log-files. A prominent method is *Thinking Aloud*. Here participants are asked to verbalize and loudly utter their thoughts (Holzinger, 2005). This might be done during the interaction or after the interaction as retrospective Thinking Aloud. For the latter, the user is confronted with video recordings of the test session and is asked to comment on them. Although retrospective Thinking Aloud is less intrusive than online Thinking Aloud, it might probably be affected by memory biases. Another version is the plural Thinking Aloud involving multiple participants using the system together. Therewith, the unnaturalness of this method should be reduced. Hackman and Biers (1992) confirmed that Thinking Aloud in double-teams is beneficial, yielding better results compared to single user Thinking Aloud.

This method can be used for free exploration of the interface as well as for conducting concrete tasks. Though it is often perceived as unnatural and confusing (Lin, Choong & Salvendy, 1997) and often the experimenter has to repeatedly advise the participants to actually think aloud as the constant verbalising can be difficult and effortful for the users (Hegner, 2003). Further problems are the systematic biases due to social desirability. Moreover, for interfaces offering speech input, the non-retrospective version of this method is inappropriate, as Thinking Aloud and speaking to the system simultaneously is not possible (Hegner, 2003).

Experiments

Experimental evaluation investigates specific aspects of the interaction behaviour under controlled conditions (Sturm, 2005). In the simplest experimental design, one hypothesis is formulated and two experimental conditions, differing only regarding the manipulated factor to be investigated (independent variable), are set-up (Dix, et al. 1993). All differences occurring in the measured variables (dependent variable) are attributed to the manipulations of the independent variable (Dix, et al. 1993). Experiments allow collecting high quality data as interfering variables are controlled and/or eliminated. Experiments provide, if carried out carefully, causal inference. Thus, experiments are essential to establish and verify theories (Hegner, 2003); accordingly, experiments are a useful method for evaluation of multimodal interfaces. However, with experiments being strongly controlled, user behaviour might be rather unnatural (Sturm, 2005). In the worst case, results can be an experimental artefact. Another drawback is the high amount of resources required to set up and conduct a proper experiment.

Interviews and Questionnaires

Questionnaire and interviews are indispensable to measure the users' judgements of the system (Holzinger, 2005) as interaction data will not necessarily reflect the users' perceptions (e.g. Naumann & Wechsung, 2008). Interviews and questionnaires are often used to assess user satisfaction, emotions or attitudes towards the system. If reliable questionnaires or interviews are available, they are relatively cheap and easy to employ. Thus, the probably most common technique applied in user-centred evaluations are questionnaires, but a standardized and validated questionnaire addressing the evaluation of multimodal systems is still not available. Even for speech-based systems the probably most common questionnaire, the Subjective Assessment of Speech Interfaces questionnaire (SASSI; Hone & Graham, 2000), still lacks final psychometric validation. Please note that the SASSI is not suitable for spoken dialogue systems, as only input but not output quality is assessed. However, the ITU-T Rec. P.851, the ITU's recommendation regarding quality evaluation of telephone-based spoken dialogue systems (ITU-T Rec. P.851, 2003), proposes an extended version of the SASSI, including items covering output quality (Möller, Engelbrecht, & Schleicher, 2008).

As standardized, well-validated questionnaires tailored to multimodal system are rare, self-made questionnaires or questionnaires developed for unimodal systems are often employed. Both approaches are problematic: Self-constructed questionnaires are usually not properly validated (Larsen, 2003a) and questionnaires developed for unimodal systems may not provide valid and reliable results for multimodal systems. A detailed comparison of relevant questionnaires is presented in Chapter 4.

A notable exception is the SUXES method presented by (Turunen et al., 2009), which is, as explicitly stated by the authors, addressing multimodal systems. SUXES aims to measure user expectation and user experience with different pre- and post-test questionnaires. Constructs measured with SUXES are speed, pleasantness, clearness, error free use, robustness, learning curve, naturalness, usefulness, and future use. It is based on the SERVQUAL method (Parasuraman, Zeithaml, & Berry, 1988), initially developed to asses perceived quality of service in service and retailing companies. Even though, the original publication of the SERVQUAL method includes psychometric validation, for SUXES no such data is available. Hence, the reliability and validity of the constructs measured with SUXES is not confirmed as the constructs measured with SUXES (see above) do not match the constructs assessed with SERVQUAL (Reliability, Assurance, Tangibles, Empathy, Responsiveness). Moreover, SERVQUAL itself has been heavily criticized one a conceptual as well as on a methodological level (Buttle, 1996; Nyeck et al. 2002). Major critique points, also highly relevant for SUXES, are related to the measurement of expectations before interacting with the system (Buttle, 1996). For example several authors (e.g. Kahneman & Miller, 1986; for a comprehensive discussion see Buttle, 1996) state, that expectations are formed after interacting with a system or service and not before. Additionally, expectations may be affected by a social desirability bias as user may want to comply with the "I have high expectations" social norm (Buttle, 1996). Moreover, asking for expectation may induce expectation, which would not have been relevant without questioning for them. Additionally, costumers tend to adapt their expectation to their actual experience. Thus, if applying SUXES experimenters need to keep in mind that asking for expectations may alter them and that participants may not have expectations before the usage of a system. Furthermore, a psychometric validation of SUXES is necessary to ensure if the constructs, which admittedly have high face validity, are actually statistically reliable and valid.

2.4 Chapter Summary

Although multimodal systems have been around for more than 25 years now and besides the rapidly increasing technical developments in this area, evaluation methods and design guidelines are still rare and evaluation of multimodal systems is considered as problematic (Jokinen, 2008). Even though the HCI literature provides a wide range of evaluation most of them were developed to assess unimodal graphical user interfaces and will not necessarily be useful for multimodal systems. As mentioned in the respective sections, most of the established methods presented above, are not instantly usable for multimodal evaluation.

The formal-analytical methods need theories of multimodal interaction, which are just emerging by now (Kühnel, 2012). The empirical methods lack the measurement instruments (e.g. questionnaires). Thus, it is not surprising that most studies evaluating multimodal interfaces employ empirical methods, using either self-constructed questionnaires, which are not or only little validated, or adapt standardized questionnaires, which were initially developed for GUI-based interfaces, and which are also not validated for multimodal systems (e.g. Bauckhage et al., 2002; Baillie, et al., 2002; Bernsen & Dybkjaer, 2004; Bornträger, et al., 2003; Damianos, et al., 2000; DeAngeli et al. 1998, Hemsen, 2004; Höllerer, 2002; Qvarfordt, Jönsson, &Dahlbäck, 2003; Sturm, 2005).

Consequently, the constructs measured are quite diverse and the results are hardly comparable. Thus, the first step towards a unified evaluation approach for multimodal interaction is a unified framework of quality aspects of multimodal interaction. The following chapter will present such a framework based on the work of Möller et al.

(2009) and Wechsung et al. (2012a), and identify measurements methods for each aspect.

3 What to Evaluate? – A Taxonomy of Quality Aspects of Multimodal Interfaces

Apart from the lack of evaluation methods specifically addressing multimodal interfaces, it is often not defined, which aspects need to be taken into consideration when evaluating a multimodal systems quality. This is partly due to the characteristic of multimodal interaction, making it necessary to measure additional concepts like the appropriateness of modality and context or interference between different modalities. Apart from these multimodality specific issues, also new topics, like emotions in the context of HCI and hedonic qualities of user interfaces, made it obvious that the traditional concept of usability might not be sufficient to ensure quality or user satisfaction.

The taxonomy depicted in Figure 3.1 aims to offer an overview of the relevant concepts and currently available measurement methods. It is based on Möller et al. (2009) und Wechsung et al. (2012a) and presents an empirically tested modification. The empirical test of the taxonomy is presented in Chapter 4. The next sections explain all concepts within the taxonomy.

Ina Wechsung, *An Evaluation Framework for Multimodal Interaction*,
T-Labs Series in Telecommunication Services,
DOI: 10.1007/978-3-319-03810-0_3, © Springer International Publishing Switzerland 2014

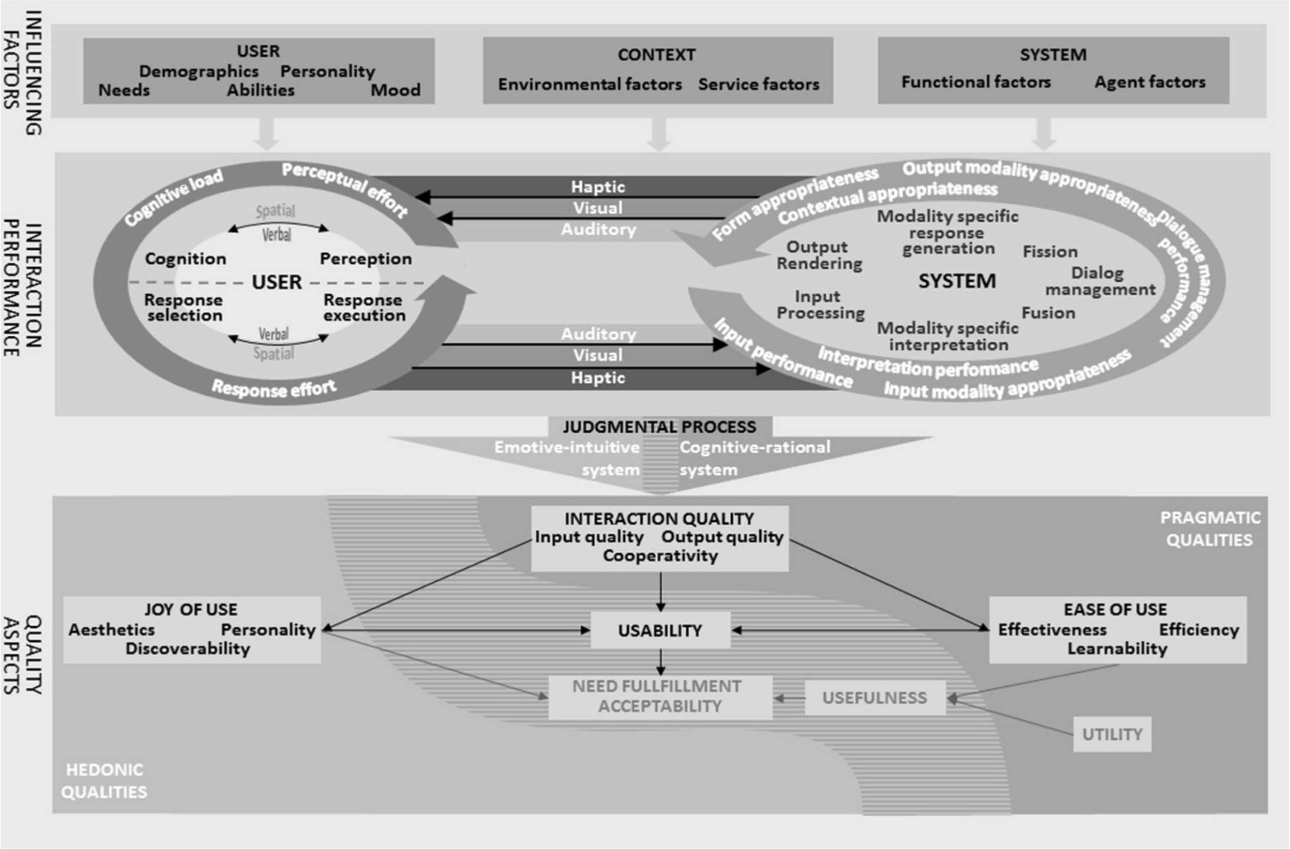


Fig. 3.1. Taxonomy of multimodal quality aspects

3.1 Influencing Factors

The first layer of the taxonomy comprises the influencing variables, namely *user characteristics*, *context* and *system* (cf. Figure 3.2).

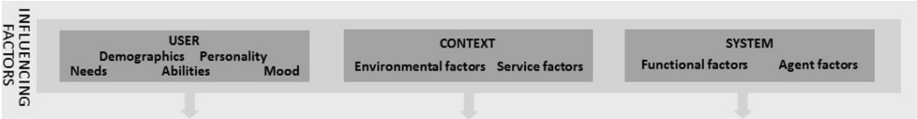


Fig. 3.2. First layer of the taxonomy – the influencing factors.

3.1.1 User Characteristics

User characteristics are the users' abilities, his/her personality, demographic variables, his/her mood, and the user's needs.

Abilities

Abilities refer to *perceptual-cognitive abilities* and *knowledge* as well as to *motor* and *physical capabilities*. Especially the impact of spatial cognitive abilities on performance has extensively been studied (for an overview see Dillon & Watson, 1996). According to Chen and colleagues (Chen, Czerwinski, & Macredie, 2000), people with high spatial abilities perform better with spatially oriented or graphic interfaces than people with low spatial abilities. In the context of multimodal interaction, such users might prefer the GUI over speech control.

Considerably less research can be found for other cognitive abilities or other interaction modalities (e.g. speech). Dillon and Watson (1996) found recall of the interaction with a spoken dialogue system to be influenced by information processing speed with users showing a lower processing speed being less likely to remember all relevant information provided by the system. Additionally, working memory span was shown to influence recall and transfer performance with a low span especially disadvantageous in a mobile instructional environment (Doolittle, Terry, & Mariano, 2009).

In addition, prior knowledge and familiarity are relevant factors for interaction behaviour: Generally, prior knowledge is reported to improve performance (Lewis, Langdon, & Clarkson 2007). Regarding multimodal systems, modality choice and preference are related to familiarity (Naumann, Wechsung, & Möller, 2008): The more familiar modality is being used more frequently than the less familiar one.

A variety of methods can be found to assess perceptual-cognitive abilities, including clinical intelligence test like the Wechsler Adult Intelligence Scale (WAIS; Wechsler, 2008). Such tests typically include sub-scales; for instance, the WAIS IV addresses perceptual abilities with the Perceptual Reasoning Index and memory capabilities with the Working Memory Index.

Validated instruments for the assessment of HCI-related abilities and knowledge are presented by Smith et al. (Smith, Caputi, & Rawstorne, 2007) and Van Vliet et al. (Van Vliet, Kletke, & Chakraborty, 1994). The instrument by Smith et al. measures computer experience, Van Vliet et al. developed a questionnaire for assessing computer literacy.

For motor capabilities the AMPS (Assessment of Motor and Process Skills) by Fisher (2004) can be employed.

Personality

Personality includes personality variables, like psychological personality traits, e.g. the so-called Big Five (Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism), and attitudes.

The influence of personality traits on interaction behaviour and judgements is documented in a study by Burnett and Ditsikas (2006), who compared introverted and extroverted user. They reported that extroverted users found more usability problems and gave slightly worse ratings on a post-hoc usability questionnaire. For attitudes, Jawahar and Elango (2001) observed, that performance is positively affected by positive attitudes towards technology. Moreover, attitudes are believed to influence also users' quality perceptions with positive attitudes resulting in better ratings (Angel, Hartmann, & Sutcliffe 2009). Additionally, positive attitudes towards technologies make adoption of new technologies more likely (Matilla, Karjaluoto, & Pento, 2003). Likewise, when using multimodal systems people may tend to use the more innovative modality more often, if they have positive attitudes towards technology.

To assess personality variables, psychometric questionnaires are available: They can be measured with the NEO-FFI (Costa & McCrae, 1992) or with the briefer Big Five Inventory (John, Donahue, & Kentle, 1991). Both questionnaires are also available in a short version (Rammstedt, & John, 2007).

For attitudes towards computers, Richter et al. (Richter, Naumann, & Groeben 2000) proposed a questionnaire, and also the technical affinity questionnaire by Karrer and colleagues (Karrer et al., 2009) includes one sub-scale for assessing positive attitudes and one sub-scale for assessing negative attitudes towards electronic devices.

Demographics

Age and gender differences are probably the demographic variables most often assessed in evaluation studies. Gender effects are documented for performance as well as for quality judgements (Morgan et al., 1994), but results are contradictory: Some studies found gender effects not to be moderated by other factors like previous experience (Brown et al. 1997, Vollmeyer, Imhof, & Beierlein, 2006); other studies report that gender effects are largely moderated by such factors (Whitley, 1996), and some studies show no or only small gender effects (Weiss et al., 2010).

For age, it is often assumed that not the chronological age "per se" causes differences, but rather characteristics like a smaller degree of previous experience (Chalmers, 2003), the age-related decrease of cognitive abilities (Wolters et al., 2010) and motor impairments (Carmichael, 1999).

While previous experience and cognitive abilities might explain differences in performance and some aspects of interaction behaviour, quality perceptions are less likely to be direct correlates of those aspects. As an exemption, a low memory span may result in forgetting relevant aspects of the interaction, and consequently in inconsistent ratings (Wechsung et al., 2009). Age effects were indeed shown for quality ratings but these might as well be moderated by other variables, for instance attitudes or performance. Moreover results are contradictory: Wolters and colleagues (2010) found older adults to be more critical than younger users, while Naumann and colleagues (Naumann, Wechsung, & Hurtienne, 2010) observed the opposite: Older users tend to rate systems more positive than younger users although their performance is worse.

Besides age and gender, many more demographic variables exist, e.g. cultural background, profession, and level of education. While results regarding the influence of such variables are inconsistent, they should be measured as they may explain outliers or differences in interaction and rating behaviour.

Demographics are usually easy to assess with simple, self-constructed questionnaires presented before, or after the test.

Mood

Moods are the affective quality of experiences, constantly experienced but often only sporadically reflected consciously (Morris, 1989; Silvia, & Warburton, 2006). Compared to emotions, moods are lacking objects, are psychologically diffuse, relatively long lasting, and are structured simply; moreover, they are not differentiated by patterns of appraisal (Silvia & Warburton, 2006).

Research suggests that people in a good mood are, compared to people in a bad mood, more likely to employ less elaborate information processing strategies like heuristics (Bless et al., 1996, for an overview see Schwarz & Clore, 2003). Thus, their information processing might lack logical consistency. For evaluative judgments this means, that evaluation is probably more context-driven than content-driven (Bless et al, 1996). More specifically in terms of usability studies, people in a positive mood might give less exact evaluation ratings, as positive mood is associated with less attention given to details and less information being considered. Moreover, they may be more influenced by the contextual factors e.g. the setting and scenario.

Additionally, memory recall is mood congruent: good moods make recall of positive experiences more likely than bad moods and vice versa (Kahneman, 1999).

A lightweight instrument to measure mood is the so-called Faces Scale (Andrews & Whitey, 1976). This non-verbal scale shows seven faces ranging from "very sad" to "very happy". Participants are asked to indicate the face, which matches their current mood best. Another questionnaire is the Brief Mood Introspection Scale (BMIS), developed by Mayer and Gaschke (1988). The BMIS consists of 16 items (plus one global item) and measures mood on four sub-scales (Pleasant-Unpleasant Mood, Arousal-Calm Mood, Positive-Tired Mood, Negative-Relaxed Mood).

Needs

According to Hassenzahl and colleagues (Hassenzahl, Diefenbach, & Göritz, 2010), the main motivation to use interactive technologies is the fulfilment of psychological needs. The most salient needs in the context of human-computer-interaction have been identified as the needs for stimulation, relatedness, competence and popularity (Hassenzahl, Diefenbach, & Göritz, 2010). For example, a very bored person may ask a speech dialogue system "stupid" questions in order to fulfil the need for stimulation. Please note, that psychological needs do not match biological-physiological needs such as hunger or thirst. However, like biological needs psychological needs are assumed to be largely invariant across human beings (Sheldon et al., 2001). Of course, the level of fulfilment and deprivation of each need changes constantly. A questionnaire to assess the level of experienced need fulfilment can be found in Hassenzahl et al. (Hassenzahl, Diefenbach & Göritz, 2012). It is an adapted version of a questionnaire, originally developed by Sheldon and colleagues in the context of personality psychology (Sheldon et al., 2001).

3.1.2 Context

Physical Environment

Here, the physical conditions in which the interaction takes place are meant (Möller et al. 2009). Examples are the lighting and acoustic conditions as well as the place and location, or parallel tasks like driving a car or walking down the street activities.

Service Factors

These factors refer to the availability of the system, privacy and security concerns and the associated costs (Möller et al., 2009).

3.1.3 System Characteristics

According to Möller et al. (2009), system factors are the *agent factors*, i.e. the characteristics of the system as an interaction partner, and the *functional factors*, i.e. its functional capabilities.

Agent Factors

Those factors comprise the technical characteristics (e.g. multimodal fusion, dialogue management, multimodal fission) of the interaction agent as well as its appearance (Möller et al., 2009). The term *agent* refers to the "system-as-an-agent-metaphor" described in (Jokinen & Raike, 2003). Following this metaphor, a system is not predominantly seen as a tool but as partner or participant, the user interacts with. According to Jokinen and Raike (2003), especially multimodal interfaces offer a possibility to realize this metaphor, as they enable the user to employ various types of input and output modalities and are thus more conversational in nature.

Functional Factors

Functional factors are the functional properties of the system, e.g. the number of available functions and their complexity, structure and usage frequency (cf. Möller, 2006). These system factors are typically defined in the earliest development stages as part of the requirement analyses (Mayhew, 1999; Möller et al., 2009). The characteristics are then listed qualitatively in specification documents. Agent factors are defined by the system developer together with design experts. Functional factors may be derived by early concept tests or focus groups (Möller et al., 2009).

3.2 Interaction Performance

This layer, depicted in Figure 3.3, describes the behaviour of and processes within the user and the system during interaction.

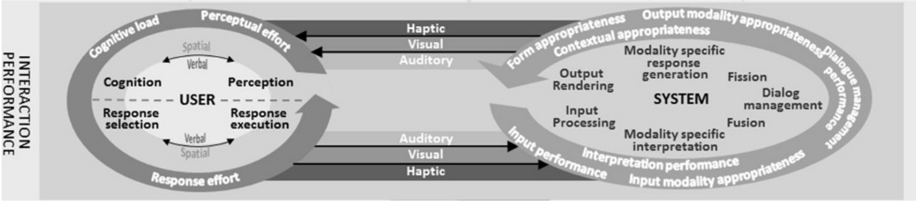


Fig. 3.3. Second layer of taxonomy – the interaction performance.

3.2.1 User

Processing Steps on the User Side

On the user side, the assumptions of the Multiple Resources Theory (MRT) by Wickens (2002) were adopted, but aligned to multimodal systems. In the MRT, *cognitive resources* are differentiated on several dimensions, each having two levels. The dimensions adopted in the taxonomy are *processing codes*, *processing stages*, and, in an adapted form, the *perceptual modalities* and the *answer codes*. For the latter two dimensions three instead of two levels are proposed.

- **Perceptual modalities.** In the MRT, perceptual modalities can be either visual or auditory. Thus, auditory input refers to other perceptual resources than visual input, although both input types may use the same cognitive resources during later stages in processing (see next bullet point). Additionally the haptic modality is suggested here, which is not proposed by Wickens (2002). It was added, as current multimodal systems often include haptic sensors (e.g. vibro-tactile feedback), which can neither described by the auditory nor the visual modality.
- **Processing codes.** The processing codes can be spatial (analog) or verbal (categorical-symbolic). Accordingly, spatial and verbal code can be processed in parallel (Wickens, 2002). Please note that different perceptual modalities might lead to the same internal representation: the internal representation of the written word "usability" is the same as for the spoken version of that word.
- **Processing stages.** Regarding the stages in processing, MRT assumes one pool of resources for perception and cognition (including working memory) and another one for response selection and response execution (Wickens, 2002).
- **Response codes.** The response codes proposed by Wickens (2002) are manual and vocal; they describe the user perspective only and do not cover facial expressions or movements other than manual gestures. To include all possible responses towards a multimodal system as well as the system perspective, the response codes are identical to the perceptual modalities and include haptic responses (e.g. responses including touching and moving the system, like manual responses), auditory responses (responses the system can hear, like verbal responses) and visual responses (responses the system can see, e.g. lip-movements or facial expressions).

Interaction Performance Aspects on the User Side

All processing steps described above, can be mapped to the *interaction performance aspects* proposed by Möller et al. (2009) and Wechsung et al. (2012a). These aspects are *perceptual effort*, *cognitive workload* and *response effort.*

- **Perceptual effort.** Perceptual effort is the effort required for decoding the system messages, and for understanding and extracting their meaning (Zimbardo, 1995), e.g. listening-effort or reading effort. This aspect refers to the perceptual modalities described above. The Borg scale can be used to assess perceptual effort (Borg, 1982).

- **Cognitive workload.** The cognitive workload is defined as the specification of the costs of task performance, such as the necessary information processing capacity and resources (De Waard, 1996). It refers to the processing codes and processing stages. An overview of methods assessing cognitive workload is given in De Waard (1996) and Jahn et al. (2005). A popular method is the NASA-TLX questionnaire (Hart & Staveland, 1988). A lightweight instrument shown to have excellent psychometric properties (Sauro & Dumas, 2009) even in comparison to more elaborate measures (De Waard, 1996) is the Rating Scale Mental Effort (RSME) by Zijlstra (1993). Note that the RSME is also known as the SMEQ (Subjective Mental Effort Questionnaire).
- **Physical response effort.** The physical response effort, is the effort required to communicate with the system (Möller et al. 2009), such as the effort required for typing in an answer or pushing a button. This aspect refers to the response codes. A scale specifically designed to measure physical response effort, is to the authors knowledge, not available. However, the questionnaire proposed in the ITU-T Recommendation P.851 (ITU-T Rec. P.851, 2003) contains items related to physical response effort. Also an adapted version of the RSME (Zijlstra, 1993) may be used.

The degree of interference (and consequently the workload and the effort) increases with the degree to which different tasks or information refer to the same processing dimensions (see Sec. "Processing Steps on the User Side").

To measure performance on the user side, peri-physiological parameters, derived from the user's body, can be used. These measures include pupil diameter, eye-tracking and psycho-physiological measures like electrocardiography (ECG), electromyography (EMG), electroencephalography (EEG) and electro-dermal activity (EDA) (Schleicher, 2009). Generally, these measures are rather unspecific and the valence of a situation (positive or negative) is not determinable even for EMG measures (Mahlke & Minge, 2006). Consequently, drawing inference based solely on these methods is difficult. Other possible data sources are log-files, which may be employed to record task success, task duration, or modality choice. Please note that for all the performance aspects, questionnaires using self-report are mentioned. Self-reports require the user to judge their performance. If such measurements are taken, the experienced workload is measured. The experienced performance and the performance assessed via indirect measurements as described above do not necessarily have to correspond.

3.2.2 System

Processing Steps on the System Side

On the system side, six processing steps have been identified based on the frameworks of Lopez Cozar and Araki (2005) and Herzog and Reithinger (2006).

- **Input Processing.** In the first step the input of the various sensors (e.g. microphones, face recognition, gesture recognition) is processed (Herzog & Reithinger, 2006). The input is decoded into a format understandable to the system, e.g. from acoustics to a text string in case of speech input.
- **Modality Specific Interpretation.** In this step, the transformed input is further transformed into symbolic information and meaning is provided to the data (Herzog & Reithinger, 2006). For example, a sequence of words is analysed to gain the meaning (Lopez Cozar, & Araki, 2005).
- **Fusion.** This is the stage, in which the meaning obtained from the different sensors is merged and combined into one coherent representation, in order to acquire the user's intention (Lopez Cozar, & Araki, 2005; Herzog & Reithinger, 2006).
- **Dialogue Management.** The dialogue management decides on the next steps or actions to be taken, in order to maintain the dialogue coherence to lead the dialogue to the intended goal (Gibbon, Moore, & Winski, 1998; Lopez Cozar, & Araki, 2005; Herzog & Reithinger, 2006).
- **Fission.** The fission operation selects the modalities presenting the output and their coordination (Lopez Cozar, & Araki, 2005; Herzog & Reithinger, 2006).
- **Modality Specific Response Generation.** After fission, modality specific responses are generated; here the abstract output information is transformed into media objects, understandable to the user (Lopez Cozar, & Araki, 2005; Herzog & Reithinger, 2006).
- **Output Rendering.** Finally, the output rendering, the actual presentation of the coordinated system response in the defined media channels like speakers and displays takes place (Lopez Cozar, & Araki, 2005; Herzog & Reithinger, 2006).

Interaction Performance Aspects on the System Side

The above presented processing steps are closely related to the interaction performance aspects proposed by Möller et al. (2009). These are *input performance*, *interpretation performance*, *input modality appropriateness*, *dialogue management performance*, *output modality appropriateness*, *contextual appropriateness*, and *form appropriateness*.

- **Input performance.** The input performance refers to the accuracy or error rate of the recognizers and is linked to input processing. Input performance can be assessed via annotation of the transcribed user input in relation to the correctly determined words, gestures or expressions in terms of substitutions, of insertions, and of deletions (Möller et al. 2009). Another indicator of input performance is the degree of coverage of the user's behavior, e.g. the facial expressions, utterances or gestures. Well-defined parameters for input performance were proposed by Kühnel (2012). These are for instance *multimodal accuracy (MA)* and *multimodal error rate (MER)*. They are defined as the

> "percentage of multimodal user inputs (words, gestures, etc.), which have been correctly recognized, based on the hypothesized and the transcribed or coded reference input, averaged over all recognition moduls. MER = 1 – MA".

- **Input modality appropriateness.** The input modality appropriateness *(IMA)* is dependent on the context, the environment, the user, as well as the information (Möller et al. 2009). It can be determined based on the modality properties by (Bernsen, 2008). Kühnel (2012) identifies three different possible values of input modality appropriateness. These are:

  Appropriate (AP)

  > "IMA:AP All input / output modalities are appropriate for the given context, environment, user and information."

  Partially appropriate (PA)

  > "IMA:PA One or more of the input / output modalities is not appropriate for the given context, environment, user or information."

  Inappropriate (IA)

  > "IMA:IA None of the input / output modalities are appropriate for the given context, environment, user or information."

  Additionally, Kühnel proposes the parameter *unsupported modality usage* (*UMU*). It is defined as:

  > "how often users tried to interact multimodally in a way not supported by the system"

  Assessing these parameters can be done via turn-wise annotation.

- **Interpretation performance.** The interpretation performance is the performance of the system to extract the semantics, the meaning, from the user input. Based on previous research (e.g. Gerbino et al. 1993) Möller (2010) suggests to quantify this in terms of the *concept accuracy*, where a concept is a semantic unit of the input, defined as *attribute value pairs* (*AVP*) (see also, Simpson & Fraser, 1993; Boros et al. 1996; Billi, Castagneri, & Danieli, 1997). To calculate the concept accuracy, a reference transcription of the actual user action is compared to the understanding result for this action. Correctly understood, inserted, substituted, and deleted AVPs are counted, and divided by the total number of AVPs in the utterance, resulting in the *concept error rate* (*CER*). The concept accuracy is then obtained as 1 minus the CER. (cf. Möller, 2010). For multimodal systems, Möller (2010) recommends these measures for each modality and, if fusion is implemented, also for the fused input after fusion. An additional parameter relevant for interpretation performance is *concept efficiency* (Kühnel, 2012). Kühnel (2012) and Möller (2009) define *concept efficiency* based on Glass et al. (2000) as

> "the average number of turns necessary for each concept to be understood by the systems."

- **Dialogue management performance.** The main function of the dialog manager is maintaining the dialogue coherence to lead the dialogue to the intended goal (Gibbon, Moore, & Winski, 1998). This meta-functionality can be assessed via task-success parameters (cf. Möller, 2005). On a lower level the dialog managers has several functionalities, e.g. the selection of the dialogue strategy and error recovery (for a comprehensive list see Gibbon, Moore, & Winski, 1998). Depending on the specific functions, additional parameters may be the *implicit recovery*, the

  > "capacity of the system to recover from user input for which the recognition or understanding process partly failed." (Kühnel, 2012)

  and the *number of system correction turns* or the *system corrections rate*, which are defined as the

  > "overall number (SCT) or percentage (SCR) of all system turns in a dialogue which are primarily concerned with rectifying 'a trouble' (caused by […] recognition or understanding errors, or by illogical, contradictory, or undefined user input), thus not contributing new propositional content and interrupting the dialogue flow. […]" (Kühnel, 2012)

  Further parameters listed in Kühnel (2012) are the *multimodal synergy* and the *fusion gain*. Multimodal synergy was initially defined by Perakakis and Potamianos (2008) and describes the improvement in completion time of the multimodal system in question, compared to a multimodal system randomly combining different modalities. Fusion gain is, according to Kühnel (2012), the sum of recognition errors for each modality compared to the recognition errors of the fused input.

- **Output modality appropriateness.** Output modality appropriateness corresponds to the input modality appropriateness and can be quantified in the same way, by using the properties offered by Bernsen (2008). The possible values are the same as for input modality appropriateness.

- **Contextual appropriateness.** Contextual appropriateness is the proportion of system utterances being judged as appropriate in the dialogue context. Judgments are based on Grice's Cooperativity Principles and quantified in terms of violations of this principle (Möller, 2005). Möller (2005) defines five possible values these are *total failure, appropriate, inappropriate, appropriate/inappropriate* (experts reach no agreement), and *incomprehensible.*

- **Form appropriateness.** Form appropriateness reflects the adequateness of the surface form of the output. Regarding spoken output, aspects of form appropriateness are *intelligibility* and *comprehensibility*. For graphical output its *readability* and *legibility* might be relevant. Metrics advised by Kühnel (2012) are for instance the *system turn duration*, and the *number of elements per system turn,* as well as asynchrony measures like the *lag of time* between corresponding modalities, and the *number of asynchronous events*.

Please note that, these interaction parameters obtained by logging user and system behaviour are referred to as *indirect* measurements in contrast to *direct* measurements like user ratings (Möller et al., 2009, Seebode et al., 2009). For a more in-depth discussion and an exhaustive list of indirect parameters of multimodal human-computer interaction, refer to Kühnel (2012). Direct measurements require the direct perception and judgement of the human subject, i.e. the participant or user acts as the measuring organ (Möller et al., 2009).

Indirect measures, in contrast, do not involve any judgmental process from the user and are not necessarily reflecting the user's experience or the perceived quality. Accordingly, such performance metrics are not necessarily correlated with a user's perceptions although this is of course possible. This assumption is supported by the findings reported in literature. For instance, a meta-analysis conducted by Nielsen and Levy (1994) showed that performance and predicted preference are indeed correlated. Sauro and Kindlund (2005) reported similar results: They found negative correlations between satisfaction (direct measurement) and time, errors (indirect measurement) and a positive correlation between satisfaction (direct measurement) and completion (indirect measurement). However, several studies reported opposing findings: Möller (2006) could not find a correlation between task duration and user judgments when evaluating speech dialogue systems. Also Frøkjær et al. (Frøkjær, Hertzum, & Hornbæk, 2000) did not observe a direct relationship between user ratings and indirect efficiency measures, such as task duration. Results from a meta-analysis by Hornbæk and Lai Chong-Law (2007) showed that the user's experience of the interaction and performance measures differ considerably from each other or show even negative correlations. Thus, it is indispensable to measure how the interaction is experienced by the user.

3.3 Quality Aspects

As mentioned before, to assess a system's quality the metrics presented above are not sufficient. The following section describes relevant aspects users might experience during the interaction and proposes assessment methods for each of them. As experience is inherently subjective, this can only be done by obtaining the data directly from the user. Thus, all aspects explained in the following section are measured directly and thus involve a user to judge the interaction (Fig. 3.4). Accordingly, they represent the user's experience of the interaction if they are assessed during interaction, or the remembered experience if assessed retrospectively.

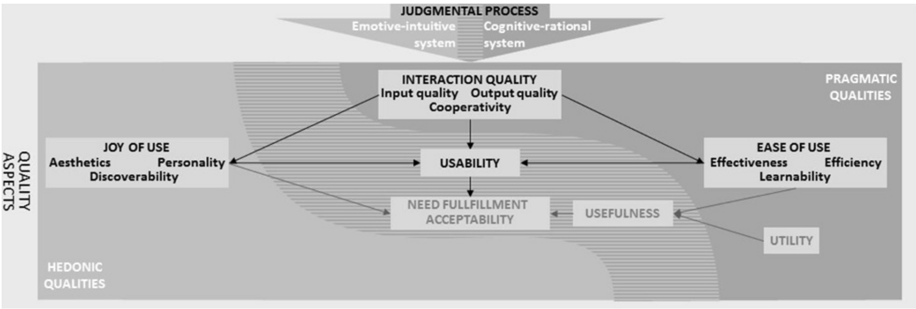


Fig. 3.4. Third layer of the taxonomy – the quality aspects.

3.3.1 Judgemental Process

Research indicates that judgment and decision-making processes involves two systems, the *cognitive-rational* and the *emotive-intuitive system* (e.g Epstein, 1994; Kahneman, 2003; Lee, Amir & Ariely, 2009, Loewenstein & O'Donoghue, 2004).

The cognitive-rational systems is, compared to the emotive system, more analytic, logical, abstract, active, controlled, rule-based and slower (Kahneman, 2003; Lee, Amir & Ariely, 2009); it is the deliberate mode of judgments (Kahneman, 2003). The emotive-intuitive system, on the other hand, is characterized by automatic, associative effortless and often emotionally charged operations (Kahneman, 2003); it is the automatic mode of judgments. These automatic, intuitive judgments of the emotive system are monitored by the cognitive system and may be corrected or overridden (Kahneman, 2003); however, the monitoring is rather loose, as the analytical conscious processing in the cognitive system requires mental resources and thus induces cognitive load (Kahneman & Frederick 2002). Hence, the judgments of the emotive-intuitive system determine preferences unless the cognitive system intervenes (Kahneman, 2003) and in every action or thought the emotional system is, at least unconsciously, engaged (Picard, 1997)

However, for a long time the only emotion considered in human-computer-interaction (HCI) was frustration, and how to prevent it. Only during the last decade, emotions and affect became a major research topic in HCI. In line with the findings reported above, it is argued that every interaction with technological systems involves a wide range of emotions (Brave & Nass, 2007). Hassenzahl et al. (2010) also emphasize the close relationship between experience, emotions, and affect proposed earlier by McCarthy and Wright (2004). Their position is, according to Hassenzahl et al. (2010), strongly influenced by the work of Dewey, who describes emotions as the "qualities of experiences". Thus, a positive experience is linked to a positive emotion and vice versa. Evidence for the above assumptions is provided by Schwarz and Clore (2003); they showed that apparent affective responses towards a target are used as information and therefore influence evaluative judgements, especially when the judgements refer to preference- or "likeability"-judgements (cf. Section 3.1.1 User Characteristics -Mood).

Accordingly, evaluative judgements of a system are not solely based on the system's attributes, but on the feelings, the user has towards the system and the mode of judgement used when forming the judgements.

3.3.2 Hedonic and Pragmatic Qualities.

Traditionally HCI was focusing on enhancing the efficiency and effectiveness of the system (Preece et al., 1994); the major concern was to prevent negative emotions (Hassenzahl & Tractinsky, 2006). With the paradigm shift from instrumental to non-instrumental qualities (cf. sub-section Usability) concepts of positive or hedonic psychology were adapted and transferred to HCI. Hedonic Psychology, as proposed by Kahneman (1999), is focusing on concepts like enjoyment, pleasantness, but also unpleasantness rather than on attention and memory, two key topics that have been the focus of psychological research. Analogous to this development, HCI research also moved away from the "classical" cognitive information processing paradigm (Kaptelinin et al., 2003) towards concepts like Affective Computing (Picard, 1997) and Emotional Design (Norman, 2004). Nowadays the aim is not only to prevent rage attacks as a result of a crashing computer, but to facilitate positive emotions while interacting with an interactive system (Hassenzahl & Tractinsky, 2006).

Thus, for the taxonomy, the differentiation between *pragmatic qualities* and *hedonic qualities* proposed by Hassenzahl et al. (2000) was used. While pragmatic qualities refer to functional aspects of a system and are closely related to the classical concept of usability, hedonic qualities cover the interfaces non-instrumental aspects (Hassenzahl, 2005). Hassenzahl et al. (Hassenzahl, Diefenbach, & Göritz, 2010) adopted the terminology of Herzberg's Two Factor Theory of Job Satisfaction (Herzberg, 1968) to describe the different characters of pragmatic and hedonic qualities. Here, *hygiene factors* and *motivators* are distinguished: Hygiene factors (job context factors such as the environmental conditions) can in the best case just prevent dissatisfaction with the job but cannot lead to satisfaction (Herzberg, 1968). However, their absence will results in dissatisfaction. The absence of motivators (job content factors, such as acknowledgment) on the other hand may not result in dissatisfaction, but their presence will facilitate satisfaction and motivation. Hassenzahl et al. (Hassenzahl, Diefenbach, & Göritz, 2010) found pragmatic qualities to be a hygiene factor removing barriers hindering the fulfilment of the users' needs. Hence, a system's pragmatic qualities enable need fulfilment, but are themselves not a source of a positive experience. Hedonic qualities are associated with a system's ability to evoke pleasure and the psychological well-being of the user (Hassenzahl, 2003a); they are motivators and reflect the products capability to create a positive experience (Hassenzahl, Diefenbach, & Göritz, 2010).

3.3.3 Interaction Quality

The construct probably closest related to the performance is the *interaction quality*. It comprises the perceived *input* and *output quality* and the perceived *cooperativity*. Möller et al. (Möller et al., 2009) define input quality "as the perceived system understanding and input comfort" and output quality as the perceived "understandability and form appropriateness". For multimodal systems, especially the perceived appropriateness of the input and output modalities needs to be taken into account. Multimedia Learning Theory proposes that words should be presented auditorily rather than visually when employing multimedia. This way an overload of the visual information processing channel can be prevented as words are processed in a separate system (Mayer & Moreno, 1999; Wickens 2002; cf. Sections 2.2 and 3.2.1).

Cooperativity refers to "the distribution of initiative between the [interaction] partners" (Möller et al., 2009) but also includes the consideration of the user's knowledge, and the system's ability for repair and clarification.

Interaction quality can be described by the interaction's speed and smoothness (Möller et al., 2009). Unfortunately, measurements methods specifically addressing interaction quality are rather rare. Möller et al. (2009) suggest the questionnaire proposed in the ITU-T Rec P.851 (ITU-T Rec P.851, 2003), which is intended for the assessment of spoken dialogue systems.

3.3.4 Ease-of-Use

Ease-of-use is closely related to pragmatic qualities. Key aspects are the traditional usability measures described by the International Organization for Standardization (ISO) in the ISO 9241-11 standard (ISO 9241-11, 1998). Namely, these are *effectiveness*, which is the accuracy and completeness of goal attainment with respect to user and context of use, and *efficiency*, i.e. the required effort and resources related to the accuracy and completeness. Although efficiency is often measured via perceived task duration, in the taxonomy mental efficiency is included. Thus, the perceived mental effort is considered as a resource. As already mentioned in Section 2.2, reducing mental effort by employing multiple modalities is one of the key advantages of multimodal systems (Sarter, 2007).

In addition, to the aspects presented above, also *learnability* determines a system's ease-of-use (Dix et al., 1993). Learnability describes how well a new user can effectively interact with a system and maximize performance. On a lower level, learnability includes *predictability, synthesizability, familiarity, generalizability* and *consistency* (Dix et al., 1993). Thus, it is largely overlapping with the concept *intuitiveness*, which is an additional aspect in the original taxonomy of Möller et al. (Möller et al. 2009, Wechsung et al. 2012). Möller et al. adopt the definition proposed by Naumann and colleagues (2007). Here, intuitiveness is described as the

> "extent to which the user is able to interact with a technical system effectively by applying knowledge unconsciously".

According to Laakkonen (2007), intuitiveness can be seen as an attribute of learnability, which is in line with the learnability definition cited above by Dix et al. (1993). Hence, learnability includes intuitiveness. Thus, intuitiveness is not considered as a separate aspect.

The vast majority of standardized usability questionnaires cover these constructs. Examples are the QUIS (Shneiderman, 1998), the SUS (Brooke, 1996), the IsoMetrics (Gediga, Hamborg, & Düntsch, 1999), the AttrakDiff (Hassenzahl, Burmester, & Koller, 2003) and the SASSI (Hone & Graham, 2000). It has to be noted that the questionnaires' sub-scales are not necessarily named efficiency, effectiveness, and learnability. The SASSI sub-scale Speed is strongly related to efficiency, the scale Pragmatic Qualities on the AttrakDiff refers to both, efficiency and effectiveness.

Besides questionnaires, expert-oriented procedures such as the Cognitive Walkthrough (Wharton et al., 1994) and modelling approaches like GOMS (Card, Newell & Moran, 1983) are regularly used for the evaluation of ease-of-use. (cf. Section 2.3.2). Please note, that the expert-oriented procedures do not involve users. Therefore, they do not assess the quality as perceived by the user. Though, they may provide an estimation of the user's perceptions.

Although there is a wide range of methods available for assessing a system's ease of use, few of them are so far particularly tailored to multimodal systems. Using established questionnaires like the SUMI (Kirakowski & Corbett, 1996) and the QUIS (Sheiderman, 1998) might be problematic as they were developed for unimodal graphical user interfaces like websites. However, the scale measuring Pragmatic Qualities of the AttrakDiff may provide meaningful results (cf. Section 4.1). A possible explanation is that the AttrakDiff measures on a relatively high level appropriate for a variety of different interfaces. The SUS questionnaire also offers rather generic questions, which might be adaptable to multimodal systems (Brooke, 1996).

3.3.5 Joy-of-Use

Joy-of-use is the positive feeling a user has when using technical or interactive systems (Schleicher & Trösterer, 2009) and is associated with *hedonic qualities*. Hassenzahl et al. (Hassenzahl, Diefenbach, & Göritz, 2010) found evidence for the assumptions that positive experiences during interactions are related to the fulfilment of human needs. Moreover, they suggest a link between *need fulfilment* and a system's hedonic, non-functional qualities under the precondition that the experience is attributed to the system and not to the context (e.g. the person can attribute a positive experience with a phone to the device itself or to the conversation or on both). In the taxonomy of Möller et al. (2009) aspects of joy-of-use are *aesthetics* and *system per-*

sonality. Aesthetics covers the “pleasure attained from sensory perception” (Hekkert, 2006). The system’s personality includes system factors like voice (e.g. gender of the voice), the wording of the voice prompts, as well as colour and icon schemes.

As a further concept Möller and colleagues suggest *appeal,* which is seen the result of aesthetics and system personality. However, the definition they provide for appeal refers to Hassenzahl's (2003) concept of *stimulation*, the extent to which a system possesses interesting, novel, and surprising features. According to this conception, a product perceived as highly aesthetic should show a large extent of novelty. Still research shows that high novelty might actually decrease the perceived aesthetics and that with increasing familiarity also perceived aesthetics increase (Sluckin, Hargreaves, & Colman, 1983). This phenomenon, known as the mere-exposure effect (Zajonc, 1968), has been shown for various stimuli in different modalities (Szpunar, Schellenberg, & Pliner, 2004). However, the relationship between novelty and aesthetics is not linear, as also an overexposure effect is reported (Williams, 1987). Hence, the findings reported above imply that the relationship between novelty and aesthetics is an inverted U-function (Sluckin, Hargreaves, & Colman, 1983). Accordingly, appeal is not adopted for the taxonomy described here; instead the concept *discoverability* is suggested. As appeal it is linked to the concept of *stimulation* suggested by Hassenzahl (2003); it describes a system’s ability to enable personal development respectively the proliferation of knowledge and the development of skills, e.g. by providing innovative and/or exciting features - by being discoverable. Please note that discoverability is not seen as the result of aesthetics and personality but is allocated on the same level.

It is noteworthy that there is an on-going debate concerning the relationship between hedonic qualities, the aspects related to joy-of-use, and pragmatic qualities, the aspects related to ease-of-use. While some findings provide evidence for the claim that “what is beautiful is usable” (Tractinsky, Katz, & Ikar, 2000) and for the underlying assumption that joy-of-use and ease-of-use are interdependent; other studies could not confirm these results (Mahlke & Lindgaard, 2007). Hassenzahl (2008) suggests that these ambiguous results are caused by different approaches in understanding and measuring aesthetics.

Accordingly, a variety of methods is available to measure joy-of use and related aspects but before deciding on a measurements method it has to be defined which aspect should be assessed. The questionnaire proposed by Lavie and Tractinsky (2004) is suitable for measuring the visual aesthetics, but not for aesthetics perceived via other modalities. The AttrakDiff (Hassenzahl, Burmester, & Koller, 2003) measures hedonic qualities on a higher level and is not limited to unimodal interfaces. For measuring hedonic qualities during the interaction the Joy-Of-Use-Button (Schleicher & Trösterer, 2009) and psycho-physiological parameters are available options, the latter being the most resource-intensive method (Schleicher, 2009).

Another well validated and widely used instrument is the Self-Assessment Manikin (Bradley & Lang, 1994), which measures the *arousal*, *pleasure* and *dominance* linked to affective reactions on three non-verbal scales.

If the aim is to measure specific emotions LemTool (Huisman & Van Hout, 2008) or PrEmo (Desmet, 2004) may be used. However, both tools are so far only validated for specific application areas: Lemtool for websites and PrEmo for non-interactive products (Huisman & Van Hout, 2008; Desmet, 2004).

Although a wide range of methods assessing hedonic, affective qualities are nowadays available, a recent review by Bargas-Avila and Hornbæk (2011) indicates that questionnaires, more specifically the hedonic qualities sub-scales of Hassenzahl's AttrakDiff (Hassenzahl, Burmester, & Koller, 2003) and the Self-Assessment Manikin by (Bradley & Lang, 1994) are by far the most popular instrument.

Please note that for evaluations of affective qualities care has to be taken, when deciding if the measurements will take place during or after the interaction. Apart from the general memory biases (e.g. consistency bias, change bias, stereotypical bias, for an overview see Schacter, 2001), several memory biases are documented regarding the retrospective assessment of emotional experiences. Kahneman and colleagues (1993) showed that retrospective reports of affective experiences are mainly based on the moment of the peak of the affect intensity and on the moment of the ending of the experience. This so-called *peak-end rule* has been shown in the context of interface evaluation (Hassenzahl & Sandweg, 2004; Hassenzahl & Ullrich, 2007). Accordingly, retrospective questionnaires reflect the remembered affective experience, but might give only little information on the specific aspects, which lead to the global evaluation (Wechsung et al. 2012b).

3.3.6 Usability

The ISO 9241-11 standard (ISO 9241-11, 1998) defines usability as the

> "extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use"

This definition sets the focus on ergonomics and hence only on the task-oriented, functional aspects of human machine interaction, which were the major themes in the early years of human-computer interaction.

This is plausible, as the initial work was mainly concerned with military and research machines. However, Shackel's (Shackel, 2009) extensive review on the history shows that while the technologies changed rapidly the main issue of research in human-computer interaction stayed the same. Although information and communication technologies became ubiquitous, with the mini and later microcomputers finding their way first into offices and soon after into almost every place, and in spite of the success of the hypertext and the World Wide Web, research was still focusing on work settings and the achievement of instrumental goals (Hassenzahl & Tractinsky, 2006).

According to Hassenzahl (2008), this traditional perspective implies that technology usage is mainly motivated to gain time for doing other (non-technology related) things – thus technology usage itself was not being considered as a pleasurable experience.

Even though (Hassenzahl & Tractinsky, 2006) show that aspects like "fun" and "experience" were already presented during the late 1980s, they also point out that it took a number of years until these ideas were adopted by the HCI community, with the new concept User eXperience (UX) becoming increasingly popular only during the last decade. The origins of the term UX probably lie in the work of Donald Normans at Apple Computers (Norman, Miller, & Henderson, 1995).

Although the term UX became omnipresent, the concept itself was neither being well defined nor well understood (Law et al., 2008). The lack of a shared view on UX (and indeed the subsequent need for one) became obvious, when many companies just exchanged the label usability with the label user experience, but kept on doing the same task-centred usability testing and engineering they did before (Hassenzahl, 2008). In academia on the other hand, lots of research was conducted aiming to define UX and its preconditions. To date, relevant literature offers numerous models, theories and definition of user experience, joy-of-use, hedonic qualities or emotional design (e.g. Hassenzahl et al., 2000; Desmet & Hekkert, 2007; McCarthy & Wright, 2004; Jordan, 2000; Forlizzi & Battarbee. 2004; Norman, 2004). A survey among researcher and practitioners, conducted by Law et al. (2009), showed how heterogeneous the views on UX are; however, the surveys' authors were able to deduce the following shared understanding: UX can be described,

> "as a dynamic, context-dependent and subjective [concept], which stems from a broad range of potential benefits the users may derive form a product."

A similarly broad definition is given in the ISO 9241-210 standard (ISO 9241-210, 2010). Here, UX is defined as

> „a person's perceptions and responses that result from the use or anticipated use of a product, system or service".

According to Bevan (2009), this definition permits three different interpretations of the term UX: First of all UX can be understood as,

> "an umbrella term for all the user's perceptions and responses [..]".

Secondly, UX can be understood as a different concept, maybe even as counter concept, to usability, as historically the focus of usability is mainly on performance. The third interpretation sees UX as

> "an elaboration of the satisfaction component of usability".

For the taxonomy the last view was adopted and thus two aspects of usability considered: The ease-of-use, which is influenced by the mentioned consequences of interaction quality, and the joy-of-use, which is often associated with UX. Please notice, that this definition of usability is relatively broad.

Following the definition above, appropriate methods need to measure both, joy-of use and ease-of-use. Although several questionnaires measure ease-of-use only few include joy-of-use. An affect scale is included in the Software Usability Measurement Inventory (SUMI; Kirakowski & Corbett, 1993). The Post-Study System Usability Questionnaire (PSSUQ; Lewis, 1995) and the Computer System Usability Questionnaire (CSUQ; Lewis, 1995) additionally measure frustration. The AttrakDiff's Attractiveness scale, measuring pragmatic (ease-of-use) as well as hedonic qualities (joy-of-use), is probably closest to the presented concept of usability (Hassenzahl, Burmester & Koller, 2003).

Apart from the questionnaires presented above, other suitable methods to assess joy-of use and ease-of-use include qualitative approaches like the Repertory Grid Technique (Kelly, 1955), the Valence Method proposed by Burmester (Burmester et al. 2010) and the UX Curve (Kujala et al., 2011). The Repertory Grid Technique has its origin in the psychology of personal constructs by Kelly (1955). Constructs in Kelly's sense are bipolar (dis-)similarity dimensions (Hassenzahl & Wessler, 2000). According to Kelly (1955), every human owns an individual and characteristic system of constructs, through which he/she perceives and evaluates his/her experiences. The Repertory Grid Technique aims to assess these individual constructs in two phases. In the elicitation phase, the persons is presented with triads of the relevant object (e.g. three websites as in Hassenzahl & Trautmann, 2001) and is asked to verbalize what two objects have in common and how they differ from the third. This way bipolar constructs in the form of a semantic differential are generated. These bipolar constructs are later used as the rating scale for all constructs. The result is an individual construct-based description of the objects (Hassenzahl & Wessler, 2000).

The Valence Method (Burmester et al. 2010) is a two-phase measure based on the theoretical work by Hassenzahl et al. (Hassenzahl, Diefenbach, & Göritz, 2010). First, the users are asked to set positive and negative valence markers while exploring the interface for up to eight minutes. The whole session is videotaped. In the next phase, the marked situations are presented to the participants again while the interviewer is asking which design aspect was the reason for setting the marker. The laddering interviewing technique is employed, to uncover the underlying need by repeating the question, why a certain attribute was mentioned until the affected need is identified.

The main limitation according to the authors (Burmester et al. 2010) is, that it is currently recommendable for first usage situations only, as the number of markers increases substantially if the product is already known to the user. Valuable insights in form of quantitative data (number of positive and negative markers) and qualitative data (interviews) can be gained with this method, one disadvantage probably being the relatively high resources required.

The UX Curve (Kujala et al., 2011) is used to retrospectively assess the system's quality over time. Participants are asked to draw curves describing their perceptions of the system's attractiveness, the system's ease-of-use, the system's utility, as well as the usage frequency and their general experience over time. In addition, users should explain major changes in the curves. According to the authors, the UX curve allows to measure the long-term experience and the influences that improve or decrease the perceived quality of the experience. A similar method is offered with iScale (Karapanos, Martens & Hassenzahl, 2012). Again, users are asked to draw a curve reflecting their experience. However, iScale is an online survey instrument while the UX curve was developed for face-to-face settings. Moreover, the assessed dimensions differ.

3.3.7 Utility and Usefulness

Utility has been defined by Grudin (1992) as the functionality or capability of the system, related to the tasks the user wants to accomplish with the system. Thus, usability and utility are separate concepts, with usability referring to the joy-of use and the ease-of-use of the system, and utility to it functionality. This also means that an interface may have zero usability but still a high utility. For example, a software program may offer all the functions a user needs, but the interface is so bad, that the user is unable to access this function. The other way around, a highly usable interface may have no utility. Usefulness comprises both usability and utility. Accordingly, a system is useful, if it provides the functionality to support the tasks, the user wants to accomplish, and if the user can employ these function in an easy (e.g. efficient, effective) and pleasurable manner.

It is noteworthy, that the distinction between usability, usefulness, and utility is often fuzzy (Landauer, 1995) and sometimes those terms are used synonymously. For example, Lindgaard (1994) defines usefulness, as the

> "degree to which a given system matches the tasks it is intended to support. "

This definition basically equals the definition of utility presented above. Thus, it is often not clear what was measured, when results regarding one of the three concepts are reported, thus such results are difficult to interpret.

Moreover, methods measuring utility and usefulness, as understood in the first paragraph of this section, are rather rare. Particularly utility is difficult to assess in classical lab tests as typically only tasks are selected that are supported by the product (Cordes, 2001). Domain relevant tasks the system is not capable of are not presented. Cordes (2001) showed that if users are told that the given tasks may not be solvable, the users tended to terminate the task earlier and terminate more tasks than users receiving the exactly same instruction, except for the hint that the tasks may not be solvable. Thus, it is likely that the user assumes that the utility to fulfil the tasks is provided by the interface, if not stated otherwise in the instructions.

Usefulness is partly covered in the PSSUQ as well as in the CSUQ. Those questionnaires are identical, except that PSSUQ was developed to use after a usability test, and is thus addressing specific tasks, whereas the CSUQ asks about the general system and is suitable for surveys (Lewis, 1995).

Within this thesis, usefulness and utility are considered only marginally, as they can best be assessed in field studies with real test users in natural settings. In contrast most of the studies conducted in the context of this thesis were experimental lab studies. Consequently, usefulness and utility could not be measured meaningfully.

3.3.8 Acceptability

Acceptability has been described by Möller et al. (Möller et al., 2009) as how readily a user will actually use the system. Quality aspects alone are not sufficient to explain whether a service will be accepted or not (Möller, 2005; Larsen, 2003b). In fact, acceptability is determined by a complex interaction of multiple aspects (Larsen, 2003b), with the influencing factors, presented in the first layer of the taxonomy assumed to be highly relevant (Möller, 2005). Möller (2005) points out that especially service factors like the costs of the system and, depending on the system, privacy and security issues are of major concern.

One of the most influential approaches in determining acceptance is the Technology Acceptance Model (TAM; Davis, 1993). Its theoretical base is the Theory of Reasoned Actions (TRA) by Ajzen and Fishbein (1980). According to the TRA, actual behaviour is determined by behavioural intentions. Behavioural intentions are dependent on attitudes towards the behaviours and subjective norms. In terms of the TAM, the attitudes towards the behaviour are the perceived usefulness and the perceived Ease of Use. Subjective norms are not included in the TAM.

It is important to note that perceived usefulness, as defined in the TAM, does not exactly match the understanding of the term presented above. More precisely, perceived usefulness is defined as

"the degree to which a person believes that using a particular system would enhance his or her job performance".

Thus the sub-scale usefulness in the TAM questionnaire is not an adequate measure of the conceptualization of usefulness given previously. However to measure acceptance the questionnaire is helpful although it will not provide detailed information about the system's quality.

As for usefulness and utility also acceptability is out of the scope of this thesis, as it can only be measured ultimately using a fully working system counting the real users (Möller, 2005).

3.4 Chapter Summary

The section above describes a taxonomy of quality aspects of multimodal systems. Although the framework provides a holistic view integrating factors of experience and performance, only the performance factors have just recently been tailored and validated for multimodal systems by Kühnel (2012).

Regarding the experience-related factors most of the methods were designed and validated for the evaluation of unimodal, GUI-based systems. Thus, it is not clear if these established methods are applicable to evaluation of multimodal systems.

The most common technique applied in quality and usability evaluations are probably questionnaires, but a standardized and validated questionnaire addressing the evaluation of multimodal systems is still not available. In the following chapter, the development of such a questionnaire is described.

4 How to Evaluate? - Development of the MMQQ (MultiModal Quality Questionnaire)

In the this chapter the development of the MultiModal Quality Questionnaire (MMQQ) is described. One aim was to build a psychometrically validated, reliable instrument for assessing the quality of multimodal interaction; second aim was to validate and, if necessary adapt, the taxonomy of Möller et al. (Möller et al. 2009, Wechsung et al. 2012a). Accordingly, the taxonomy presented in Möller et al. (2009) served as the theoretical base of the questionnaire.

At first, an initial study comparing established questionnaires regarding their suitability for multimodal interfaces is presented. Based on the results of the study, the AttrakDiff questionnaire (Hassenzahl, Burmester, & Koller, 2003) was selected as a starting point for developing the new instrument. Next, a preliminary item pool was generated. This item pool was used in a pilot study and evaluated by experts. Based on the psychometric properties, on the feedback of the participants, and on the results of an expert evaluation, some items were dropped. The remaining items were then used for further evaluation studies with different systems. Finally the remaining item set was validated using a structural equitation model. Each of the aforementioned steps is explained in the following sections.

4.1 Study 4.1 - Comparison of Established Questionnaires

A first study was conducted comparing well-known and widely used questionnaires for the evaluation of graphical user interfaces and speech-based systems. The general objective was to investigate if those questionnaires are appropriate for the evaluation of multimodal systems. It was examined to which extent different questionnaires show agreement, indicating the measurement of similar constructs. After that, the direct measurements, assessed through the aforementioned questionnaires, and the performance measures assessed via log files during the interaction, were compared to see how these data types relate to each other.

The questionnaires used in the study were

- the Software Usability Measurement Inventory (SUMI; Kirakowski & Corbett, 1993) questionnaire,
- the System Usability Scale (SUS; Brooke, 1996),
- the tool for Subjective Assessment of Speech System Interfaces (SASSI; Hone & Graham, 2000) and
- the AttrakDiff (Hassenzahl, Burmester & Koller, 2003).

The SUMI was chosen as it is one of the most mature usability questionnaires, it is also recommended in the ISO standard 9241-11 (ISO 9241-11, 1998). The SUMI

Ina Wechsung, *An Evaluation Framework for Multimodal Interaction*,
T-Labs Series in Telecommunication Services,
DOI: 10.1007/978-3-319-03810-0_4, © Springer International Publishing Switzerland 2014

consists of five scales, which are named Efficiency, Affect, Learnability, Helpfulness and Control. Each scale comprises ten items. Additionally a global scale can be calculated using 25 of the 50 items. Those items showed the highest loading on a general Usability factor (Kirakowski & Corbett, 1993).

The SUS is a single scale questionnaire comprising ten items. Reasons for including the SUS were that the SUS is, according to a search within the ACM Digital Library, the most widespread usability questionnaire after the SUMI (33 matches), yielding 22 matches, and that it is free of charge. Moreover, the SUS was shown to be more sensitive than the Questionnaire for User Interaction Satisfaction (QUIS; Shneiderman, 1998), although it is far shorter (Tullis & Stetson, 2004). Hence, the QUIS was not included. Moreover, the QUIS yielded less matches in the ACM Digital Library than the SUS and it has to be paid for.

Other questionnaires considered were the IBM Usability Questionnaires (Lewis, 1995), namely the Post-Study System Usability Questionnaire (PSSUQ), the After-Scenario Questionnaire (ASQ) and the Computer System Usability Questionnaire (CSUQ). Like the QUIS they yielded less matches in the ACM Digital Library than the SUS, which indicates that they are less popular. Consequently, they were excluded. For the same reason the German questionnaires ISONORM (Prümper, 1997) and IsoMetrics (Gediga, Hamborg, & Düntsch, 1999) were not investigated, since the aim was to compare widespread questionnaires.

The two other questionnaires, the AttrakDiff (Hassenzahl, Burmester & Koller, 2003) and the SASSI (Hone & Graham, 2000), were chosen due to contextual considerations. To the author's knowledge, the AttrakDiff was at the time of the study the only questionnaire explicitly measuring hedonic qualities of the interface. The theoretical basis of the AttrakDiff questionnaire is Hassenzahl's model of user experience (Hassenzahl, 2003a). This model postulates that a product's attributes can be divided in hedonic and pragmatic attributes. Hedonic refers to the products ability to evoke pleasure and emphasize the psychological well-being of the user. Hedonic attributes can be differentiated into attributes providing stimulation, attributes communicating identity and attributes provoking memory. The AttrakDiff aims to measure two of these hedonic attributes: stimulation and identity. A product can provide stimulation when it presents new interaction styles, for example new modalities. Thus, multimodality should be measureable with the scale Hedonic Qualities-Stimulation. The scale Hedonic Qualities-Identity measures a product's ability to express the owners self. Pragmatic Qualities refer to a product's functionality and the access to the functionality. The fourth scale is named Attractiveness and is the AttrakDiff's global scale. It measures both hedonic and pragmatic qualities.

The SASSI is considered to be the most widely used (Larsen, 2003a) questionnaire for speech-based interfaces (Hone & Graham, 2000). It consists of 34 items with Likert-scales as the answer format. The items form six factors, which are System Response Accuracy, Likeability, Cognitive Demand, Annoyance, Habitability and Speed.

All of the multimodal interfaces offered speech control; accordingly, it was assumed that a questionnaire for speech-based interfaces might be appropriate for their evaluation. The SUXES (Turunen et al., 2009), a questionnaire tailored to multimodal interaction, was not yet published and thus unavailable at the time the study was conducted.

4.1.1 Method

Participants

Twenty-one German-speaking individuals (11 male, 10 female) aged between 19 and 69 years (*M* = 31.24) took part in the study. All users participated in return for an Amazon voucher.

Due to technical problems, log-data was not recorded for three participants. Three further cases were identified as outliers[1] based on the observed task duration, and were therefore excluded from all analyses involving task duration.

Devices

The multimodal devices adopted for the test were a PDA (Fujitsu-Siemens Pocket LOOX T830) and a Tablet PC (Samsung Q1-Pro 900 Casomii). Both devices could be operated via speech control (IBM embedded Via Voice) as well as via a graphical user interface with touch screen. Additionally, the PDA could be operated via motion control. Furthermore, a unimodal device, a conventional Desktop PC with mouse and keyboard input, was used as a control condition. The application, a web-based media recommender system called MediaScout, was the same for all devices. The MediaScout allows the users to search through different media content such as internet videos, movie trailers and the TV programme. Search queries can be filtered by content type or time (e.g. “TV programme tomorrow between 4 and 6 p.m.). Results will be shown in a list, ordered by the assumed preferences of the user, which are based on a pre-defined user profile. The suggested results can be rejected, accepted, or marked as “favourites” in order to improve the user profile and eventually the recommendations. Moreover, it is possible to invite friends to the application. All functionalities of the MediaScout were available on all devices and for all modalities. For more information on the MediaScout see Stegmann et al. (Stegmann, Henke, & Kirchherr, 2008).

Procedure

[1] Outliers are defined as $X_{individual}$ >3rd Quartile of X_{total} + 1.5*Inter Quartile Range of X_{total}

The users performed five different types of tasks: seven navigation tasks, six tasks where checkboxes had to be marked or unmarked, four tasks where an option from a drop-down list had to be selected, three tasks where a button had to be pressed, and one task where a phone number had to be entered.

The SUMI, the AttrakDiff and the SASSI were used in their original wording. But for the SASSI the original answer scale, a 7-point scale (*strongly agree, agree, slightly agree, neutral, slightly disagree, disagree, strongly disagree*), was exchanged with a 5-point scale (*strongly agree, agree, neutral, disagree and strongly disagree)* in order to be consistent with the ITU-T Recommendation P.851 (ITU-T Rec. P.851, 2003).

As the SASSI was designed to evaluate speech-based applications, SASSI ratings were collected only for both multimodal systems (PDA and Tablet PC). The SUS was adapted for voice control by exchanging the word “system” with “voice control” and was hence used only for the Tablet PC and the PDA.

Each test session took approximately three hours. Each participant performed the series of tasks with each device. Participants were first instructed to perform the tasks with a given modality. This was repeated for every modality supported by that specific device. After that, the tasks were presented again and the participants could freely choose the interaction modality. Finally, they were asked to fill in the SUMI, the AttrakDiff, the SUS and the SASSI questionnaire to rate the previously tested device.

This procedure was repeated for each of the three devices. After the third device, a final questionnaire regarding the overall impressions and preferences had to be filled in. The sequence of the questionnaires as well as the sequence of the devices were randomized for each subject in order to avoid fatigue or learning effects. An illustration of the procedure is presented in Figure 4.1.

Dialogue duration as a measure of efficiency was assessed task-wise via the log-files and was, for each system, averaged over all tasks.

The scales and sub-scales for each questionnaire were calculated according to the instructions in the specific handbook or manual. All questionnaire items which were negatively poled were recoded; accordingly higher values indicate better ratings.

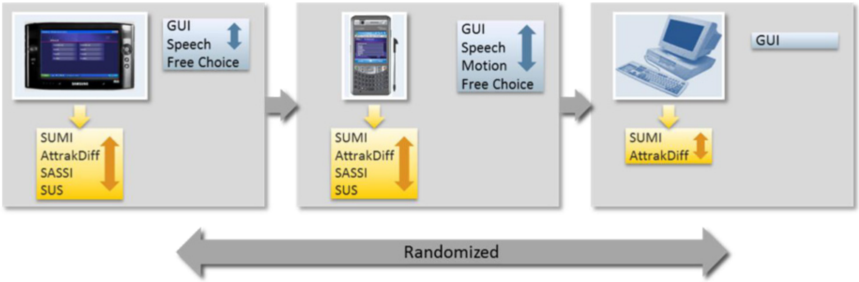


Fig.4.1. Sample procedure and devices.

4.1.2 Results

In this section, the results of the questionnaire comparisons are reported. The detailed comparisons of the different devices are not in the focus of the current study; for completeness, those results are reported in Appendix A.1.

Overall Scales

For a first overview, the raw scores of the overall scales of the questionnaires were transformed into ranks. The SUMI has a scale explicitly measuring "global usability". For the AttrakDiff, results of the Attractiveness scale were transformed into ranks, since this scale is reflecting the overall attractiveness of the system. The SASSI has no such overall scale; hence, the mean of all items was used as a global assessment. The SUS is a single-scale instrument, and thus has no overall scale, like for the SASSI, a global scale based on all ten items was calculated.

The device with the highest value got rank one, the one with lowest score rank three, or if data was available only for two devices, rank two. In Table 4.1 the rankings for each device and each questionnaire are presented. The comparison shows that the results of the different questionnaires are inconsistent. No device got the same ranking on all questionnaires. Especially the SUMI-ratings are not supported by any other questionnaire. On all questionnaires, except for the SUMI, the Tablet PC was ranked best and the PDA least. However, these ranks do not reflect statistical significant differences. Thus in a next step the ratings for the different devices were statistically analysed.

Table 4.1. Ranks based on raw data of overall scales.

	Ranks based on			
	SUMI Global	AttrakDiff Attractiveness	SASSI Overall	SUS Overall
Tablet PC	2	1	1	1
PDA	1	3	2	2
Desktop PC	3	2	n.a.	n.a.

The SUMI's Global scale showed significant differences between the devices. The PDA was rated best, the Tablet PC was rated second, and the unimodal Desktop PC got the worst rating. For the AttrakDiff's overall scale, the Attractiveness scale, differences could also be found. But here the order was different: The Tablet PC was rated most attractive while the PDA was rated least attractive. The SUS and the SASSI showed no significant differences. The detailed results, including means and standard deviations are presented in Table 4.2.

Table 4.2. Comparison of systems on overall scales.

Scale	System	*M*	*SD*	*F*-value†/*t*-value‡	*p*
				(df)	(part.
SUMI Global	Tablet PC	40.19	7.87	6.56† (2,40)	.003** (.247)
	PDA	45.29	10.14		
	Desktop PC	38.04	7.06		
AttrakDiff Attractiveness	Tablet PC	.99	1.04	4.04† (2,38)	.026** (.175)
	PDA	.34	.88		
	Desktop PC	.74	.66		
SASSI Overall	Tablet PC	2.24	.52	2.00‡	.059
	PDA	1.96	.48	(20)	(n.a.)
SUS Overall	Tablet PC	53.93	16.59	1.32‡	.232
	PDA	50.12	13.38	(20)	(n.a.)

Note. ** $p_{2\text{-tailed}}$<.01; *$p_{2\text{-tailed}}$ <.05, †F-values and partial eta² values are reported for the SUMI and the AttrakDiff scales, ‡t-values are reported for the SASSI and the SUS scales

In a further step, the overall scales of the different questionnaires were correlated (s. Table 4.3). As indicated by the previous results, the SUMI results are least consistent with the results of the other questionnaires: For all devices the SUMI's Global scale correlated negatively with all other overall scales, except for the PDA's SUS scale. In other words, results of the SUMI's Global scale are contradictory to the results of most other questionnaires. Regarding the Tablet PC, the SASSI is highly correlated with the AttrakDiff and the SUS. For the PDA, a significant correlation could be found between SASSI and AttrakDiff.

Table 4.3. Correlations (Pearson's *r*) between overall scales.

System	Scale	AttrakDiff Attractiveness	SASSI Global	SUS Overall
Tablet PC	SUMI Global	-.71**	-.82**	-.49*
	AttrakDiff Attractiveness	-	.66**	.19
	SASSI Global	-	-	.58**
PDA	SUMI Global	-.77**	-.77**	.08
	AttrakDiff Attractiveness	-	.52*	-.34
	SASSI Global	-	-	.19
Desktop PC	SUMI Global	-.50*	n.a.	n.a.

Note. ** $p_{2\text{-tailed}}$<.01; *$p_{2\text{-tailed}}$ <.05, N=20

Pragmatic Aspects

Additionally to the overall scales, scales measuring similar constructs were investigated. The Efficiency scale of the SUMI, the Pragmatic Qualities scale of the AttrakDiff and the Speed scale of the SASSI were correlated with each other.

Again, significant differences were observed on the SUMI and on the AttrakDiff. As for the overall scales, the SUMI results and the AttrakDiff results are not in line with each other. According to the SUMI, the Desktop PC is the least efficient system and the PDA the most efficient. On the AttrakDiff the order is reversed; here, the Desktop PC was rated best and the PDA was rated worst. The SASSI indicated no significant differences (Table 4.4).

Table 4.4. Comparison of systems on scales measuring pragmatic qualities.

Scale	System	*M*	*SD*	*F*-value†/*t*-value‡ (df)	*p* (part. eta²)†
SUMI Efficiency	Tablet PC	19.00	3.39	6.19† (2,40)	.005** (.236)
	PDA	19.90	3.48		
	Desktop PC	16.67	3.15		
AttrakDiff Pragmatic Qualities	Tablet PC	0.91	.84	16.80† (2,38)	.000** (.469)
	PDA	0.01	.89		
	Desktop PC	1.34	.61		
SASSI Speed	Tablet PC	1.64	.50	.40‡ (20)	.693
	PDA	1.59	.46		

Note. ** $p_{2\text{-tailed}}$<.01; *$p_{2\text{-tailed}}$ <.05, †*F*-values and partial eta² values are reported for the SUMI and AttrakDiff scales, ‡*t*-values are reported for the SASSI scale

The results of the correlation are similar. Significant negative correlations were found for the Tablet PC and the PDA between the SUMI Efficiency scale and the AttrakDiff's Pragmatic Qualities scale (s. Table 4.5). This means that the more efficient a system is rated on the SUMI, the worse it is rated on the respective AttrakDiff scale.

For all devices, the SASSI Speed scale was neither correlated with the AttrakDiff Pragmatic Qualities scale nor with the SUMI Efficiency scale. Thus these scales do not seem to measures the same construct.

Table 4.5. Correlations (Pearson's *r*) between scales measuring pragmatic aspects.

System	Scale	AttrakDiff Pragmatic Qualities	SASSI Speed scale
Tablet PC	SUMI Efficiency	-.56**	-.22
	AttrakDiff Pragmatic Qualities	-	.31
PDA	SUMI Efficiency	-.66**	.05
	AttrakDiff Pragmatic Qualities	-	-.12
Desktop PC	SUMI Efficiency	-.38	n.a.

Note. **$p_{2\text{-tailed}}$ <.01; *$p_{2\text{-tailed}}$ <.05, N=21

Hedonic Aspects

Regarding hedonic aspects, ratings on the Affect scale from the SUMI, the Hedonic scales from the AttrakDiff, as well as the Annoyance and the Likeability scale from the SASSI were compared. Differences were observed for the SUMI Affect scale and for both Hedonic Qualities scales from the AttrakDiff. No differences were shown for both SASSI scales. However, in contrast to the results reported above, the scales measuring emotional responses are somewhat consistent across the different questionnaires. The Desktop PC was rated worst on the SUMI and the AttrakDiff (cf. Table 4.6).

Table 4.6. Comparison of systems on scales measuring hedonic qualities.

Scale	System	*M*	*SD*	*F*-value†/*t*-value‡ (df)	*p* (part. eta²)†
SUMI Affect	Tablet PC	19.33	2.13	10.02† (2,40)	.000 (.334)
	PDA	19.95	2.13		
	Desktop PC	17.38	2.73		
AttrakDiff Hedonic Qualities - Stimulation	Tablet PC	.63	.84	3.59† (2,38)	.037 (.159)
	PDA	.32	.57		
	Desktop PC	.27	.64		
AttrakDiff Hedonic Qualities - Identity	Tablet PC	.81	.81	23.03† (2,38)	.000 (.548)
	PDA	.64	.76		
	Desktop PC	-.33	.89		
SASSI Likeability	Tablet PC	2.60	.62	5.00‡ (20)	.065
	PDA	2.23	.61		
SASSI Annoyance	Tablet PC	2.49	.59	1.79‡ (20)	.089
	PDA	2.18	.64		

Note. ** $p_{2\text{-tailed}}$<.01; *$p_{2\text{-tailed}}$ <.05, †*F*-values and partial eta² values are reported for the SUMI and AttrakDiff scales, ‡*t*-values are reported for the SASSI scales

Again correlations were calculated between the Affect scale of the SUMI, the hedonic scales of the AttrakDiff, as well as the Annoyance and the Likeability scale of the SASSI.

Regarding the Tablet PC, no correlations were found between the SUMI Affect scale and all other scales. Both hedonic scales of the AttrakDiff and the SASSI Likeability scale and Annoyance scale (recoded so that higher values indicate less annoyance and hence more favourable ratings) correlated positively with each other. This means these scales measure similar constructs.

The results for the PDA show significant positive correlation within but not between the questionnaires. Both AttrakDiff scales correlated with each other as well as both SASSI scales.

The ratings of the Desktop PC system show significant positive correlations between the Hedonic Qualities-Identity scale from the AttrakDiff and the SUMI Affect scale. Only for this measurement the SUMI was in agreement with results from other questionnaires.

Compared to the scales measuring other constructs, the scales measuring emotional aspects show the highest agreement across the different questionnaires. All results are shown in Table 4.7.

Table 4.7. Correlations (Pearson's r) between scales measuring hedonic aspects.

		AttrakDiff Hedonic Qualities -Stimulation	AttrakDiff Hedonic Qualities -Identity	SASSI Likeability	SASSI Annoyance
Tablet PC	SUMI Affect	.12	.02	.13	-.48
	AttrakDiff Hedonic Qualities-Stimulation	-	.735**	.627**	.62**
	AttrakDiff Hedonic Qualities-Identity	-	-	.775**	.67**
	SASSI Likeability	-	-	-	.81**
PDA	SUMI Affect	.20	.05	.04	-.12
	AttrakDiff Hedonic Qualities- Stimulation	-	.68**	.40	.13
	AttrakDiff Hedonic Qualities-Identity	-	-	.10	.31
	SASSI Likeability	-	-	-	.73**
Desktop PC	SUMI Affect	.31	.49*	n.a.	n.a.
	AttrakDiff Hedonic Qualities-Stimulation	-	.64**	n.a.	n.a.

Note. $^{**}p_{\text{2-tailed}} <.01$; $^{*}p_{\text{2-tailed}} <.05$, N=21

Questionnaire Data and Interaction Data

Although interaction data do not necessarily match the quality judgments (cf. Section 3.2.2), the efficiency-related scales from the different questionnaires were correlated with task duration. The purpose was to investigate possible relations between the different data types and to integrate these results with the results reported above.

Due to the large variance between the participants regarding the task duration, all questionnaire ratings and task durations were transformed into ranks for each participant. Concerning the questionnaire ratings, rank one was assigned to the system with the best (most favourable) rating. For task duration, the system with the shortest task duration got rank one. Thus, positive correlations show concordance between the performance measures (task duration) and the direct quality measures (questionnaire ratings).

The ranks based on SUMI's Efficiency scale correlated negatively with task duration. A positive correlation could be observed between the ranks based on the AttrakDiff's Pragmatic Qualities scale and task duration. Also the rank transformed SASSI Speed scale was positively correlated with task duration. Table 4.8 shows the detailed results.

Table 4.8. Correlations (Kendall's τ_b) between ranks based on task duration and ranks based on sub-scales measuring efficiency.

Ranks based on	Task Duration
SUMI Efficiency scale (N=45)	-.58**
AttrakDiff Pragmatic Qualities scale (N=45)	.53**
SASSI Speed scale(N=30)	.32*

Note. **$p_{\text{2-tailed}}$ <.01; *$p_{\text{2-tailed}}$ <.05

Regarding the overall scales, the SUMI Global scale showed a negative correlation with task duration. All other overall scales were not significantly correlated with task duration (see Table 4.9).

Table 4.9. Correlations (Kendall's τ_b) between ranks based on task duration and ranks based on overall scales.

Ranks based on	Task Duration
SUMI Global scale (N=45)	-.42**
AttrakDiff Attractiveness scale (N=45)	.02
SASSI Overall scale (N=30)	.23
SUS Overall scale (N=30)	.26

Note. **$p_{\text{2-tailed}}$ <.01; *$p_{\text{2-tailed}}$ <.05

4.1.3 Discussion

The questionnaire ratings most inconsistent to the results of all other questionnaires as well as to the task duration data were the SUMI ratings.

Regarding the overall scales the best system according to the SUMI is the PDA. On the overall scales of the AttrakDiff and the SASSI, the Tablet PC got the highest ratings. The SUS revealed no significant difference between the systems.

Further differences where shown for the sub-scales: Solely on the SUMI the PDA was rated best regarding efficiency. The results of the AttrakDiff's Pragmatic Qualities scale, which also aims to measure efficiency-related aspects, and of the SASSI's Speed scale, are in contrast to the SUMI results. Both, AttrakDiff as well as SASSI, indicate that the Tablet PC is most efficient and the PDA least efficient.

In addition, concerning the SUMI Affect scale, where the PDA was rated best too, the results are inconsistent with similar scales of the other questionnaires. The AttrakDiff implies that the Tablet PC has more Hedonic Qualities than the other systems. The SASSI scales Likeability and Annoyance also point to the Tablet PC as the system most fun to use. Only regarding the Desktop PC system consistency between the SUMI and the other questionnaires could be shown.

Results from the comparison between the interaction data and the quality judgement support these findings: The SUMI results showed correlations in the "wrong" direction for the Global scale and task duration and for the Efficiency scale and task duration. That means the longer the task duration, the better was the rating on the SUMI. Regarding the other questionnaires' global scales, no significant correlation with task duration was observed. According to these results, the global usability is hardly affected by the system's actual efficiency in terms of task duration. But, as mentioned previously, the actual efficiency might not equal the perceived efficiency (Hornbæk, & Law, 2007). However, the specific scales associated with efficiency of the AttrakDiff and the SASSI were in agreement with the actual task duration. Thus, it might as well be the case that the subject's perceptions of efficiency were indeed corresponding to the actual efficiency, but efficiency itself is not contributing very much to the overall judgment.

In summary, based on the questionnaire results it remains unclear which system is the one with the best usability. Rather it is shown that questionnaires designed for unimodal systems are not applicable for usability evaluation of multimodal systems, since they seem to measure different constructs. The questionnaires with the most concordance were the AttrakDiff and the SASSI. Moreover, these two questionnaires were also in highest agreement with the interaction data. In addition, the AttrakDiff was more sensitive regarding differences between the systems than the other questionnaires.

A possible explanation could be that the kind of rating scale used in the AttrakDiff, the semantic differential, is applicable to all systems. The semantic differential uses no direct questions but pairs of bipolar adjectives, which are not linked to special functions of a system. The SASSI uses direct questions, but was specifically developed for the evaluation of systems with speech input, and may therefore be more suitable for multimodal systems including voice control than questionnaires developed for GUI-based systems.

Furthermore, all SUMI questions were included although some of them are only appropriate for the evaluation of market-ready interfaces. These inappropriate questions may have affected the SUMI results.

Regarding the relation with the interaction data, two explanations are possible: Either the SUMI reflects just the perceptions of efficiency, which differed largely from the actual efficiency, or with a lack of construct validity of the SUMI scale. The latter is more likely, as the SASSI and the AttrakDiff indicated that for the reported experiment the perceptions matched the performance measures.

4.1.4 Conclusion

To conclude, the questionnaire assessing quality perceptions should be chosen carefully as an unsuitable instrument might provide inconclusive or misleading results. Furthermore, a reliable, valid, and more specific questionnaire for multimodal interfaces is desirable. In view of the results reported, the AttrakDiff provides a proper basis for this.

Yet, one of the AttrakDiff's advantages, its relatively high level of information due to the items not being linked to specific functions, may turn out to be a disadvantage when the results are reported to the developers or to the project management. Although the AttrakDiff seems to be a reliable instrument, it is difficult to derive concrete design recommendation to improve the interface; e.g., if the results indicate the interface lacks pragmatic qualities, little is known about the exact feature, which needs to be improved.

One might argue, that using a quantitative questionnaire as the only measuring instrument in an evaluation study may just be insufficient for an in-depth evaluation aiming to result in design recommendations, and that a multi-method approach including quantitative and qualitative data should rather be applied.

However, according to an analysis of papers published at the ACM CHI conference, most researchers gather quantitative data only (Barkhuus & Rode, 2007); qualitative evaluations or studies assessing both kinds of data are considerably less common. As the CHI conference is considered as one of the key conference (Greenberg & Buxton, 2008), the "gold standard" in the field of HCI, it is plausible to assume that these results show a general tendency in usability research. Please note, that as reported in (Bartneck & Hu, 2009) the typical CHI author's major affiliation is a university (~62%) rather than a company (~21%) or another type of institute (~16 %). Hence, the studies reported above reflect academic practice; usability evaluation may be conducted differently and maybe with more employment of qualitative methods in industry. Unfortunately, the methodological approaches actually used in industry are rarely published (Roto, Obrist, & Väänänen-Vainio-Mattila, 2009), one of the few studies including participants from industrial contexts indicates that the evaluation strategy indeed differ. Nevertheless questionnaires are assumed to be a satisfying (Väätäjä & Roto, 2009) and popular (Wechsung, Naumann, & Schleicher, 2008) instrument for both, industry and academia

Thus, the AttrakDiff was chosen as a starting point for the development of a new questionnaire for the evaluation of multimodal interfaces for which a major requirement was to provide more specific information than the AttrakDiff. At the same time, the new instrument should be widely applicable to a large variety of multimodal systems.

4.2 Item Generation and Item Selection

Based on the results reported in the previous section, a first set of items was generated. The AttrakDiff's item format, the semantic differential, was adopted. To avoid the information being too high level, the items were grouped into five different blocks. The purpose was to be able to track down problems to specific properties of the system. The first block asks to rate the system, the second block refers to the interaction, the third to the feedback, the fourth to the design and the fifth two the input modalities. The aspects to rate should reflect the directly perceivable part of the system, which Donald Norman (1988) labelled as the *system image*. According to Norman (1988), the designer and the user of a system can only communicate via the system image. It comprises the system's appearance, its way of response and its operation as well as the manuals. Except for manuals all parts of the system's image are covered; the appearance corresponds to the design block, the ways of response to the feedback block, the system's operation to the interaction and input modalities block. The system block contained attributes, which are linked to several parts of the system's image. Note that the blocks do not equal the dimensions, or the assumed scales, respectively; the blocks contained items of different dimensions or scales.

As a theoretical framework, the taxonomy presented in Möller et al. (2009) was used. The taxonomy by Möller et al. was also the starting point for the taxonomy presented in Chapter 3. Hence, the concepts related to the quality apects, the third layer of the taxonomy, explained in Chapter 3 are largely the same as the concepts by Möller et al. (2009) except for the following differences: Möller and colleagues assume the concept appeal, which is seen as a result of the system's personality and aesthetics. For reasons explained in Section 3.3.5, the concept appeal was not adopted but another concept, namely discoverability, was introduced. In contrast to appeal, discoverability is allocated on the same level as personality and aesthetics.

Furthermore Möller et al. (2009) suggest intuitivity as a separate aspect of ease of use. However, in the taxonomy presented in Chapter 3, intuitivitiy is understood as a sub-aspect of learnability, an understanding that is based partly on theory (cf. Section 3.3.4) as well as on the results of the empirical validation of the taxonomy, which is presented in the following sections. Note, that in the initial item pool, items for both constructs, learnability and intuitivity, were formulated. However, later in the development process, it was discovered that merging both constructs into one offered a better fitting model, based on the empirical data.

In a first step, 74 items in the form of a semantic differential, i.e. each item consists of a bipolar pair of adjectives, were generated covering all quality-of-experience concepts of the taxonomy. One reason for choosing the semantic differential format was that also the AttrakDiff uses this format. Another cause was that relevant literature shows that respondents use formal properties of a scale to interpret the intended meaning (Armitage & Deeprose, 2004). If a unipolar scale is presented, respondents are more likely to recognize the underlying construct as unipolar whereas a bipolar scale indicates a bipolar construct. Gannon and Ostrom (1996) showed that an explicitly unipolar scale (e.g. *not all honest – completely honest*) activates only one category, namely honesty, whereas an explicitly labelled bipolar scale (e.g. *very dishonest – very honest*) activates two different categories, namely honesty and dishonesty. According to Gannon and Ostrom (1996), judging someone's honesty may involve other cognitive processes than judging someone's dishonesty. Consequently, the same question may communicate different contents, if the answer format is changed.

Moreover, the perception of a scale is context dependent. In the context of affect measurement, unipolar scales measuring affect are perceived as being bipolar (Russell & Carroll, 1999a), as positive feelings (e.g. happiness) are believed to be the opposite of negative feelings (e.g. sadness). However, even though emotions might typically be bipolar, research indicates that people can feel happy and sad at the same time (Larsen, McGraw, & Carpaccio, 2001). At this point one of the main disadvantages of bipolar scales becomes apparent. Such scales cannot distinguish between ambivalent, more complex emotions and neutrality (Kaplan, 1972). Hence, there are different positions in how to measure affect: Schimmack and colleagues (Schimmack, Boeckenholt, & Reisenzein, 2002) advocate strict unipolar answer formats. They argue that if positive affect and negative affect are not bipolar, bipolar answer formats are not applicable as these scales force the participants to integrate negative and positive affect into a single judgement. Russell and Caroll (1999b) on the other hand recommend bipolar scales for measuring affect for the reason already mentioned above: If unipolar scales might be perceived as bipolar by the respondents, the researcher cannot clearly interpret the results. Thus using a bipolar scale maximizes the chances that results can be interpreted in a meaningful way.

A seven-point answer format was chosen, because, as indicated by prior research (for an overview see Krosnick & Fabrigar, 1997), reliability is highest for seven, and validity for five to nine response categories.

With these first 74 items a small expert evaluation and a pilot study were conducted. Aim of the expert evaluation was to ensure that the pairs are actually bipolar adjectives. The expert evaluators were asked to indicate all items, which they considered as not clearly bipolar. An example pair rated as not bipolar was *aufheiternd* vs. *deprimierend* (*exhilarating* vs. *depressing*). *Deprimierend* was considered as having a too strong negative meaning compared to the meaning of *aufheiternd*. Additionally, the experts were asked to sort out ambiguous items, items using unusual wording, or items not matching the intended quality-of-experience aspect clearly enough. An example for an item marked as problematic was *stressig* vs. *unstressig* (*stressful* vs. *unstressful*), as *unstressig* was considered as non-standard German. Furthermore, items intended to measure cooperativity were assigned to other constructs or no consensus was found whether the items measure cooperativity or another construct. Hence, the concept was dropped, as it seems to be very difficult to formulate respective items in the form of the semantic differential. The items initially intended to measure cooperativity were mostly assigned to input quality and output quality, and to the concepts associated with ease-of-use. The initial list of items can be found in Appendix B.1.

In parallel, the initial questionnaire was used in two laboratory studies to assess empirical characteristics of the items. Both studies were conducted with the same system, namely the *Sprachbox* system, which is a multimodal voice and mailbox. Detailed descriptions of the studies are presented in the Appendix A.2. The aim was to get a first assessment of the items statistical properties and to identify items which were not easily understandable by non-experts. This was the case for the item *ergonomisch* vs. *unergonomisch* (*ergonomic* vs. *non-ergonomic*). It was eliminated, as some participants did not know the meaning of *ergonomic*.

Overall, 38 completely filled in questionnaires were gathered (18 female, 20 male, M_{age} = 38 years, SD_{age}= 18 y.). The items were tested for normality using the Jarque-Bera-Test. This test was chosen as the Shapiro-Wilk test, the probably most common test of normality, is very sensitive to ties (DeCarlo, 1997). Using a 7-point scale, many ties were expected in the data. Analysis showed that only one item significantly differed from the Gaussian normal distribution. Jarque-Bera values and *p*-values are presented in Appendix B.1.

In a further step the item difficulty indices were calculated in accordance with Moosbrugger and Kelava (2007), as follows.

$$P_i = \frac{\sum_{v=1}^{N} x_{vi}}{N * \max(x_i)} \cdot 100 \qquad (4.1)$$

with

P_i = difficulty of item *i*

$\sum_{v=1}^{N} x_{vi}$ = sum of score actually achieved by all *N* participants on item *i*

$N * \max(x_i)$ = maximum score achievable by all *N* participants on item *i*

The item difficulty indices indicate the rate of approval towards an item and can range between 0 and 100. For the current sample, the indices varied between 47 and 72 and were thus all within the recommended range between 20 and 80 (cf. Bortz & Döring, 2007).

Furthermore, item discrimination indices were calculated as the corrected item-scale correlation (Bühner, 2011). According to Bühner (2011), the corrected item discrimination of item *i* is defined as:

$$r_{i(s-i)} = \frac{r_{is} \cdot SD_S - SD_i}{\sqrt{SD_S^2 + SD_i^2 - 2 \cdot r_{is} \cdot SD_S \cdot SD_i}} \quad (4.2)$$

with

$r_{i(s\text{-}i)}$ = item discrimination coefficient of *i* and scale *s*, where item *i* is not included in the scale`s *s* score

r_{is} = correlation between item *i* and scale *s*

SD_s = standard deviation of scale *s*

SD_i = standard deviation of item *i*

Item discrimination indices were between .30 and .87 indicating medium to high discrimination. Hence, no further items were excluded.

Overall 27 items were eliminated, most of them due to the experts comments. Thus, 47 items remained in the pool. The complete list of these items is given in Appendix B.1, excluded items are indicated.

4.3 Construction of the Final Questionnaire

4.3.1 Validation of the Constructs

The 47 item questionnaire was subsequently used in three different laboratory studies involving three different multimodal systems. In all studies, ratings were collected after participants interacted with those systems. Details on the studies are presented in Section 6.2, Section 6.3, and in Appendix A.2.3.

All prototypes offered at least speech and touch control. Together with the initial studies, 244 completely filled in questionnaire were gathered. Again, incomplete questionnaires were checked for missing data patterns

As the questionnaire was based on the theoretical assumptions of Möller et al. (2009), a mixed method approach, employing exploratory and confirmatory methods, was chosen to identify items matching the intended dimensions best. Confirmatory approaches are designed to test pre-defined factor structures, like the structure of the taxonomy, whereas exploratory analyses do not require any hypothesized factors structures (Bühner 2010). The development process as described in Homburg and Giering (1996) was largely followed.

All of the following analyses were conducted using a random sample (25%, N=57) of the whole data set, to be able to validate the final model with the other part of the data set.

Based on the recommendation by Homburg and Giering (1996) Cronbach's α, the internal consistency, was calculated for each quality aspect of the taxonomy by Möller et al. (2009), except for Cooperativity and Stimulation. All items originally intended to measure Cooperativity were assigned to other constructs (cf. Section 4.2.). The aspect Stimulation was renamed into Discoverability for reasons explained in Section 3.3.5. Moreover, the concept Efficiency was split into Temporal and Mental Efficiency.

As shown in the Table 4.10, values for Cronbach's *a* were higher than the .7 which is defined as the minimum by Homburg and Giering (1996). Accordingly, none of the items had to be removed.

As the next step, Homburg and Giering (1996) advise calculating an exploratory factor analysis for each construct, to ensure that the items for each construct actually form only one factor. Furthermore, each of the single factors is required to explain at least 50% of the variance. Accordingly, exploratory Maximum-Likelihood (ML) factor analyses were calculated for each construct. The ML methods allows for testing the fit between the hypothesized factor structure (defined number of factors) and the data structure via the χ^2-goodness-of-fit test. If the test shows a non-significant result ($p > .05$) the null hypothesis, which assumes that the hypothesized factor structure fits the data structure, is kept. Hence, if all items of one construct are loading on one factor, the χ^2-goodness-of-fit should be non-significant.

For constructs assessed with three or less items, the χ^2-goodness-of-fit could not be calculated, as the degrees of freedom would have been zero or negative.

For Output Quality and Intuivity the χ^2-goodness-of-fit test was significant, indicating a poor fit for the single factor solution. To identify the items not loading on the same factor, a ML factor analysis with two factors was calculated. As the two-factor solution showed a good fit, a three-factor solution was not investigated. Items loading highest on the second factor and/or lowest on the first factor were excluded. After the respective items were removed, the fit of another single-factor solution was investigated for the remaining items in order to check that they show a single factor structure. For both constructs, Output Quality and Intuitivity, one item[2] had to be excluded to achieve the recommended criteria. Results for all constructs are shown in Table 4.10. After this step, 45 items remained.

[2] The excluded items were INTUI1 ("*Das System ist schwierig zu bedienen - einfach zu bedienen.*"; translation: "The system is difficult to operate - easy to operate.") and OQ2 *(„Die Rückmeldungen des Systems sind unnötig - notwendig."*; translation: "The system's feedback is unnecessary - necessary.").

Table 4.10. Cronbach's α, explained variance and results of χ^2-test for all sub-constructs

	Cronbach's α	Explained Variance	N of Items	χ^2 (df)	p
Personality	.87	74.15	4	2.43 (2)	.290
Discoverability	.93	73.29	6	11.67 (9)	.233
Aesthetics	.89	82.03	3	n.a.	n.a.
Temporal Efficiency	.92	82.67	4	5.94 (2)	.051
Mental Efficiency	.87	88.09	2	n.a.	n.a.
Intuitivity	.81	65.08	4	1.82 (2)	.403
Effectiveness	.82	73.30	3	n.a.	n.a.
Learnability	.83	74.42	3	n.a.	n.a.
Output Quality	.92	65.09	8	31.13 (20)	.054
Input Quality	.96	78.45	8	25.68 (20)	.176

Next, all sub-constructs were modeled using AMOS, a software package for structural equation modelling. Structural equation models can be seen as a combination of factor analyses and multiple regression analyses (Amelang et al., 2006). In contrast to exploratory methods, structural equation models allow to not only investigate relations between manifest, observable variables, but to also investigate and test relations between latent, not directly observable, constructs, like intelligence or perceived quality (Amelang et al., 2006). Based on previous theoretical assumptions, a model specifying the relations between the latent and manifest variables is postulated. Then, it is tested if the hypothesized structure of the model fits the empirical data.

In line with the assumptions of the taxonomy for each of the models, a single-factor structure was employed. As only the 25% of the sample was used, the sample size is rather small. However, according to Iacobucci (2010), the number of cases (N=57) is still large enough for well-performing models.

For all concepts, the following criteria, suggested by Homburg and Giering (1996), were employed:

- **Indicator reliability (IR) ≥ 0.4**
  The indicator reliability is the squared factor loading (regression weight) which equals the squared multiple correlation of the item with the factor. It is the part of the item's variance explained by the factor (Bühner, 2011). Indicator reliability ranges from 0 to 1. Higher values indicate higher reliability.

- **Composite reliability (CR) ≥ 0.6**
  The composite reliability (also known as factor reliability) describes how well the factor can be measured via the items and is accordingly a measure of consistency of the factor (Ruge, 2011). Composite reliability ranges from 0 to 1. Higher values indicate higher reliability.

- **Quotient between χ^2 and the degrees of freedom (df) ($\frac{\chi^2}{\text{df}}$) ≤ 3**
  The χ^2-test checks the null hypothesis ('the theoretical model fits the data structure') against the alternative hypothesis ('the theoretical model does not fit the data structure'). Accordingly, a not significant χ^2-test and a high *p*-value respectively are desired (Bühner 2010). As *p*-values are decreasing with increasing sample sizes, the quotient between the χ^2 and the degrees of freedom (df) is used (Homburg & Giering, 1996).

- **Average variance extracted (AVE) ≥ 0.5**
  The average variance extracted indicates the part of the variance explained by the factor compared to the part of the variance resulting from the measurement error (Fornell & Larcker, 1981). Like the CR, the AVE is a measure of the internal consistency of the factor. The AVE ranges between 0 and 1. Higher values indicate higher consistency.

Homburg and Giering (1996) suggest using two further criteria. These are:

- **Goodness of fit index (GFI) ≥ 0.9**
  The GFI compares the proposed model with a saturated model, a model perfectly fitting the data (Bühner 2011). The GFI is a squared multiple correlation coefficient (R^2) indicating the amount of variance explained by the model. The GFI is not adjusted for the degrees of freedom; or, in other words, adding a parameter to a model, for instance a correlation between two factors, would automatically enhance its fit. The GFI ranges between 0 and 1 with 1 indicating a perfect fit.
- **Adjusted goodness of fit index (AGFI) ≥ .9**
  The AGFI is, like the GFI, a descriptive measure of goodness of fit but in contrast to the GFI the AGFI uses *penalty terms* dependent on the number of parameters. (Homburg & Giering, 1996). The AGFI ranges between 0 and 1 with 1 indicating a perfect fit.

However, both measures have been shown to be heavily dependent on the sample size and some authors strongly advise against them (Bühner, 2011; Hu & Bentler, 1999). Hence, they will only be reported but will not serve as criteria. Instead the following criteria recommended by Bühner (2011) will be used:

- **Comparative fit index (CFI) ≥ .95**
  The CFI compares the proposed model with a *null model*, where all variables are

uncorrelated (Bühner, 2011). If the proposed model is better than the null model the CFI increases, i.e. the fit gets better. The CFI ranges between 0 and 1, with 1 indicating a perfect fit.

- **Root mean square error of approximation (RMSEA) ≤ 0.08**
  The RMSEA is a so-called badness-of-fit- measure, reflecting the deviation of the observed variance from the hypothesized variance (Bühner, 2011).

As some of the concepts were measured by three or less items, the corresponding models were *just identified* or *under-identified*. If three items were used, models could be just identified and the applicable criteria were employed. The GFI, AGFI, CFI and RMSEA could not be used, as the fit is always perfect for just identified models. For Mental Efficiency, represented with only two items, no model could be built as it would have been under-identified. Under-identified models are models, which do not have enough observed parameters (e.g measured variables) compared to parameters to be estimated.

If several of the above criteria were violated, items showing low indicator reliability were excluded until the criteria were met. For Personality, Discoverability, Aesthetics, Mental Efficiency, Effectiveness, Learnability and Intuitivity no items had to be excluded. Although for each, Effectiveness and Learnability, one item showed a slightly to low indicator reliability, the item was kept as the values for the two other criteria assessed were sufficient. For Temporal Efficiency one item was excluded and for Output Quality three items were removed[3]. Table 4.11 summarizes the criteria explained above for each construct after the removal. After this step, 41 items were left in the item pool for further analyses.

Table 4.11. Fit indices for all constructs.

	IR Min./Max	CR	$\frac{\chi^2}{df}$ *(p)*	AVE	CFI	*RSMEA*	GFI	AGFI	N of Items
Personality	.48/.83	.89	1.28 (.28)	.66	1	.07	.98	.89	4
Discoverability	.50/.95	.93	1.38 (.32)	.68	.99	.08	.94	.86	6

[3] The excluded items were EFFIT1 (*„Die Interaktion mit dem System ist umständlich - direkt.“*; translation: The interaction with the system is cumbersome - direct.), OQ1 (*„Die Rückmeldungen des Systems sind dumm - klug.”*; translation: “The system’s feedback is dumb - smart.”). OQ7 (*„Die Rückmeldungen des Systems sind zweckdienlich -zwecklos“*; translation: “The system’s feedback is expedient – futile.”), and OQ8 (*„Die Rückmeldungen des Systems sind konstruktiv-destruktiv.”*; translation: “The system’s feedback is constructive - destructive.”).

Aesthetics	.60/.92	.89	n.a.	.74	n.a.	n.a.	n.a.	n.a.	3
Temporal Efficiency	.71/.92	.94	n.a.	.85	n.a.	n.a.	n.a.	n.a.	3
Mental Efficiency	n.a.	n.a	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	2
Intuitivity	.40/.75	.72	.96 (.38)	.54	1	<.001	.98	.91	4
Learnability	.39/.77	.77	n.a.	.63	n.a.	n.a.	n.a.	n.a.	3
Effectiveness	.39/.89	.83	n.a.	.62	n.a.	n.a.	n.a.	n.a.	3
Input Quality	.64/.85	.95	1.39 (.12)	.75	.98	.08	.91	.80	8
Output Quality	.53/.79	.88	.99 (.43)	.68	1	<.001	.97	.90	5

4.3.2 Validation of the Meta-Constructs

In the next development stage, for each of the three meta-constructs, which are Joy of Use, Ease of Use and Interaction Quality, the structure was tested using exploratory factor analyses. Aim was to investigate if the proposed factor structure was found in the data. ML factor analyses with Promax rotation were carried out. Promax rotation was chosen over Varimax because Promax rotation allows the factors to correlate (Bühner, 2011). Correlations between the constructs were expected, for example a higher degree of effectiveness may lead to more efficient interactions (Frøkjær, Hertzum, & Hornbæk, 2009).

For each construct, the number of factors to be extracted was fixed to the number of constructs. For example, three factors were extracted for Joy of Use as three constructs (Personality, Discoverability, and Aesthetics) were assumed. Cross- and non-loading items were removed. For Joy of Use, three items, one item of each sub-construct, were removed[4]. A χ^2-goodness-of-fit test showed that the supposed three-factorial structure fits the remaining data well, $\chi^2(25, N=57)=25.57$, $p=.431$. The pattern matrix of the resulting solution is presented in Table 4.12.

[4] The excluded items were PER4 (*"Die Interaktion mit dem System ist nervig - spaßig."*; translation: "The interaction with the system is nerved - amusing."), DISC4 (*"Die Gestaltung des Systems ist konventionell - originell."*; translation: "The design of the system is conventional - original.") and AEST1 (*"Die Gestaltung des Systems ist unansehnlich - ansehnlich."*; translation: "The design of the system is unsightly - sightly.").

Table 4.12. Pattern matrix of final solution for Joy of Use (three items removed, N=57)

	Factor 1 (Discoverability)	Factor 2 (Personality)	Factor 3 (Aesthetics)
PER1	.02	**.93**	.02
PER2	.22	**.66**	.03
PER3	.04	**.51**	.29
DISC1	**.56**	.18	.03
DISC2	**.82**	.13	-.21
DISC3	**.66**	.10	.08
DISC5	**.76**	-.07	.27
DISC6	**.81**	-.05	.23
AEST1	.00	.17	**.64**
AEST2	-.01	.00	**1.00**

Note. Highest loadings for each item are in bold-face.

Ease of Use was assumed to comprise the five sub-constructs Temporal Efficiency, Mental Efficiency, Intuitivity, Learnability and Effectiveness. Hence, five factors were extracted. However, most of the Intuitivity items were loading highly on the same factor as the Learnability items and one item showed a high loading on the factor Mental Efficiency (see Table 4.13). However, Intuitivity may be interpreted as an attribute of Learnability (Laakkonen, 2007). Moreover, interacting with a highly intuitive interface should keep the mental effort low. According to Mohs et al. (2006) a system is intuitive to use if effective interaction is possible by applying prior knowledge unconsciously. Unconscious information processing imposes only minimal cognitive load. Thus, Intuitivity may not be understood as an independent quality aspect.

These assumptions were tested and a four factors solution was calculated. The χ^2-goodness-of-fit indicated that the hypothesized four factor structure does not differ significantly from the data structure, $\chi^2(51, N=57)=56.92$, $p=.264$. Consequently, the concept Intuitivity was dropped. The items were assigned to Learnability and Mental Efficiency respectively.

The four factors solution was further investigated regarding cross- and non-loading items. Subsequently three items were removed: One Effectiveness item and two items of the merged Learnability/Intuitivity construct[5]. The goodness-of-fit-test

[5] The excluded items were EFFEC3 ("*Die Gestaltung des Systems ist ablenkend - zielführend.";* translation: "The design of the system is distracting - targeted."), LEARN3 (*"Die Gestaltung des Systems ist unangebracht - angebracht.";* translation: "The design of the system is inappropriate - appropriate.") and INTUI5 (*"Die Gestaltung des Systems ist*

of the final solution (cf. Table 4.14) indicated a satisfactory fit, $\chi^2(24, N=57)=18.76$, $p=.765$.

Table 4.13. Pattern matrix of initial five factors solution for Ease of Use (no items removed, N=57)

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
EFFIT2	.06	**1.01**	-.11	-.08	.09
EFFIT3	-.07	**.60**	.13	.32	.05
EFFIT4	-.05	**.99**	.12	-.08	-.08
EFFIM1	.08	.00	**.78**	.14	-.16
EFFIM2	-.05	.14	**.77**	.06	.03
INTUI2	.00	-.02	**.87**	-.14	.14
INTUI3	**.71**	-.22	.01	.10	.16
INTUI4	**.92**	.00	.06	-.07	-.06
INTUI5	.24	.12	.03	.11	**.34**
LEARN1	**.90**	.25	-.09	-.09	-.06
LEARN2	**.58**	.21	-.01	.16	.00
LEARN3	.02	.02	.04	-.01	**.96**
EFFEC1	.32	.03	.02	**.55**	-.07
EFFEC2	-.08	-.07	-.03	**1.09**	.03
EFFEC3	.52	-.12	**.43**	-.05	.07

Note. Highest loadings for each item are in bold-face.

inkonsistent - konsistent."; translation: "The design of the system is consistent - inconsistent.").

Table 4.14. Pattern matrix of final solution for Ease of Use (three items removed, N=57)

	Factor 1	Factor 2	Factor 3	Factor 4
	(Learnability)	(Temporal Efficiency)	(Mental Efficiency)	(Effectiveness)
EFFIT2	.05	**1.01**	-.07	-.06
EFFIT3	-.05	**.61**	.14	.32
EFFIT4	-.09	**1.01**	.07	-.07
EFFIM1	.07	-.04	**.73**	.13
EFFIM2	-.02	.12	**.81**	.04
INTUI2	.08	-.01	**.88**	-.14
INTUI3	**.75**	-.23	.11	.09
INTUI4	**.93**	-.03	.06	-.09
LEARN1	**.82**	.27	-.08	-.07
LEARN2	**.57**	.21	.02	.15
EFFEC1	.27	.05	-.03	**.56**
EFFEC2	-.06	-.06	-.02	**1.09**

Note. Highest loadings for each items are in bold-face.

For Interaction Quality, one item was removed for each, Output Quality and Input Quality.[6] The χ^2- test indicated a good fit between the hypothesized two-factorial structure and the data , $\chi^2(34, N=57)=46.84$, p =.070. The pattern matrix is presented in Table 4.15.

Table 4.15. Pattern matrix of final solution for Interaction Quality (two items removed, N=57)

	Factor 1	Factor 2
	(Input Quality)	(Output Quality)
OQ3	.18	**.73**
OQ4	-.02	**.92**
Q5	.11	**.77**
OQ6	-.04	**.83**
IQ1	**.79**	.16
IQ2	**1.03**	-.11
IQ3	**.72**	.16
IQ4	**.89**	-.05
IQ5	**.68**	.16
IQ7	**.81**	.07
IQ8	**.83**	.09

Note. Highest loadings for each items are in bold-face.

[6] The excluded items were OQ9 (*"Die Rückmeldungen des Systems sind zureichend – unzureichend."*; translation: "The system's feedback is sufficient - insufficient") and IQ6 (*"Die verschiedenen Eingabemöglichkeiten sind schlecht zu koordinieren - sind gut zu koordinieren."*; translation: "The different input modalities are poorly to coordinate - well to coordinate.").

As a next step in the procedure, the 33 remaining items were used to model each meta-construct (Joy of Use, Ease of Use, Interaction Quality) using structural equation modelling. Again, the criteria explained above were checked. Additionally, the *discriminant validity* was investigated with the Fornell-Larcker-criterion. With the discriminant validity, it is tested if concepts (e.g. Ease of Use and Joy of Use), which are assumed to be different, actually measure different constructs:

- **Fornell-Larcker-criterion: AVE (Factor$_x$) > R² (Factor$_x$, Factor$_y$)**
  According to the Fornell-Larcker-criterion the average variance extracted (AVE, see above) of a factor has to be higher than each squared correlation (R^2) of this factor with another factor.

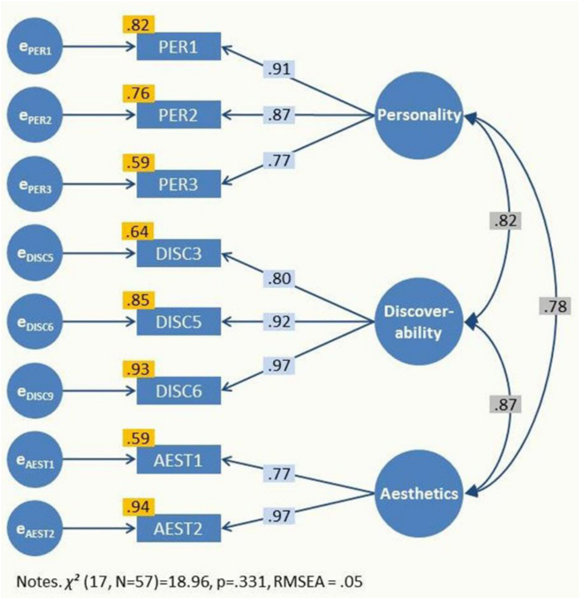
Another method to assess discriminate validity is the χ^2-difference-test. The Fornell-Larcker criterion was chosen over the χ^2-difference-test as it is the more rigorous criterion (Homburg & Giering, 1996).

Among the concepts of Joy of Use, the average variance explained was not sufficient for Discoverability. In particular, the squared correlation between Discoverability and Aesthetics was higher than the average variance explained of Discoverability (cf. Table 4.16). Hence, the Discoverability items with the lowest loadings were successively excluded.[7] After two items were removed, the Fornell-Larcker-criterion was met. The final model is given in Figure 4.2.

[7] The excluded items were DISC1 (*"Die Interaktion mit dem System ist entmutigend - motivierend.*"; translation: "The interaction with the system is discouraging - encouraging.") and DISC2 (*"Die Interaktion mit dem System ist langweilig - unterhaltsam."*; translation: "The interaction with the system is boring - entertaining.").

Table 4.16. Fornell-Larcker-criteria for Joy of Use

	Factor_X	N of Items	AVE (Factor_X)	Factor_y	R^2 $(\text{Factor}_X, \text{Factor}_y)$	Fornell-Larcker-Criterion met? $(\text{AVE} > R^2)$
Model 1	Personality	3	.73	Aesthetics	.62	Yes
				Discoverability	.67	Yes
	Aesthetics	2	.77	Personality	.62	Yes
				Discoverability	.76	Yes
	Discoverability	5	.69	Personality	.67	Yes
				Aesthetics	.76	No
Model 2	Personality	3	.73	Aesthetics	.62	Yes
				Discoverability	.66	Yes
	Aesthetics	2	.77	Personality	.62	Yes
				Discoverability	.76	Yes
	Discoverability	4	.74	Personality	.66	Yes
				Aesthetics	.76	No
Model 3	Personality	3	.73	Aesthetics	.61	Yes
				Discoverability	.65	Yes
	Aesthetics	2	.77	Personality	.61	Yes
				Discoverability	.76	Yes
	Discoverability	3	.81	Personality	.65	Yes
				Aesthetics	.76	Yes



Notes. χ^2 (17, N=57)=18.96, p=.331, RMSEA = .05

e = error (e.g. measurement error)

↔ = correlations

← = regression path

■ = observed variable (item)

● = unobserved, latent variable (underlying factor /assumed construct or error variable),

■ = multiple squared correlation coefficients (R^2, item communalities, item reliabilities)

■ = partial standardised regression weights (factor loadings),

■ = factor correlation coefficients.

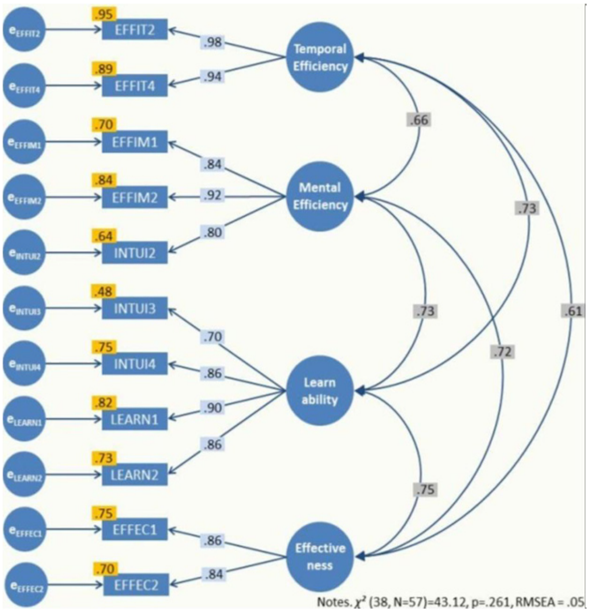
Fig. 4.2. Final model of Joy of Use after removal of items DISC1 and DISC2.

For Ease of Use the Fornell-Larcker-criterion was met for each sub-construct. However, the model fit was insufficient, $\chi^2(48, N=57)=74.79$, p = .008, RMSEA = .10, CFI = .95. Hence, the modification indices, which are automatically computed by AMOS, were investigated. Modification indices describe which additional parameters (i.e. paths, loadings or regression weights and correlations or covariances) to specify, in order to improve the model fit (Bühner, 2011). Highest modification indices were observed for the item EFFIT3. This item shared high error covariance with the item EFFEC2. Moreover, the modification indices suggested adding regression paths from the factor Mental Efficiency to the item EFFIT3, and from the item EFFEC2 to the item EFFIT3. As such unidirectional relations were not in line with the theoretical assumptions, the item EFFIT3 was excluded.[8] After the exclusion of EFFIT3 the model fit was satisfactory, $\chi^2(38, N=57)=43.12$, p=.261, RMSEA = .05, CFI = .99 and the Fornell-Larcker-criterion was met (Table 4.17). The final model is presented in Figure 4.3.

[8] The wording of the item EFFIT3 is *"Die Interaktion mit dem System ist lahm - flott."* (translation: "The interaction with the system is sluggish - responsive.").

Table 4.17. Fornell-Larcker-criteria for Ease of Use

	Factor$_X$	N of Items	AVE (Factor$_X$)	Factor$_y$	R^2 (Factor$_X$, Factor$_y$)	Fornell-Larcker-criterion met? (AVE > R^2)
Model 1	Mental Efficiency	3	.73	Learnability	.53	yes
				Temporal Efficiency	.48	yes
				Effectiveness	.52	yes
	Learnability	4	.70	Mental Efficiency	.53	yes
				Temporal Efficiency	.55	yes
				Effectiveness	.56	yes
	Temporal Efficiency	3	.85	Mental Efficiency	.48	yes
				Learnability	.55	yes
				Effectiveness	.41	yes
	Effectiveness	2	.72	Mental Efficiency	.52	yes
				Learnability	.56	yes
				Temporal Efficiency	.41	yes
Model 2	Mental Efficiency	3	.73	Learnability	.53	yes
				Temporal Efficiency	.44	yes
				Effectiveness	.52	yes
	Learnability	4	.70	Mental Efficiency	.53	yes
				Temporal Efficiency	.54	yes
				Effectiveness	.56	yes
	Temporal Efficiency	2	.92	Mental Efficiency	.44	yes
				Learnability	.54	yes
				Effectiveness	.37	yes
	Effectiveness	2	.72	Mental Efficiency	.52	yes
				Learnability	.56	yes
				Temporal Efficiency	.37	yes



Notes. χ^2 (38, N=57)=43.12, p=.261, RMSEA = .05

e = error (e.g. measurement error)
↔ = correlations
← = regression path
▬ = observed variable (item)
● = unobserved, latent variable (underlying factor /assumed construct or error variable),
▬ = multiple squared correlation coefficients (R², item communalities, item reliabilities)
▬ = partial standardised regression weights (factor loadings),
▬ = factor correlation coefficients.

Fig. 4.3. Final model of Ease of Use after removal of item EFFIT3

Regarding Interaction Quality the Fornell-Larcker-criterion was met for both factors, Input Quality and Output Quality (see Table 4.18). However, as for the initial Ease of Use model, the fit was rather poor, χ^2(53, N=57)=88.65, *p*=.002, RMSEA = .11, CFI = .95. Hence, modification indices were investigated for this construct as well. The highest modification index indicated that the items IQ5 and OQ6 share substantial covariance. As the item IQ5 also was the Input Quality item showing the lowest loading, it was excluded.[9] The fit of the resulting model was still unsatisfactory, χ^2(43, N=57)=63.00, *p*=.025, RMSEA = .09, CFI = .97. Thus, modification indices were inspected a further time. The modification indices showed, that the item IQ6 shared a relatively high amount of variance with the factor Output Quality. Accordingly, the item IQ6 was excluded from the analysis[10]. After exclusion of IQ6 the model fit was sufficient, χ^2(34, N=57)=39.48, *p*=.238, RMSEA = .05, CFI = .99. The final model is given in Figure 4.4.

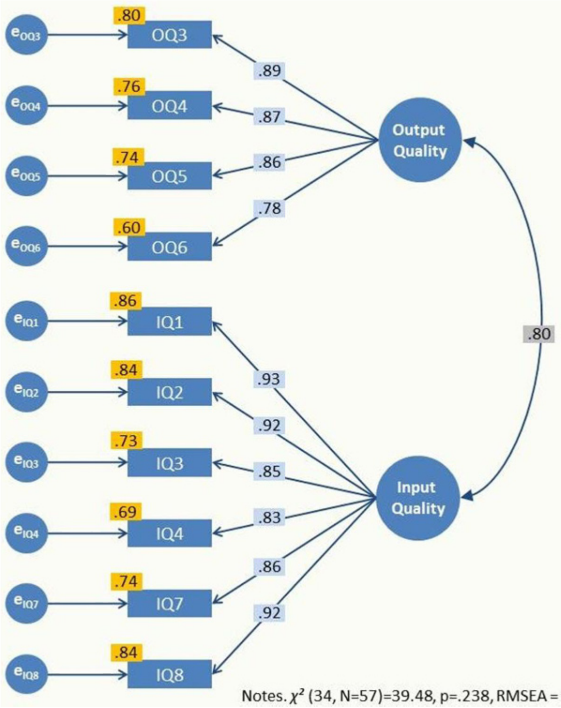
At this point 29 items were still included in the analysis. The fit criteria for all meta-constructs are given in Table 4.19.

[9] The wording of the item IQ5 is *„Die verschiedenen Eingabemöglichkeiten: sind schlecht aufeinander abgestimmt - sind gut aufeinander abgestimmt.“* (translation: „The different input modalities are poorly aligned with each other - are well aligned with each other.”)

[10] The wording of the item IQ6 is *„Die verschiedenen Eingabemöglichkeiten sind schlecht zu koordinieren - sind gut zu koordinieren.“* (translation: „The different input modalities are poorly to coordinate - well to coordinate.“)

Table 4.18. Fornell-Larcker-Criteria for Interaction Quality

	Factor$_X$	N of Items	AVE (Factor$_X$)	Factor$_y$	R² (Factor$_X$, Factor$_y$)	Fornell-Larcker-criterion met? (AVE > R²)
Model 1	Input Quality	8	.76	Output Quality	.67	yes
	Output Quality	4	.73	Input Quality		yes
Model 2	Input Quality	7	.77	Output Quality	.67	yes
	Output Quality	4	.73	Input Quality		yes
Model 3	Input Quality	6	.78	Output Quality	.64	yes
	Output Quality	4	.73	Input Quality		yes



e = error (e.g. measurement error)

↔ = correlations

← = regression path

▬ = observed variable (item)

● = unobserved, latent variable (underlying factor /assumed construct or error variable),

▬ = multiple squared correlation coefficients (R², item communalities, item reliabilities)

▬ = partial standardised regression weights (factor loadings),

▬ = factor correlation coefficients.

Fig. 4.4. Final model of Interaction Quality after removal of items IQ5 and IQ6

Table 4.19. Fit criteria for final models of meta-constructs

	$\frac{\chi^2}{df}$	CFI	RSMEA	GFI	AGFI	IR Min./Max	CR Min./Mx	N of Items
Joy of Use	1.12	.99	.05	.93	.85	.58/.95	.87/.93	8
Ease of Use	1.14	.99	.05	.88	.80	.48/.95	.84/.95	11
Interaction Quality	1.16	.99	.05	.89	.82	.60/.86	.91/.96	10

4.3.3 Validation of the Global Model – the Final Questionnaire

Next, the global model, the whole questionnaire, was investigated using exploratory factor analysis. Again, maximum-likelihood factor analysis was intended to be employed. But the solution did not converge because a Heywood case occurred, probably due to the low sample size. Heywood cases describe impossible parameter estimates, such as correlations larger than 1 (Kolenikov & Bollen, 2012). The ML method is particularly prone to the occurrence of Heywood cases; therefore another method, a Principal Axis Factor analysis (PAF), was carried out. Like the ML method, PAF aims to explain relationship between the items while the aim of the popular Principal Component Analyis (PCA) is data reduction (Bühner, 2011).

One Mental Efficiency item and one Personality item were loading on the "wrong" factor and were thus excluded.[11] The resulting pattern matrix is presented in Table 4.20. Most of the loadings are in line with the expected structure. However, one Aesthetic item and one Mental Efficiency showed high cross-loadings on the factor Discoverability. Moreover, two Learnability items were cross-loading on Aesthetics and Temporal Efficiency respectively. Nevertheless, the theoretical structure was largely in line with the empirical data.

[11] The excluded items were INTUI2 (*"Die Interaktion mit dem System ist kompliziert - unkompliziert."*; translation: "The interaction with the system is complicated - uncomplicated.") and PER3 (*"Die Interaktion mit dem System ist unangenehm - angenehm."*; translation: "The interaction with the system is unpleasant - pleasant.")

Table 4.20. Pattern matrix for whole questionnaire (final solution, N=57, only loadings > 0.3)

	Stimulation	Input Quality	Output Quality	Temporal Efficiency	Learnability	Effectiveness	Aesthetics	Personality	Mental Efficiency
PER1	-	-	-	-	-	-	-	.64	-
PER2	-	-	-	-	-	-	-	.50	-
DISC3	.71	-	-	-	-	-	-	-	-
DISC5	.94	-	-	-	-	-	-	-	-
DISC6	.93	-	-	-	-	-	-	-	-
AEST1	.38	-	-	-	-	-	.77	-	-
AEST2	.66	-	-	-	-	-	.33	-	-
EFFIT2	-	-	-	.92	-	-	-	-	-
EFFIT4	-	-	-	1.00	-	-	-	-	-
EFFIM1	.31	-	-	-	-	-	-	-	.44
EFFIM2	-	-	-	-	-	-	-	-	.54
INTUI3	-	-	-	-	-	-	-	-	-
INTUI4	-	-	-	-	1.05	-	-	-	-
LEARN1	-	-	-	.31	.56	-	-	-	-
LEARN2	-	-	-	-	.47	-	.36	-	-
EFFEC1	-	-	-	-	-	.73	-	-	-
EFFEC2	-	-	-	-	-	.77	-	-	-
OQ3	-	-	.81	-	-	-	-	-	-
OQ4	-	-	.86	-	-	-	-	-	-
OQ5	-	-	.73	-	-	-	-	-	-
OQ6	-	-	.76	-	-	-	-	-	-
IQ1	-	.79	-	-	-	-	-	-	-
IQ2	-	.97	-	-	-	-	-	-	-
IQ3	-	.66	-	-	-	-	-	-	-
IQ4	-	.85	-	-	-	-	-	-	-
IQ7	-	.84	-	-	-	-	-	-	-
IQ8	-	.82	-	-	-	-	-	-	-

The remaining 27 items were modelled with AMOS and evaluated using the criteria explained earlier. The initial model showed insufficient fit, χ^2(312, N=57)=466.23, p<.001, RMSEA = .09, CFI = .90. High modification indices were found for one of the Learnability items (LEARN2): It was highly correlated with one of the Aesthetic items; furthermore the modification indices suggested adding regression paths from several other items to the item LEARN2. Accordingly the item was excluded.[12] The resulting model showed a better but still unsatisfactory fit, χ^2(287, N=57)=406.04, p<.001, RMSEA = .09, CFI = .92. Thus modification indices were

[12] The wording of the item LEARN2 is *"Die Gestaltung des Systems ist ungeeignet - geeignet."* (translation: "The design of the system is unsuitable - suitable".).

inspected again. For the Output Quality item OQ3 modification indices were highest. It was highly correlated with the factor Joy of Use and, as for the previously excluded item LEARN2, also several regression paths were indicated. Consequently also the item OQ3 was dropped.[13] Except for the slightly to low CFI, the fit criteria were then satisfactory (cf. Table 4.21).

Table 4.21. Fit criteria for global model (Intermediate solution, N=57)

$\frac{\chi^2}{df}$	CFI	RSMEA	GFI	AGFI	IR Min./Max	CR Min./Max.	N of Items
1.35	.93	.08	.64	.58	.53/.94	.85/.93	25

For the remaining 25 items, the Fornell-Larcker criterion was investigated. It was not fulfilled for Ease of Use (see Table 4.22). This indicates that the concepts measured with this factor are not theoretically unrelated to the concepts measured by the other factors.

Table 4.22. Fornell-Larcker criterion for global model (Intermediate solution, N=57)

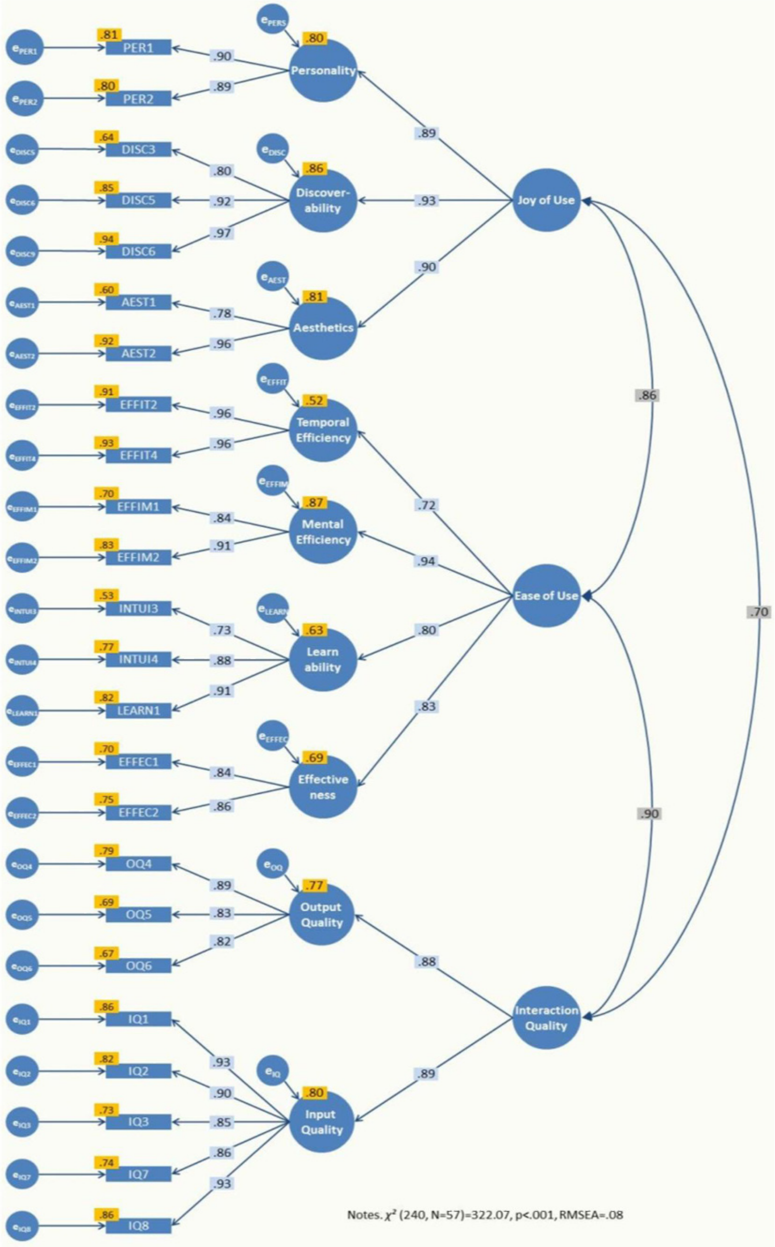
Factor$_X$	N of Items	AVE (Factor$_X$)	Factor$_y$	R^2 (Factor$_X$, Factor$_y$)	Fornell-Larcker criterion met? (AVE > R^2)
Joy of Use	7	.82	Interaction Quality	.48	yes
Ease of Use	9	.68	Joy of Use	.74	no
Interaction Quality	9	.78	Ease of Use	.81	no

However as already explained in Chapter 4.3.2, unrelated factors were not expected. Moreover, according to Backhaus et al. (2006) sufficient discriminate validity can be assumed, despite a non-fulfilled Fornell-Larcker criterion, as long as the correlation between the factors is not higher as 0.9. Between Interaction Quality and Ease of Use, the correlation coefficient was slightly higher (*r*=.902). Hence, all remaining items for Interaction Quality and Ease of Use were correlated with each other in order to remove items, showing low correlations with the items on the same factor and high correlations with the items on the other factor. As for Input Quality a high number of items was still in the item pool, compared to the number of items of the other factors, the analysis was started with this factor. The item IQ4 showed the lowest difference between the mean correlation with the related items (Interaction Quality items) and the mean correlation with the unrelated items (Ease of Use items).

[13] The wording of the item OQ3 is *„Die Rückmeldungen des Systems sind hemmend - unterstützend.“* (translation: „The system's feedback is hindering - supporting.")

Consequently, the item was excluded.[14] After exclusion of IQ4, the correlation was reduced to .898 and thus below 0.9. Fit criteria were inspected again and indicated sufficient fit (Table 4.23). The final model is given in Figure 4.5.

[14] The wording of the item IQ4 is *„Die verschiedenen Eingabemöglichkeiten sind schlecht miteinander integriert - sind gut miteinander integriert.“* (translation: „The different input modalities are poorly integrated with each other - are well integrated with each other.”)



Notes. χ^2 (240, N=57)=322.07, p<.001, RMSEA=.08

e = error (e.g. measurement error) ↔ = correlations

← = regression path ▬ = observed variable (item)

● = unobserved, latent variable (underlying factor /assumed construct or error variable)

🟨 = multiple squared correlation coefficients (R², item communalities, item reliabilities)

🟦 = partial standardised regression weights (factor loadings)

⬜ = factor correlation coefficients.

Fig. 4.5. Final global model after removal of items LEARN2, item OQ3 and item IQ4.

Table 4.23. Fit criteria for global model (final solution, N=57)

$\frac{\chi^2}{df}$	CFI	RSMEA	GFI	AGFI	IR Min./Max	CR Min./Max.	N of Items	AVE Min./Max.	Cronbach`s α Min./Max.
1.34	.94	.08	.72	.65	.52/.94	.85/.93	24	.68/.82	.92/.95

Next, the model was validated with the remaining cases of the initial sample. Aim was to ensure that the data structure is not sample dependent and is also valid for other samples. Homburg and Giering (1996) recommend collecting a new, larger sample. As only a small part (N=57) of the initial sample was used to develop the above model, the larger part was available for the following validation. Moreover, the samples include different studies and population; accordingly, the resulting models should not be population-dependent. Thus, a new data collection was not necessary.

The model based on the held out sample, largely matched the results of the smaller sample: The CFI was slightly too low and the factor Ease of Use showed high correlations with the other factors. The latter was expected as all factors were assumed to be related. Fit indices are presented in Table 4.24, the validated model is shown in Appendix B.2, the wording of the items of the final questionnaire can be found in Table B.1.

Table 4.24. Fit criteria for validated model (final solution, N=188)

$\frac{\chi^2}{df}$	CFI	RSMEA	GFI	AGFI	IR Min./Max	CR Min./Max.	N of Items	AVE Min./Max.	Cronbach's α
1.88	.94	.70	.84	.80	.53/.91	.74/.91	24	.67/.76	.92/.92

In the last step, the questionnaire was validated with the AttrakDiff questionnaire (Hassenzahl, Burmester, & Koller, 2003). Both questionnaires were used in a study evaluating a multimodal remote control app (Weiss, Wechsung & Marquardt, submitted). The questionnaires' sub-scales (constructs or dimensions) were correlated with each other (cf. Table 4.25). The results confirm the MMQQ`s convergent validity. As expected, the scale Joy of Use was highly correlated with the hedonic qualities scales of the AttrakDiff and the global Attractiveness scale; the correlation between Joy of Use and Pragmatic Qualities was not significant. Also in line with the theoretical assumptions are the significant correlations between Ease of Use and Pragmatic Qualities, and between Ease of Use and Attractiveness. Also, a significant correlation for Hedonic Qualities-Identity was observed, but this was also the case for the AttrakDiff scale Pragmatic Qualities. Hence, apparently for the tested app those aspects were related. Interaction Quality was significantly correlated with Pragmatic Qualities only, which is consistent with expectation, given the definitions of Input Quality and Output Quality presented in Chapter 3.3.3. Concepts like perceived understanding or understandability are related to a system`s functionality more than to its hedonic, not task-related properties. As also the global models of the MMQQ indicated a stronger relation between Ease of Use and Interaction Quality than between Joy of Use and Interaction Quality, the position of Interaction Quality in the underlying taxonomy was changed: While in the original taxonomy (Möller, 2009) it is positioned within both, the pragmatic and the hedonic dimension. In the taxonomy presented in Chapter 3 Interaction Quality is located within the pragmatic dimension,

Table 4.25. Correlations (Pearson's *r*) between MMQQ sub-scales and AttrakDiff sub-scales.

		Ease of Use	Interaction Quality	Pragmatic Qualities	Hedonic Qualities Stimulation	Hedonic Qualities Identity	Attractiveness
Joy of Use	Pearson's *R*	.49*	.54*	.43	.80**	.78**	.78**
	N	17	16	17	17	17	17
Ease of Use	Pearson's *R*	-	.52*	.77**	.44	.63**	.79**
	N	-	16	17	17	17	17
Interaction Quality	Pearson's *R*	-	-	.64**	.40	.39	.49
	N	-	-	16	16	16	16
Pragmatic Qualities	Pearson's *R*	-	-	-	.42	.65**	.73**
	N	-	-	-	17	17	17
Hedonic Qualities Stimulation	Pearson's *R*	-	-	-	-	.85**	.79**
	N	-	-	-	-	17	17
Hedonic Qualities Identity	Pearson's *R*	-	-	-	-	-	.89**
	N	-	-	-	-	-	17

Note. ** $p_{\text{two-tailed}}$<.01; *$p_{\text{two--tailed}}$ <.05,

4.4 Chapter Summary

In this chapter a new questionnaire for the evaluation of multimodal systems was developed. At first it was shown that standardized usability questionnaires designed for unimodal systems are not appropriate for post-interaction overall evaluation of multimodal systems and that of the tested questionnaires, only the AttrakDiff (Hassenzahl, Burmester, & Koller, 2003) yielded valid and reliable results. However, the AttrakDiff provides rather unspecific information, which may make it difficult to track and fix specific problems based on its results. Hence, this questionnaire was used as a starting point for the development of a new questionnaire tailored to multimodal systems; a major requirement for the new questionnaire was to provide more specific information than the AttrakDiff. As the theoretical basis for the questionnaire, the taxonomy of quality aspects of multimodal systems by Möller and colleagues (2009) was chosen. In parallel to the questionnaire development process, the quality aspect layer of the taxonomy was validated and altered according to the empirical data. The validation was achieved with the employment of confirmatory modelling approaches in addition to the usual exploratory approaches. As a result of the analyses, the construct Intuitivity was dropped; moreover the position of the concept Interaction Quality was altered. While in the original taxonomy Interaction Quality was located in between both, the pragmatic and the hedonic dimensions, results of the current chapter suggest that its relation to the pragmatic dimension is stronger compared to its relation to the hedonic dimension. The taxonomy was adjusted accordingly. Thus, the taxonomy presented in Chapter 3 is an empirically validated version of the taxonomy published in Möller et al. (2009).

The MMQQ questionnaire developed in the current chapter covers all quality aspects except Cooperativity. It contains 24 items, nine quality aspects and three metascales. The MMQQ was extensively validated; however, its correlations with interaction data have not been investigated yet. Correlations with the AttrakDiff are promising that the constructs are valid.

Even with this questionnaire being available, the assessment of quality of the complete multimodal system is still only possible near the end of the development cycle, when all components are implemented. Furthermore, such global evaluations of multimodal systems tell little about how the different modalities relate to each other. For example, it may be the case that if one modality is implemented extremely poorly, quality perceptions of the whole systems are dominated by this modality regardless of the other modality. Also the opposite is imaginable, an extremely well working, fun to use modality may overshadow less proper modalities.

Thus, the next chapter investigates how ratings for single modalities relate to the global evaluation of multimodal systems. Aim was to see if an estimation of the quality of the multimodal systems is possible, based solely on an evaluation of its component modalities.

5 Is the Whole the Sum of its Parts? - Predicting the Quality of Multimodal Systems Based on Judgments of Single Modalities

In this chapter, three studies are described aiming to investigate whether ratings of the individual components of multimodal systems are suitable to estimate the quality of the whole system. With respect to early development stages such an approach could be especially beneficial in getting a rough approximation of the quality of a system, without the need of deploying complex evaluation procedures, which currently are often required for testing multimodal systems.

In the first study, a wall-mounted information and room management system offering speech and touch input was used. The different modalities could interfere with each other for this system, for instance the speech recognizer was occasionally unintentionally switched on by off-talk although the user intended to operate the system via touch input. The multimodal condition was always presented last. The second study used an extended version of the system already used in the initial study but the multimodal condition was always presented first. Aim was to ensure that the results of Study 1 are not an artefact of the test design. In order to investigate the generalizability of the results a different system, a mobile jukebox, was used in the third study. Besides speech and touch also motion control was offered. Moreover, a different user group, elderly users, was taken into account.

The work presented here is grounded in some previous research. In terms of the statistical method, all studies follow the PARADISE approach (Walker et al. 1997), and thus linear regression is applied. But while PARADISE predicts user satisfaction ratings based on interaction parameters quantifying dialog costs and task success, the studies described in this chapter make predictions regarding the perceived quality of multimodal systems based on the quality ratings for the individual components. PARADISE, initially developed for spoken dialogue systems, achieves predictions explaining up to 50 % of the variance in the user satisfaction ratings for unseen data (Engelbrecht, 2012). Note that PARADISE models with high prediction performance usually contain, apart from the interaction parameters, quality ratings for success (Kühnel, 2012).

Regarding multimodal systems, a similar approach was defined with PROMISE (Beringer et al., 2002) but to the best of the authors knowledge has never been applied. However, the PARADISE framework itself has been adopted (e.g. Hjalmarsson, 2002) and extended for multimodal systems (Kühnel 2012). To assess user satisfaction, Kühnel does not use the questionnaire proposed in PARADISE, but the AttrakDiffMini (Hassenzahl & Monk, 2010), a short version of the AttrakDiff questionnaire already explained in Section 4.1.1. The PARADISE questionnaire was discarded due to its lack of psychometric validation, a criticism that has also been point-

DOI: 10.1007/978-3-319-03810-0_5, © Springer International Publishing Switzerland 2014

ed out by Larsen (2003a) and Hajdinjak and Mihelic (2006). Moreover, based on the work of Möller (2005), Kühnel (2012) suggests a new set of parameters for modelling the perceived quality of unimodal as well as multimodal interaction. The explained variance achieved with Kühnel's approach was as high as 64 %, using only interaction parameters. However, Kühnel did not cross-validate the models. In addition, Kühnel (2012) found that performance varied largely for different systems and input modalities, and that a generalized model suitable for different systems was not possible. Thus, although Kühnel's results are very encouraging, they also indicate that accurate modelling of user satisfaction of multimodal interaction cannot be easily realized. However, the approach proposed in the following may be useful for the judgement prediction from interaction parameters as well, since understanding of the judgemental process is gained.

5.1 Study 5.1

5.1.1 Method

Participants and Material

Thirty-six German-speaking individuals (17 male, 19 female) between the age of 21 and 39 (M = 31.24) took part in the study.

The system tested was a wall-mounted information and room management system controllable via a graphical user interface (GUI) with touch input, via speech input and via a combination of both (cf. Figure 5.1). The output was always given via GUI. Functionalities of the system include searching for room locations, searching for employees, booking of rooms and retrieving information on bookings and events. For all functionalities tested in the study both modalities, touch and speech, were implemented.

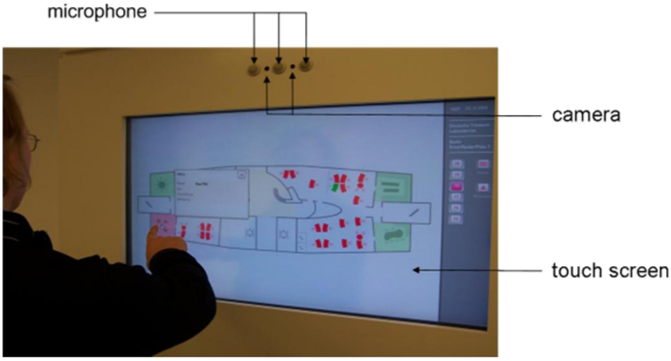


Fig. 5.1. Information and room management system used in Study 5.1 and Study 5.2.

Procedure

The users performed six different tasks with the system, including two search tasks (e.g.:‘Search for employee X’ or ‘Search for room Y’) and four navigation tasks (e.g.: ‘Go to the main screen’ or “Go to the event screen’). User ratings were collected with the AttrakDiff questionnaire (Hassenzahl, Burmester, & Koller, 2003).

Each test session took approximately one hour. Each participant performed the tasks with each input modality (touch and speech). First, they were instructed to perform all tasks with a given modality. After that, they were asked to fill out the AttrakDiff. This was repeated for the other modality. In order to balance fatigue and learning effects, the order of the modalities was randomized. Then, the tasks were presented again and the participants could freely choose the interaction modality. It was possible to switch or combine modalities after each task and also within a task. Again the AttrakDiff had to be filled out to rate the multimodal system.

The four AttrakDiff sub-scales (Pragmatic Qualities, Hedonic Qualities-Stimulation, Hedonic Qualities-Identity, Attractiveness), comprising seven items each were calculated according to the instructions provided by its authors. Furthermore, an overall scale was calculated based on the mean of all 28 items. All questionnaire items which were negatively poled were recoded so that higher values indicate better ratings.

To analyse which modality the participants preferred when using the multimodal system version, the modality chosen first to perform the task was annotated. This way, the frequencies of modality usage were assessed. An example of the procedure is presented in Figure 5.2.

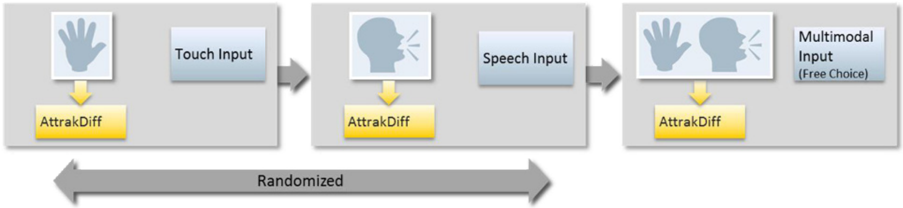


Fig. 5.2. Procedure for Study 5.1

5.1.2 Results

Differences between the three system versions were not in the focus of the reported study, for completeness those results are given in Appendix A.3.

To investigate if and how the ratings of the unimodal system versions relate to ratings for the multimodal system version, stepwise multiple linear regression analysis was conducted for each sub-scale and the overall scale. The judgments assessed after the interaction with the unimodal system versions were used as predictor variables, the judgments collected after interacting with the multimodal system version were used as the response variable.

The results show that for the Attractiveness scale and the overall scale the judgments of the unimodal system are very good predictors of the judgments of the multimodal version (Table 5.1). For both scales, the β–coefficients (also called standardized coefficient) were higher for the judgments of the touch-controlled version of the system. Higher β-coefficients indicate a stronger influence. This is in line with the modality usage for the multimodal system: Touch input was used more frequently for 50% of the tasks, speech was preferred for 33% of the tasks. For 17% of the tasks, touch and speech were used equally often. Thus, the overall and global judgments of the multimodal system should be more influenced by the interaction with the touch input.

Regarding the scales Hedonic Qualities-Stimulation, Hedonic Qualities-Identity and the Pragmatic Qualities, between 61% and 69% of the variance could be explained by using the ratings of the unimodal systems as predictors of the ratings for the multimodal system. For both hedonic scales, the β-coefficients of speech were higher than those of touch; therefore, the ratings of speech had a larger impact on the multimodal system judgments than the touch ratings had.

Table 5.1. Results of multiple linear regression analysis using all data. Only significant parameters (p<.05) are included in prediction.

Scale	Touch				Speech				*Adj.* *R^2*	*RMSE*	*F* *(df)*
	B	*SE B*	*β*	*t* *(df)*	*B*	*SE B*	*β*	*t* *(df)*			
Overall	.81	.11	.57	0.21* (32)	.68	.10	.55	6.91* (32)	.82	.37	74.94* (2,31)
Attractiveness	.85	.09	.68	9.02* (32)	.48	.09	.42	5.52* (32)	.83	.41	81.99* (2,32)
Pragmatic Qualities	.80	.17	.54	4.57* (31)	.47	.13	.42	3.59* (31)	.60	.70	26.19* (2,31)
Hedonic Qualities-Stimulation	.69	.13	.52	5.13* (32)	.63	.12	.54	5.31* (32)	.67	.53	36.05* (2,32)
Hedonic Qualities-Identity	.28	.11	.31	2.66* (32)	.66	.14	.57	4.60* (32)	.59	.51	25.24* (2,32)

Leave-one-out cross-validation[15] was conducted to test for overfitting effects. For the Attractiveness scale and the overall scale the determination coefficient R^2 is still

[15] For this procedure one case is excluded from the dataset and is used as the "test set" while the model is trained on the remaining cases. This is repeated for every case. Based on the prediction error of each "left-out" model the mean prediction error can be calculated (Steiner & Weber, 2009).

around 0.8 indicating that about 80% of the variance of the multimodal ratings can be explained with the unimodal ratings. For the other scales the overfitting effects are larger, with the worst accuracy for Hedonic Qualities-Identity. Note that in the model for Hedonic Qualities-Identity the β-coefficient of the touch ratings is relatively low compared to the β-coefficient of the speech ratings; it is also relatively low compared to the β-coefficient of the touch ratings in the models for the other scales although touch was used more often than speech. For the scales with the most stable models, Attractiveness and the overall scale, on the other hand, the β-coefficients of the ratings for touch and speech were in largely in line with the usage rates. The detailed results are given in Table 5.2.

Table 5.2. Results of multiple linear regression analysis using leave-one-out cross-validation.

	Overall	Attractiveness	Pragmatic Qualities	Hedonic Qualities-Stimulation	Hedonic Qualities-Identity
R^2	.89	.81	.56	.61	.47
RMSEA	.39	.43	.74	.58	.58

5.1.3 Intermediate Discussion

The reported study investigated how judgments of unimodal system versions relate to judgments of the multimodal version of the same system. It was shown that for overall and global measures (Attractiveness) the judgments of the unimodal versions are very good predictors for judgments of the multimodal version. For more specific quality aspects, the prediction performance is lower. Additionally, the results indicate that for the stable models, the modality used more frequently has a higher influence on the judgment of the multimodal version than the less frequently used modality.

Furthermore, in accordance with Oviatt (1999) it could be observed that adding a modality to a unimodal system does not automatically lead to better quality judgments. For the present study this means, that regarding overall and global judgments the whole is actually the sum of its parts. Ratings for the multimodal system are the (weighted) sum of the ratings of the unimodal systems. However, for scales measuring more specific constructs, this assumption is not valid; here, the accuracy was lower compared to the models for the more general scales. This may be related to the observation already mentioned above: for the more specific scales the influence of the judgements of the individual modalities is not in line with the actual usage rates.

The reported findings are currently limited to the tested system and the test design. For the multimodal system version, interference between the modalities was possible (e.g. the speech recognizer was occasionally unintentionally switched on by off-talk). It is therefore plausible, that without this interference different results may have been obtained. Moreover, the multimodal version was always tested last. Hence, it is possible that the participants tried to rate consistently by mentally adding up their single-modality judgments. Consequently, the judgments of the multimodal version would not represent the actual quality of that system. Facing these limitations, two follow-up studies were conducted using a different test design (Study 5.2) and a different system (Study 5.3).

5.2 Study 5.2

The aim of this experiment was to examine if the results obtained in the previous study were a consequence of the test design and the participants effort to rate consistently. Therefore, the order of the system versions was changed with the multimodal system presented first. Thus, mentally adding up the single-modality judgments to the multimodal judgments was not possible. The participants could only pre-estimate the quality of the unimodal system versions and their impact on the multimodal systems quality. Furthermore, addition is a less effortful mental operation compared to subtraction (Kamii, Lewis, & Kirkland, 2001; Dixon, Deets, & Bangert, 2001). This means that single modalities ratings are more difficult to derive from multimodal ratings. Consequently, less accurate predictions are expected.

5.2.1 Method

Participants and Material

Eighteen German-speaking individuals (9 male, 9 female) between the age of 22 and 30 (M = 26.7) took part in the study. The tested system was similar to system in Study 5.1. However, it was extended with face recognition to activate the speech recognizer.

All measurements were taken like in Study 5.1.

Procedure

The procedure was very similar to the procedure in Study 5.2. Only the order of the conditions (touch input, speech input, multimodal input/free choice) was reversed. This time, the multimodal block, where participants could freely choose and switch the input modality, was presented first. In the following blocks, they were instructed to use a given modality.

5.2.2 Results

Differences between the three system versions were not in the focus of the reported study, those results are published in Seebode et al. (2009).

Stepwise multiple linear regression analysis was conducted for each sub-scale and the overall scale of the AtrrakDiff, using all cases. The questionnaire ratings obtained after the multimodal test block were used as response variable, the ratings obtained after the unimodal test blocks were used as predictors.

Prediction was not possible for two sub-scales, as no significant predictor could be found by the stepwise inclusion algorithm (Hedonic Qualities-Identity, Pragmatic Qualities). The highest accuracy was observed for the scale Hedonic Qualities-Stimulation (cf. Table 5.3).

For two of the three scales for which prediction was possible, *β*-coefficients were not differing much between the modalities. This is in line with the modality usage: Both modalities were used equally frequently (Touch: 51.7%; Speech: 48.3%). Thus, it is plausible to assume, that both modalities had a similar impact on the judgments of the multimodal system.

Table 5.3. Results of multiple linear regression analyses using all cases. Only significant parameters (p<.05) are included in prediction.

Scale	Touch				Speech				Adj. R^2	*RMSE*	*F* (2, 15)
	B	*SE B*	*β*	*t* (15)	*B*	*SE B*	*β*	*t* (15)			
Overall	.68	.22	.53	3.11	.55	.19	.50	2.93	.58	.45	12.56
Attractiveness	.65	.24	.46	2.78	.55	.16	.55	3.32	.54	.73	11.05
Hedonic Qualities - Stimulation	.66	.12	.73	5.64	.49	.15	.43	3.35	.72	.41	22.97

Thus to test for over-fitting effects, leave-one-out cross-validation was conducted. Only for Hedonic Qualities-Stimulation the model was stable (R^2=.53, *RMSE*=.54). Very large over-fitting effects were observed for Attractiveness (R^2=.06, *RMSE*=.88) and for the overall scale (R^2=.133, *RMSE*=.52), but for the latter the prediction error was rather low. A visual inspection of the scatter plot showed (cf. Fig. 5.3), that the variance in the ratings for the overall scale was relatively low, it is not clear if the R^2 or the *RMSE* is the more reliable performance indicator in this case.

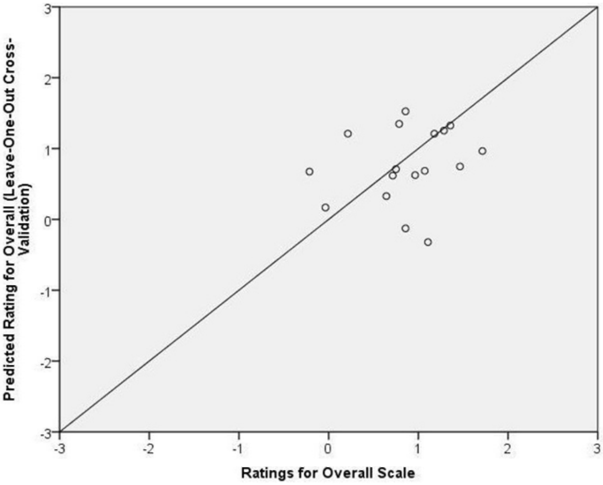


Fig. 5.3. Scatter plot for overall scale after leave-one-out cross-validation.

5.2.3 Intermediate Discussion

Although the results indicate that the order of test conditions influences the accuracy and stability of the models, the same questionnaire scales as in Study 5.1 were suitable for prediction. Thus, predicting judgments for multimodal systems based on judgements of single modalities might primarily be possible for global judgements and for the sub-scale measuring stimulation. Additionally only data of 18 users was analysed which besides the order effect may be an explanation for the poorer prediction performance.

5.3 Study 5.3

This experiment was conducted to examine if the results in Study 5.1 and Study 5.2 are valid when using a system where there is no interaction (and thus no possible interference) between the different modalities. In such cases, participants can use the multimodal system like a unimodal one, by using only the modality they like most. For such systems, the modality not or rarely used should have a lower impact on the prediction of the multimodal ratings.

Furthermore, Study 5.2 showed that changing the order is of negative effect for the predictions: Accuracy and stability were lower than in Study 5.1. If users are just summing up the unimodal judgments this should be influenced by memory capacity. Memory performance, especially processes related to the working memory decreases with age (Spencer & Raz, 1995). Therefore, young and old users were compared, better models were expected for the younger participants.

5.3.1 Method

Participants and Material

Fifteen younger (<25 years/ *M* =29 y.) and fifteen older (>55 years/ *M* = 66 y.) users took part in the study.

The application tested, the *Sprachbox* app, was a multimodal mailbox system capable of handling speech-, e-mail- and fax-messages as well as call forwarding and notifications of mailbox messages. It was implemented on a smart-phone (HTC Touch Diamond) controllable via motion (tilt and twist), speech (IBM embedded Via Voice) and touch screen. For speech and motion input activation, a respective button had to be pressed before starting to speak or tilt and twist. The push-to-talk trigger is placed on the left hand side and the push-to-move button on the front of the device (cf. Figure 5.4).

System output was graphical for all modalities. For motion control, additional tactile feedback (vibration) and for speech control, additional auditive feedback - either beeping sounds characterizing the events ‘active’, ‘match’, ‘no match’, and speech output - was given.

The main screen of the *Sprachbox* had a simple structure. It consisted of four menu options: voice messages, e-mail inbox, fax messages, and an option for settings. For the first three options, there was a list of the messages or e-mails with the information about the sender, date, and time. The messages could be opened in order to be read or to be listened to. There were also options for answering to e-mails or forwarding fax messages. Messages or e-mails could also be deleted. The inbox or message list could be sorted with respect to a specific criterion (e.g. alphabetically or chronologically). With the ‘settings’ option, for example, a call forwarding to another telephone number could be set up, activated, and deactivated. The default setting of the application for the study contained fictitious voice messages, e-mails, and fax messages.

The interaction with the system mainly comprised navigation between menu options and messages and the selection of messages and of actions to be performed on the messages. In the modality touch this was achieved by swiping gestures for scrolling and tapping gestures for selecting. The modality speech allowed users to directly select messages and commands by saying short commands like “Show me the message of Tom”. The modality motion allowed users to navigate within lists by slightly tilting the device forward and backward and to switch to different levels of the menu hierarchy by tilting the device left and right. Menu items were opened by tilting to the right. Hierarchy levels were switched by tilting to the left.

Again, the AttrakDiff was used to assess quality perceptions. Additional information on this study is provided in Appendix A.2.

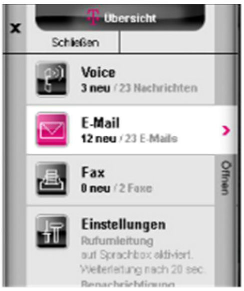


Fig. 5.4. Start screen of application used in Study 5.3.

Procedure

The procedure was adopted from Study 5.1. Participants had to execute 14 tasks (get messages, reply to them, forward, and sort messages as well as changing notification options) with each input modality (touch, speech, motion). After the single modality conditions, the tasks had to be executed again. In this condition (multimodal/free choice), participants could freely choose the input modality and switch the modality at any time. The order of single modality conditions was balanced (Latin square).

5.3.2 Results

Again, stepwise multiple linear regression analysis was carried out for each scale and sub-scale (cf. Table 5.4). Similar to the previous results, best predictions were observed for overall judgments and the scale measuring Hedonic Qualities-Stimulation. But in contrast to the previous experiments, prediction for the scale Hedonic Qualities-Identity was relatively good and prediction for the global scale Attractiveness was relatively bad compared to the other scales. Poorest results were obtained for the scale Pragmatic Qualities. Overall, the models were more accurate than in Study 5.2 but not as precise as in Study 5.1.

For none of the predictions, motion was included, and speech was only included for one scale (Hedonic Qualities - Stimulation).

Thus, for four of the five scales, only the judgments for touch were included in the predictions. Accordingly, judgments for speech and motion could not explain a significant part of the variance in the judgments for the multimodal system version. The actual modality usage is in accordance with these results: Touch was used most often (68%). All other modalities were chosen considerably less frequently (speech 19 %; motion 7%; combination 6%). This supports the assumption that for multimodal systems with no potential interference between modalities, the modality not used should have no impact on the multimodal judgments.

Table 5.4. Results of multiple linear regression analyses using all cases. Only significant parameters (p<.05) are included.

Scale	Touch				Speech				Adj. R^2	*RMSE*	*F* (*df*)
	B	*SE B*	*β*	*t* (*df*)	*B*	*SE B*	*β*	*t* (*df*)			
Overall	.69	.11	.84	7.20 (23)	-	-	-	-	.69	.48	51.85 (1,23)
Attractiveness	.79	.14	.75	5.66 (25)	-	-	-	-	.56	.68	32.10 (1,25)
Hedonic Qualities - Stimulation	.60	.11	.61	5.24 (24)	.35	.11	.39	3.37 (24)	.86	.40	68.95 (2,24)
Hedonic Qualities - Identity	.75	.20	.83	7.44 (26)	-	-	-	-	.69	.45	55.36 (1,26)
Pragmatic Qualities	.49	.13	.60	3.83 (24)	-	-	-	-	.36	.77	14.70 (2,24)

Leave-one-out cross-validation showed, as in the previous studies, the least stable model for the scale Pragmatic Qualities. Again, prediction based on the scale Hedonic Qualities- Stimulation was most stable (cf. Table 5.6).

Table 5.6. Results of leave-one-out cross-validation by age group and overall cases.

Scale	R^2			*RMSE*		
	Young	Old	All	Young	Old	All
Overall	.62	.40	.58	.23	.54	.32
Attractiveness	.52	.07	.47	.45	.96	.45
Hedonic Qualities - Stimulation	.75	.62	.85	.15	.74	.17
Hedonic Qualities - Identity	.70	.31	.65	.17	.44	.21
Pragmatic Qualities	.06	-.23^2=.05	.22	.65	1.54	.61

To test for age effects, multiple linear regressions were conducted for each age group separately. As hypothesized, predictions were less accurate for older participants (cf. Table 5.7), except for the scale Hedonic Qualities-Stimulation in terms of the R^2. However, the cross-validated models were always better for the younger users (cf. Table 5.6).

Table 5.7. Results of multiple regression analysis by age group

Scale		Touch				Speech				Motion				R^2	RMSE	F (df)
		B	*SE B*	*β*	*t (df)*	*B*	*SE B*	*β*	*t (df)*	*B*	*SE B*	*β*	t (df)			
Overall	Young	.83	.14	.87	5.99 (12)	-	-	-	-	-	-	-	-	.75	.41	35.85 (1,12)
	Old	.80	.21	.80	3.78 (8)	-	-	-	-	-	-	-	-	.64	.57	14,26 (1,8)
Attractive-ness	Young	.90	.18	.81	4.91 (13)	-	-	-	-	-	-	-	-	.65	.61	24.14 (1,13)
	Old	.85	.26	.72	3.25 (11)	-	-	-	-	-	-	-	-	.51	.74	1.58 (1,11)
Hedonic Qualities - Stimulation	Young	.40	.10	.44	3.87 (12)	.45	.08	.63	5.49 (12)	-	-	-	-	.88	.28	45.93 (2,12)
	Old	1.06	.09	1.07	11.94 (7)	-	-	-		-.36	.14	-.24	-2.61 (7)	.96	.29	78.52 (2,7)
Hedonic Qualities - Identity	Young	.85	.13	.88	6.51 (13)	-	-	-	-	-	-	-	-	.77	.39	42.34 (1,13)
	Old	.71	.18	.78	3.91 (10)	-	-	-	-	-	-	-	-	.60	.53	15.25 (1,10)
Pragmatic Qualities	Young	.54	.18	.65	2.96 (12)	-	-	-	-	-	-	-	-	.42	.75	8.74 (1,12)
	Old	.50	.21	.59	2.39 (11)	-	-	-	-	-	-	-	-	.34	.82	5.73 (1,11)

5.3.3 Intermediate Discussion

Results showed that the age has an influence on the prediction accuracy of the proposed modelling. To which extent these results are due to the lower memory capacity of the older participants cannot be determined as the actual memory capacity has not been measured. However, it may have been the case that the older users had more difficulties in memorizing their previous judgments and/or in remembering and judging the interaction correctly, which would consequently decrease the prediction accuracy.

Furthermore, the hypothesis that for systems with no interaction between the modalities the judgments of the most used modality match to a large extent the judgment for the overall system, is supported also by this study.

5.4 Chapter Discussion and Chapter Summary

In this chapter, three studies were reported aiming to predict quality ratings of a multimodal system based on the ratings of its individual component modalities. In all three studies, prediction was best for overall and general judgments and judgments on the scale Hedonic Quality-Stimulation.

The poor prediction performance for the scale Hedonic Quality-Identity in Study 5.1 and Study 5.2 might be explained by the underlying construct measured via this scale. As mentioned before the scale Hedonic Qualities-Identity measures a product's ability to express the owner's self. The room information and management system, which was tested in Study 5.1 and Study 5.2 was not developed (and is not available) for personal use. Moreover, it was custom tailored for *Deutsche Telekom Innovation Laboratories* (T-Labs). Thus, this system was in none of its versions designed to promote self-expression or identification by communicating personal values. Furthermore the test was very task-oriented and goal-oriented. Data confirmed that users rated neutral on this scale. The mean differed between 0.3 and 1.2 on a scale from -3 to 3.

However, no such explanation can be found regarding the scale Pragmatic Qualities, measuring largely the "classical" concept of usability. This is in line with the results reported in Kühnel (2012), who also reports relatively poor prediction performance for this scale in the context of multimodal interaction. Kühnel (2012) hypothesizes that the Pragmatic Qualities sub-scale might not be adequate for multimodal interfaces, as she observed prediction to be rather accurate for graphical user interfaces. However, the results reported in Study 4.1 indicate that at least regarding the (temporal) efficiency, the Pragmatic Qualities scale of the AttrakDiff provides reliable results. Though, Kühnel used the short AttrakDiffMini (Hassenzahl & Monk, 2010), while in the study reported in Chapter 4.1 the long version was used. Still, also the studies in the current chapter employed the longer version, nevertheless leading to relatively poor results. Therefore, it remains unsettled why accurate predictions of perceived Pragmatic Qualities are not possible for multimodal systems. The results of the above studies imply that the perceived Pragmatic Qualities of a multimodal system is not just a simple linear combination of the perceived quality of its individual components.

The hypothesis that prediction performance is partially a consequence of the participants' effort to rate consistently was supported. Both the order of the system version and the age of the participants (as an indicator of memory capacity) seemed to be of influence. The effect of order implicates that repeated measure test designs (within designs) have to be considered as critical for evaluation studies.

Generally, it has to be remarked that for all experiments the sample size was very small. More accurate results may have been obtained with a larger sample. Furthermore, all results are based on only one questionnaire. As shown for the scale Hedonic Quality-Identity, results are heavily dependent on the appropriateness of the construct intended to be measured.

However, the results for the overall scales indicate that a rough estimation of a multimodal system's quality is possible, based on the quality of its individual components.

All of the studies imply that the modality usage is crucial. Ratings of the modalities used more often have a larger impact on the ratings for the multimodal system. Hence, if also modality choice could be predicted, models should be more accurate. This issue, factors influencing modality choice, will be addressed in the following chapter.

6 What Determines Modality Selection Strategies? - Identifying Factors Influencing Modality Selection and Perceived Quality

The following chapter is divided into three sections, each of which focuses on a specific factor assumed to influence modality selection strategies, as well as performance and quality perceptions. Each section reviews relevant previous work in the field that provides a theoretical basis for the empirical studies which follow.

Mobile devices are nowadays equipped with computational power and functionalities comparable to earlier desktop PCs, their immobile counterparts, but offer less visual bandwidth due to the small screen size. This trend, together with the major improvements in speech recognition accuracy, will likely foster overcoming the reluctance to use speech as a fully-fledged input modality. According to Google around 1 out of 4 search queries on Android devices are entered using Google Voice Search (Kincaid, 2011). Although the usage rate is lower for Apple's iPhone, the implementation of Google Voice Actions with Android Froyo shows that speech is becoming an important input modality for mobile devices. Thus, there is a need to investigate when and why users prefer speech over touch or vice versa in order to improve multimodal devices (Kamvar & Beeferman, 2010).

While multimodal output is a major research topic, e.g. (Prewett et al., 2006; Sarter, 2006), research on multimodal input has been focusing on specific modality combinations, namely speech-/pen- or speech/gesture input and highly specific tasks like spatial-verbal navigation on maps (Cohen, McGee, & Clow, 2000; Oviatt, 2003). For these tasks and systems the benefits of multimodality and flexible modality selection strategies are well documented.

However, for different tasks, for multimodal systems using other input modalities, or for systems offering serial (instead of simultaneous) input the findings are less clear: For instance, it has been shown that, although multimodality may lead to a higher perceived quality, users often do not employ multimodal interaction strategies and instead stick to one input modality (Wechsung, Naumann, & Hurtienne, 2009; Naumann, Wechsung, & Möller, 2008). Moreover, only few studies explicitly address modality selection strategies. Next to predominately exploratory work (Chen & Tremain, 2006; Althoff et al., 2003; Lemmelä et al., 2008), the experimental studies systematically investigating modality selection have mainly been concerned with error-recovery (Chen & Tremaine, 2006; Sturm & Boves, 2005; Suhm, Myers & Waibel 2001; Lai, Mitchell, & Pavlovski, 2007).

Those studies indicate that at least for the initial correction attempt users tend to stay in the current modality and do not switch the modality. An explanation might be that modality switches require the development of a new problem solving strategy,

Ina Wechsung, *An Evaluation Framework for Multimodal Interaction*,
T-Labs Series in Telecommunication Services,
DOI: 10.1007/978-3-319-03810-0_6, © Springer International Publishing Switzerland 2014

which involves more cognitive effort than reapplying the same modality and problem solving strategy.

Although some studies provide ambiguous results, a clear advantage of offering speech as an additional input modality could rarely be observed. A preference for speech was only shown for situations where the visual channel was already occupied. For example, Lemmelä et al. (2008) showed that speech is preferred in an in-car scenario. Also (Bilici et al., 2000) found speech input to be superior over touch input while driving. Rudnicky (1993) reported speech being preferred by the users and being more efficient than the other input modalities in situations with limited visual feedback. However, if none of the human perception channels is busy due to the situational constraints, the majority of studies showed that the most frequently chosen and often also best rated modality is input via the touch screen (Wechsung, Naumann, & Hurtienne, 2009; Naumann, Wechsung, & Möller, 2008; Raisamo, 1999; Sturm et al., 2002; Lamel et al., 2002).

Interestingly, in most of the studies mentioned above an increase of speech input was observed for specific tasks. For example, speech was preferred for selecting an element out of a very long list (Raisamo, 1999), for entering a long telephone number (Naumann, Wechsung, & Möller, 2008), or for searching for specific titles (Wechsung, Naumann, & Hurtienne, 2009; Naumann, Wechsung, & Möller, 2008; Metze et al., 2009). All these tasks have in common that speech input offered a shortcut in terms of a reduced number of necessary interaction steps and was thus more efficient than the other input modalities.

According to the findings reported above, the preference of one modality over another seems to be highly dependent on the efficiency of that very modality, situational demands and individual user preferences. Hence, in the following sections, studies systematically investigating those three factors are presented.

6.1 Efficiency

The assumption of a modality being favoured when more efficient is supported by Bilici et al. (2000), who compared text-input to speech input: Speech input was reported to be more efficient and more preferred. It has to be noted that efficiency was not systematically varied in this study. The conclusion of speech being preferred if more efficient was based on the observation that numbers were less likely to be entered via speech than input requiring more keystrokes (Bilici et al., 2000).

These findings are opposed by the results of Rudnicky (1993), where speech input was the less efficient but most preferred modality compared to keyboard and a scrolling bar. However, efficiency was measured in task-completion time and not in interaction steps. Hence, the explanation of the results presented by Rudnicky (1993) might be that speech actually was more efficient in terms of interaction steps. Also Kamvar and Beeferman (2010) report longer inputs to be less likely being entered with speech in a web search task. However, they used real-life log data and could thus not check if input was entered using the copy and paste-function. In this case, the GUI would be the most efficient modality in terms of interaction step.

Hence, although research indicates that shortcuts have an influence on modality choice, a systematic investigation is still missing. In view of these findings an initial study was conducted, aiming to systematically investigate how many interaction steps a shortcut must offer to skip in order to lead to a change in modality selection. Furthermore, the influence of efficiency on perceived mental effort was examined.

Since differences in interacting strategies between expert and novice users are well documented (e.g. Kamm, Litman, & Walker, 1998; Lazonder, Harm, & Wopereis, 2000; Petrelli, et al. 1997), expertise was taken into account as an additional variable.

6.1.1 Study 6.1

Method

Participants

Twenty-six German-speaking subjects voluntarily participated in our study. All participants were either PhD students or student workers in the area of human-computer-interaction. Thus, all of them were assumed to be expert users regarding information and communication technology. One of the participants was excluded from further analysis as he did not follow the experimenter's instructions. The remaining 25 participants were aged between 22 and 39 years (*M*=29 years, *SD*= 3 y., 7 female). Twelve of them were experts regarding speech dialogue systems, working or having previously worked in the area of speech recognition or voice user interface design, and were frequently using such systems. The other 13 participants were novice users with a different research background (e.g. tactile interaction, brain-computer interfaces). These users were using such systems seldom. All of them were familiar with virtual keyboards and were regularly using them.

Material

The application, a mobile jukebox, was installed on an Android-based smart-phone (G1 HTC Dream, cf. Figure 6.1). The available input modalities were touch using the virtual keyboard and speech (Nuance Vocon). Speech recognition was started via a virtual push-to-talk button. Different types of auditory feedback were implemented for the system states recognition active, match, and no match. The end of the utterance was detected automatically by the recognizer. If more than one result was found in the grammar, an n-best list was presented. If only one artist was recognized, a direct forwarding to the available tracks and albums took place. If no artist was found, a pop-up reading ‘I did not understand you’ was presented.

For input via the virtual keyboard only two scenarios were possible: Either the direct forwarding took place, or a pop-up reading ‘Nothing found’ was shown. Substring search (e.g. just typing QUE instead of QUEEN) was not supported and the systems had zero error tolerance; i.e. all misspelled search request were rejected.

The participants had to search for ten different artists. The length of the artists’ names increased by one character from one task to the next. The shortest artist name had two characters, the longest had eleven. Thus, the characters plus one (pressing the search-button) equalled the minimal number of necessary interactions steps in the keyboard conditions. For speech, the minimal number was always one (plus speaking the name) as only pressing the push-to-talk button was required.

To measure perceived mental effort, the SEA-scale (*Subjektiv Erlebte Anstrengung*; Eilers, Nachreiner, & Hänecke, 1986) was employed. The SEA-scale is the German version of the Rating Scale Mental Effort (RMSE also known as the SMEQ; Zijlstra & Van Doorn, 1985). The SEA was chosen as it is a lightweight instrument shown to have excellent psychometric properties (Sauro & Kindlund, 2009) even in comparison to more elaborate measures (De Waard, 1996).

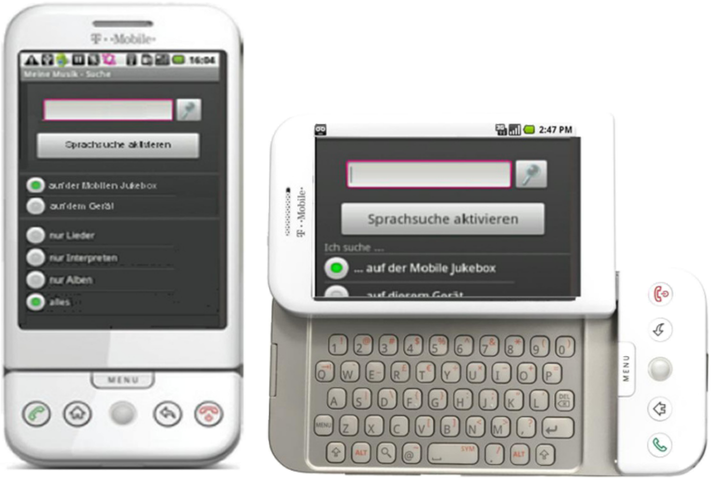


Fig. 6.1. Device used in Study 5.3.

Procedure

The participants were first asked to answer demographic questions. Although all individuals were recruited based on their research background, they were asked to self-assess their expertise with spoken dialogue systems. Only subjects where the research background and matched the self-assessed expertise were entitled experts.

Next, the device, the virtual keyboard, the speech control, and the tasks were explained. For each task, the artists were presented in written and oral form. Every participant received the same order of tasks. The test always started with the shortest name (two characters) and ended with the longest (11 characters). The task was accomplished if the artist was in the results list or if the page displaying the artist's albums and tracks was shown. If the artist was not in the list, the participants had to search again. This was repeated until the name was in the list, or until the artist's page was shown. After each accomplished task, the SEA-scale had to be filled in. The participants could choose and switch the input modality at any time. The modality switches and the number of trials were logged.

Results

Perceived Mental Effort

Regarding the perceived overall mental effort (averaged over all tasks), differences between expert and novice users were not observed. When analysing the single tasks differences between the user groups were found for Task 7 (8 characters, t (23) = 1.92, p =.034) and task 10 (11 characters, t (17.57) = 2.53, p =.011). Novice users reported higher mental effort than expert users (cf. Figure 6.2).

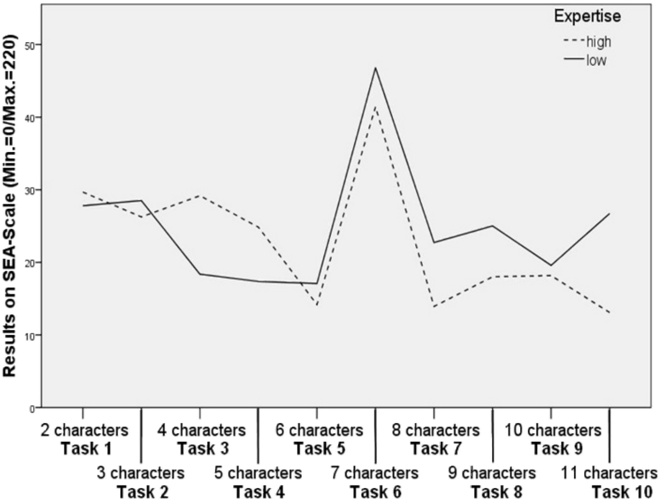


Fig. 6.2. Perceived mental effort by expertise and task

Differences were also found between the tasks: Task 6 (7 characters) was perceived as most demanding (F (4.25, 97.83) =4.28, p=.002, cf. Fig. 6.2).

Task 6 was also the task requiring the most attempts and thus being most error prone (F (5.40, 124.28) =3.06, p=.05, cf. Figure 6.3). The number of attempts and the perceived mental effort were significantly correlated (Pearson's r_{overall} (248) = .55, p<0.01, Pearson's r_{novice} (118) = .53, p<0.01, Pearson's r_{experts} (128) = .58, p<0.01).

Furthermore, participants using more speech tended to report less mental effort (Pearson's $r_{\text{overall}}(26)$= -.32, p =.056, Pearson's r_{novice} (14) = -.31, p=.140, Pearson's $r_{\text{experts}}(12)$ =-.48, p=.059).

No correlation was shown between perceived mental effort and the number of characters.

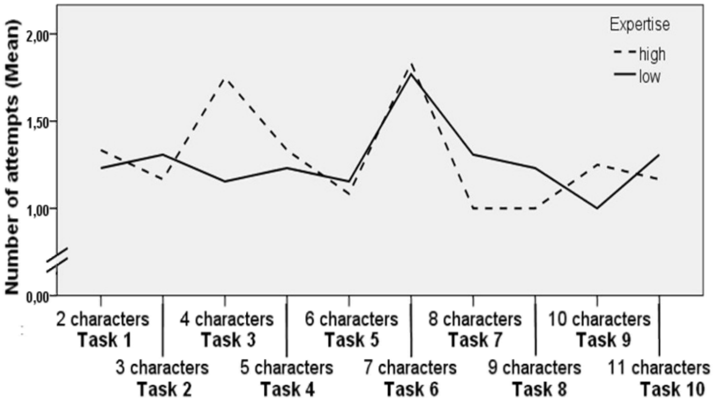


Fig. 6.3. Number of attempts by expertise and task

Interaction Steps and Modality Selection

The modality selection strategy for the first attempt did not differ significantly between the first three tasks. For Task 4 (5 characters) speech usage increased and stayed similar for the last six tasks (cf. Fig. 6.4). Thus, if the speech short-cut offered to skip at least four interaction steps the modality selection strategy changed resulting in an increase of speech usage (McNemar Test: χ^2 (1, N=25) = 2.77, p=.002).

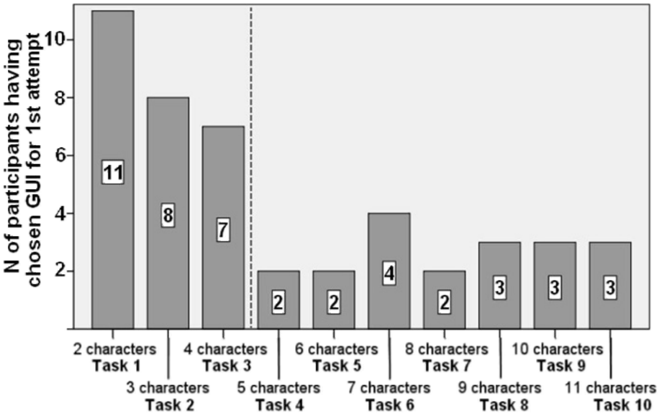


Fig. 6.4. Number of participants choosing GUI (touch) for first attempt by task. Dotted line displays tasks differing in modality selection strategy.

A separate analysis of expert and novice users revealed that for experts the threshold is one interaction step lower: The shortcut needs to offer only three interaction steps that can be skipped to significantly change the modality preference (cf. Table 6.1). For novice users the results are ambiguous: Although the modality selection strategy switched after Task 3 (4 characters), another switch was observed before and after Task 6 (7 characters).

Table 6.1. Results for modality selection strategies by expertise and task.

Task	Expert			Novice		
	% Speech (N)	% Touch (N)	χ^2 (*p*)	% Speech (N)	% Touch (N)	χ^2 (*p*)
1 (2 char.)	58.3 (7)	41.7 (5)	.33 (.387)	53.8 (7)	46.2 (6)	.077 (.500)
2 (3 char.)	66.7 (8)	33.3 (4)	1.33 (.194)	69.2 (9)	30.8 (4)	1.92 (.134)
3 (4 char.)	83.3 (10)	16.7 (2)	5.33 (.015)*	61.5 (8)	38.5 (5)	.69 (.291)
4 (5 char.)	100 (12)	0 (0)	-	84.6 (11)	15.4 (2)	6.23 (.011)*
5 (6 char.)	100 (12)	0 (0)	-	84.6 (11)	15.4 (2)	6.23 (.011)*
6 (7 char.)	100 (12)	0 (0)	-	69.2 (9)	30.8 (4)	1.92 (.134)
7 (8 char.)	100 (12)	0 (0)	-	84.6 (11)	15.4 (2)	6.23 (.011)*
8 (9 char.)	91.7 (11)	8.3 (1)	8.33 (.003)*	84.6 (11)	15.4 (2)	6.31 (.011)*
9 (10 char.)	100 (12)	0 (0)	-	76.9 (10)	23.1 (3)	3.77 (.046)*
10 (11 char.)	100 (12)	0 (0)	-	76. (10)	23.1 (3)	3.77 (.046)*

Note. **p*<.05

Task Success and Modality Switches

The majority of participants did not change the input modality after the first unsuccessful attempt. Even after the second unsuccessful attempt, a multitude of participants did not switch the input modality. If a third attempt was necessary, nearly all participants switched the modality. This pattern was the same for expert and novice users (cf. Table 6.2). Interestingly none of the novice users needed a fourth attempt.

Table 6.2. Modality switches for additional attempts by expertise.

% of Modality Switches	Expert		Novice	
	M	*SD*	*M*	*SD*
1st to 2nd attempt	27.3	38.9	40.2	42.3
2nd to 3rd attempt	75.0	41.8	64.3	47.6
3rd to 4th attempt	100	0	-	-

Intermediate Discussion

The results show that perceived mental effort is hardly affected by expertise but strongly related to effectiveness (task success). Moreover, speech input tends to reduce cognitive load. However, it was shown that task success (number of attempts) and mental effort are closely related. Therefore, the correlation between speech usage and mental effort may be confounded with task success. This was confirmed in a post-hoc analysis: A partial correlation between speech usage and mental effort with task success as a control variable showed no significant effect.

Also for modality selection, the number of necessary interaction steps is more relevant than expertise. In more detail, the results indicate that if at least three to four interaction steps can be skipped, speech is strongly preferred over touch. An influence of effectiveness on modality selection could be shown, but mainly if a modality failed more than twice. After switching the modality once (from keyboard to speech), participants tend to stick to that modality, aiming for some kind of internal interaction consistency. Thus, in the current study the most important factor is the number of interactions steps.

A point of criticism about the study is that participants could anticipate the increase of necessary interactions steps as the order of tasks was not randomized and the length of the artists' names increased by one character for each task. If this increase would have been less apparent the threshold probably would not have been this clear. Moreover, the virtual keyboard of the used device is rather small; a more convenient option may have led to a higher threshold.

Additionally, all of the participants were researchers in the area of human-computer-interaction and can hence be seen as a very specific user group.

To overcome these limitations and verify the initial results from Study 6.1, a follow-up study was conducted. In this study, a physical keyboard was included, the tasks were randomized and the user group was not limited to HCI researchers.

6.1.2 Study 6.2

Method

Participants

Thirty-four German-speaking subjects (*M* = 25 years, *SD* = 5 y, 13 female) participated in the study. In return, they received small gift. The majority of participants were students of various disciplines like engineering, phonetics, business, etc. None of them had any experience with the application. Half of the participants were offered speech and virtual keyboard as input modalities. The other half was instructed to use either the physical keyboard or speech.

Material

The device and application were the same as in Study 6.1. But in this study, also the physical keyboard was used.

The participants' task was to search for ten different artists. The length of the artists' names varied between three and 12 characters[16].

All artist names were either German names or names easy to pronounce for Germans. The names were: *PUR, Juli, Heino, Markus, Madonna, Nazareth, Kraftwerk, Extrabreit, Fettes Brot, Revolverheld.*

To measure perceived mental effort, again the SEA-scale (Eilers, Nachreiner, & Hänecke, 1986) was employed.

Procedure

The procedure was the same as in Study 6.1 except for the order of tasks, which was randomized in this study.

Results

Perceived Mental Effort

Regarding the SEA-scale values, neither differences between the tasks ($F(3.44, 106.81) = 1.06$, $p = .390$, part. eta^2=.033) nor an interaction effect between task and keyboard condition ($F(3.44, 106.81) = 1.44$, $p = .233$, part. eta^2=.044) could be observed. For keyboard a main effect was shown: the condition with the virtual keyboard was rated as more effortful, $F(1,31)= 5.31$, $p = 0.28$, part. eta^2 = .146.

The SEA-scale was filled in only once after the successful task completion. In Study 6.1, it was observed that effectiveness (task success) had a major influence on mental effort. To control for this effect and to investigate if speech usage "per se" reduces cognitive load, SEA-values of tasks not being solved in the first trial were excluded.

[16] To verify that the results of Study 6.1 are not dependent on the artist names, different names were used. In the application's database, only one artist name was two characters long (U2). Hence, for Study 6.2 the names varied between 3 and 12 characters, and not as in Study 6.1 between 2 and 11 characters.

A correlation between speech usage and perceived mental effort could not be observed, Pearson's $r_{\text{overall}}(10) = .25$, $p = .252$, Pearson's $r_{\text{virtual}}(10) = -.22$, $p = .271$, Pearson's $r_{\text{physical}}(10) = .48$, $p = .079$. This may have been caused by a trade-off between the task requirements and speech usage: longer names are more complex and might thus be more demanding to process. At the same time speech usage is correlated with the length of the names (cf. Section *Interaction Steps and Modality Selection*). Thus, the higher complexity of the longer names might have masked an effect for speech usage on mental effort. Hence, a partial correlation (*pr*) between modality usage and perceived mental effort was calculated, controlled for the length of the names resp. the necessary interaction steps. But again no significant results were obtained, $pr_{\text{overall}} = -.33$, $p = .196$, $pr_{\text{virtual}} = -.25$, $p = .259$, $pr_{\text{physical}} =.04$, $p = .457$.

Interaction Steps and Modality Selection

In both conditions, a correlation between interaction steps and modality usage was observed, Pearson's r_{overall} $(10) = .88$, $p = .000$. The more interaction steps necessary in the keyboard condition, the more participants selected the speech input (cf. Fig. 6.5). The correlation was slightly higher for the condition with virtual keyboard than for the condition with physical keyboard, Pearson's $r_{\text{virtual}}(10) = .76$, $p = .005$, Pearson's $r_{\text{physical}}(10) =.72$, $p = .009$).

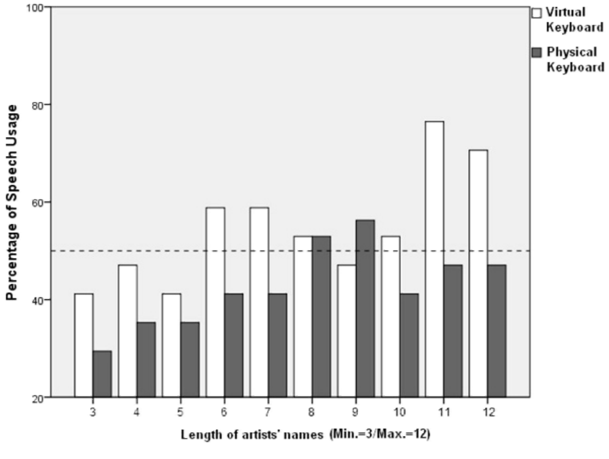


Fig. 6.5. Speech usage for first trial by length of artists' names (necessary interaction steps) and keyboard condition

To ensure that not efficiency in terms of task duration (instead in terms of interaction steps) caused these results, partial correlations controlling for task duration were calculated. Except for the condition with physical keyboard, the results support the hypothesis that the necessary interaction steps, and not the task duration, determine modality selection, $pr_{\text{overall}} = .66$, $p = .027$, $pr_{\text{virtual}} = .77$, $p = .008$, $pr_{\text{physical}} =.43$, $p =$

.123. Moreover, task duration was positively correlated with speech usage overall (including both conditions) and for the condition with physical keyboard, Pearson's $r_{\mathrm{overall}}(10) = .89$, $p = .000$, Pearson's $r_{\mathrm{virtual}}(10) = .30$, $p = .197$, Pearson's $r_{\mathrm{physical}}(10) =.73$, $p = .008$. When applying partial correlation between speech usage and task duration controlling for interaction steps, the correlation vanished, $pr_{\mathrm{overall}} = .26$, $p = .26$, $pr_{\mathrm{virtual}} = -.33$, $p = .19$, $pr_{\mathrm{physical}} =.45$, $p = .115$. Thus, efficiency in terms of task duration had no influence on modality selection.

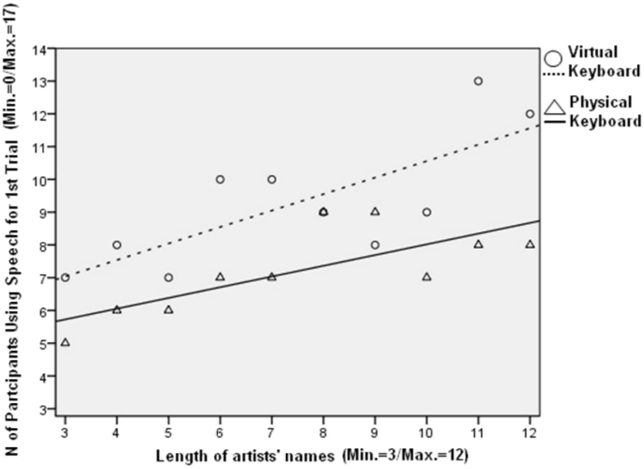


Fig. 6.6. Correlations between number of participants using speech in the first trial and length of artists' names (resp. minimum of necessary interaction steps) by keyboard condition. Lines displays regression lines.

To compare the magnitude of the increase in speech usage with increasing number of interaction steps, linear regression was applied for both conditions separately (virtual vs. physical keyboard). As expected, the regression slope was higher for the condition with virtual keyboard than for the condition with physical keyboard (cf. Fig. 6.6). However, the confidence intervals of the regression coefficients B (which are equivalent to the slope of the regression line) did overlap. Thus, the difference in slope was not significant (Table 6.3).

Table 6.3. Results of linear regression analyses. Dependent variable: percentage of speech usage. Independent variable: length of names (resp. necessary interactions steps).

	Virtual Keyboard	Physical Keyboard
SE B	2.96 (.89)	1.99 (.67)
CI for B lower bound	.90	.45
CI for B upper bound	5.01	3.53
ß	.76**	.72**
R^2 (Adjusted R^2)	.58 (.53)	.53 (.47)

Note.**p<.01

In a next step, χ^2-tests were calculated to determine the exact threshold of a preference for speech (cf. Figure 6.5). For both conditions, only for the task with the shortest artist name (3 char.), the keyboard was slightly preferred over speech (χ^2_{overall} (1, *N*=34) =2.94, p= .043) whereas for all other tasks, speech and keyboard were used equally frequently.

A separate analysis for the different keyboard conditions showed that with virtual keyboard, speech and keyboard are used equally often for the names between three and ten characters. If the artists' name was longer than ten characters, speech was significantly preferred, 11 characters: χ^2_{virtual} (1, *N*= 34) = 4e.77, p= .029/ 12 characters: χ^2_{virtual} (1, *N* = 34) = 2.88, p = .045. The virtual keyboard was used equally as often as speech for eight of the tasks, but was never significantly preferred in this condition.

In the physical keyboard condition, the keyboard was significantly preferred over speech for the task with the shortest name, 3 characters: χ^2_{virtual} (1, *N* = 34) = 2.88, p= .045. For all other tasks the modality usage pattern was the same. In this condition speech, although it was used equally often for the majority of tasks, was never significantly preferred over keyboard.

Task Success and Modality Switches

Our results are in line with the observation in Study 6.1: After the first task failure, the majority of participants (71 %) did not switch the modality (Wilcoxon Z = 2.21, p =. 027). After the second task failure, the frequency of switches equalled the frequencies of non-switches (Wilcoxon Z = .33, p =. 739).

Intermediate Discussion

The results confirm the earlier findings of Study 6.1: Modality selection is strongly influenced by the number of necessary interaction steps. Furthermore, task failure in one modality does not immediately result in switching to the other modality.

Thus, for longer inputs speech is advantageous over keyboard input. This preference for speech was stronger for virtual keyboard than for physical keyboards. As physical keyboard are more and more replaced by virtual keyboards, speech input might get more popular.

However, the threshold of speech being preferred over keyboard was higher than in the first study. A possible explanation is the task order: In Study 6.1, participants could anticipate that the required inputs get longer each task. Thus, they could be sure that, once they switched from GUI to speech, for all the following tasks speech input offers a clear shortcut. This means, they could be sure that they did not have to switch the modality again for very short artist names. In addition, the different sample of participants may have caused an increase of the threshold. Whereas the previous study included only HCI researcher, the current study tested a wider range of users (although most of them were students). It may be the case that HCI researchers are more familiar to systems offering speech input and are thus less hesitant to use it.

Both reported studies imply that the number of necessary interaction steps is crucial for modality choice. However, in both of the studies only text and not numbers had to be entered. Since research reports that numbers are less likely to be entered via speech (Bilici et al. 2000), the results may have been different for telephone numbers instead of artists.

Furthermore, the context of the usage situation was not considered: Even if speech input offers shortcuts, privacy concerns or parallel tasks might deter users from using speech input when using the device in public. Thus, in the following section the influence of such situational demands will be examined.

6.2 Situational Demands

Besides the classical usability metrics of efficiency and effectiveness discussed above, other relevant factors which have been identified are the setting, the domain (Kamvar & Beeferman, 2010), the task (Chen & Tremaine, 2006), and the prior knowledge about the system (Jokinen & Hurtig, 2006).

Althoff and colleagues (2003) showed that in an automotive setting, speech is the dominant input modality whereas in a desktop environment keyboard was preferred. Nevertheless, it is not clear if their results are actually due to the different settings or due to other factors e.g. the different task types they employed: In the desktop setting users had to carry out navigational tasks, while in the automotive setting an entertainment system had to be controlled while driving in a simulator. However, in the automotive setting, the task of driving, obviously leads to a strong preference of speech. For instance, Lemmelä et al. (2008) report that speech is preferred over gestures in an in-car scenario, whereas gestures are preferred while walking. 2D gestures, the input method requiring the most visual attention, were in both scenarios rarely used.

Regarding the domain, as one might expect, the likelihood of speech input decreases for topics which are considered as confidential or private domain (Kamvar & Beeferman, 2010).

Evidence for the influence of task type is provided by a study by Chen and Tremaine (2006). In a test set-up using an audio browser, they report in line with Althoff et al. (2003), that navigational task tend to be less likely being performed via speech than other tasks. Also, Gong (2003) reported task-specific modality preferences (speech vs. stylus) when interacting with a PDA: Again for navigation, stylus was preferred over speech. In a study by Jokinen and Hurtig (2006), the stated modality preference differed depending on the prior knowledge about the system. Participants received either the information that the tactile modality is supplemental to the speech modality, or that speech is supplementing the tactile modality. Both groups preferred the respective supplemental modality.

The majority of the studies including situational demands (Chen & Tremaine, 2006; Althoff et al., 2003; Lemmelä et al. 2008) were of exploratory character. Cox and colleagues (2008) provide a controlled experiment: To imitate a situation with the visual channel being busy (e.g. walking or driving) they offered only limited visual feedback for a text creation task and found speech being preferred over keyboard. It has to be noted that the crucial aspect for (visual) mobile interaction might not only be that the visual feedback is limited due to small displays, but that the visual channel might get overloaded and that visual attention needs to be shared with other concurrent tasks, e.g. walking.

In summary, situational demands seem to influence modality choice and are often related to the allocation of attentional resources. This assumption is supported by cognitive theories (Baddeley & Hitch, 1974; Baddeley, 1992, 2003; Paivio, 1986, Wickens, 1984, 2002) proposing multiple resources for different sensory modalities which may be helpful in explaining the findings above (cf. Chapter 2.2).

Those theories assume multiple cognitive resources and imply that attention can be shared between tasks if the tasks refer to different resources. All three models differentiate between modality-specific subsystems. E.g. Paivio (1986) proposes two different long term storage systems for verbal and non-verbal information. Baddeley's (Baddeley & Hitch, 1974; Baddeley, 2003) model of the working memory has in its original version three different, capacity-limited components:

- the phonological loop, processing verbal information,
- the visual-spatial sketchpad, the short-term storage system for visual and spatial information, and
- the modality-unspecific central executive, which is a general processing capacity controlling the two aforementioned sub-systems and additionally dividing and switching attention.

To link these subsystems, another component, whose function initially was ascribed to the central executive, was later added, namely the episodic buffer. It serves to integrate information from different sensory modalities into one coherent experience or episode. Regarding the two different modules, visual sketchpad and phonological loop, the theory predicts that visual-spatial information can still be processed when the phonological loop is occupied (Mayer & Moreno, 1998). However, the theory is less specific for other situations where different perceptual channels are employed. For example, auditory-verbal information is supposed to be ’recorded’ on the phonological loop automatically, whereas verbal information presented visually (e.g. text) might also enter the phonological loop, if it is recoded into phonological code by silent sub-vocalisation (Baddeley, 1992). The Multiple Resource Theory proposed by Wickens (1984) would make the same prediction regarding the parallel processing of visual-spatial (e.g. driving) and phonological information (e.g. listening). However, while Baddeley mainly specifies the influence of the (later) processing codes, Wickens also differentiates perceptual modalities. Here, auditory and visual input refer to different perceptual resources, although they might use the same cognitive resources on later stages in processing, since the different perceptual modalities might lead to the same internal representation or code, as already explained for spoken and written text.

These theories can be especially helpful when designing mobile systems, as the aforementioned situational demands of (mobile) interaction can be taken into account. However, these theories apparently have mainly been employed for investigating and designing multimodal output for in-car or in-cockpit scenarios (Sarter, 2006). In this section, they are applied to multimodal input, adapting the parallel-task paradigm from cognitive psychology.

6.2.1 Study 6.3

The purpose of the current study is to investigate multimodal input from the user and the predictions based on the previous work of Baddeley (1992) and the Multiple Resource Theory (Wickens, 1984). Specifically, it was aimed to simulate shared visual or auditory attention, which is characteristic for mobile situations, through a secondary task. It was expected that a task requiring visual attention will decrease performance and increase mental effort while using touch input, and that a task requiring auditory attention will decrease performance and increase mental effort while performing speech input. These predictions are derived from the Multiple Resource Theory: The touch interaction is visual demanding, as the screen is the input modality, and should thus interfere with the visual secondary task. On the other hand, speech input is auditory demanding and should then interfere with the auditory secondary task. If participants are free to choose an input modality, speech should be used more often in the visual attention condition than in the auditory attention condition, while for touch the results should be vice versa.

Moreover, the study investigated the influence of situational demands on perceived quality. Hassenzahl et al. (Hassenzahl, Kekez, & Burmester, 2002) showed the influence of different usage situations on global quality ratings. For different usage modes, i.e. goal mode vs. non-goal mode, different qualities determine a product's overall appeal. Whereas the influence of a systems non-functional hedonic qualities (e.g. novelty) was constant in both modes, the influence of the functional pragmatic qualities (efficiency) on overall appeal differed between the modes. Pragmatic qualities were less important in the non-goal mode, where no task had to be accomplished, compared to the goal mode, where the system was used to complete a certain task. Accordingly, the judgments of hedonic qualities were expected to be more or less stable, which means that an interaction effect between secondary task and input modality of the primary task was not expected for hedonic qualities. Still the input modalities may differ regarding their perceived hedonic quality. Pragmatic qualities refer to usability-related attributes like efficiency and effectiveness. If the secondary task is presented in the same modality as the input quality, the efficiency and effectiveness of the system should be impaired. Hence, lower ratings for pragmatic qualities were expected in the same modality conditions compared to the cross-modal conditions.

Method

Participants

Twenty-four German-speaking subjects participated in our study. To control for previous experience with the input modalities, it was stated in the invitation that only persons without experience with spoken dialogue systems could participate, whereas prior experience with touch input was permitted. Those criteria were chosen, as experience with touch screen is more or less unavoidable. Speech input, on the other hand, is becoming more popular, but only few people in our subject database use it frequently. However, five of the invited subjects did not meet the criteria according to the information they provided in the introductory questionnaire and were excluded from further analysis. The remaining participants were aged between 23 and 33 years (*M* = 25.8 years, *SD* = 2.4, 10 female). The majority of participants were students. None of them had any experience with the application used in the test.

Design

In a 2x3 design the factors secondary task (visual vs. auditory) and input modality (speech vs. touch vs. multimodal) were manipulated. Secondary task was a between factor with nine participants in the auditory condition and ten participants in the visual condition; input modality was a within (repeated measurements) factor. All participants were presented with two unimodal blocks, speech input and touch input, and a multimodal block, where they could freely choose the input modality. The order of the two unimodal blocks was alternated for each participant. The multimodal block was always presented at the end, in order to have the participants trained on each input modality.

Device

The application and the device were the same as in Study 6.1 and Study 6.2. However, for this study an additional push to talk-button was implemented on the back of the device. Aim was to prevent the potential visual load imposed by the virtual push to talk-button. Furthermore, output was presented via additional loudspeakers, as the devices' internal loudspeakers were rather weak. None of the auditory feedback provided for the speech interface was similar to the stimuli of the secondary task described later. Visual output was always available, as the display was not turned off in the speech block. In addition, auditory feedback, like the music that had to be played in the last set of interaction tasks (see next section), was always presented. Thus, output was multimodal in all conditions, and just the input modalities were varied.

Primary Tasks

Overall, 13 different tasks grouped into three types had to be carried out. Amongst them were four menu navigation tasks requiring to switch the different hierarchy levels in the music library, and after each level to go back to the main menu. In order to keep memory load due to memorizing the instruction as low as possible, the first task was on the first level, the second task was on the second level and so on. The next task was searching for five specific entries in the library. Again, to prevent memory load they had to enter the search terms 'first', 'second', 'third', 'fourth', 'fifth'. The last group of tasks dealt with controlling the music player, namely to call the commands 'play', 'next', 'shuffle' and 'stop'. The tasks were presented in this order to ensure a logical, easy to remember sequence.

Secondary Tasks

The secondary task was a "go/no-go task" requiring selective attention. In random intervals between three and five seconds, either the target stimulus or the distractor was presented for 500 ms, and the participants were instructed to hit a button whenever the target appeared, and to inhibit a reaction when the distractor was presented. In the visual condition, the target was a black circle and the distractor a black square; in the auditory condition, a high tone (1000hz) was the target stimulus and a low tone (200hz) was the distractor stimulus. The visual stimuli were presented on a 20 inch computer screen. Participants sat centred to the screen with circa 60 cm distance. The size of the square stimuli was 3x3 cm, the circle had a diameter of 3 cm. One loudspeaker, placed centrally (above the screen), was used for the presentation of the auditory stimuli. The loudness level was around 72 dB(A) at the participant's position.

The go/no-go task was chosen, as it requires sharing attention with the primary task, a typical demand in mobile interaction. The memory load induced by this kind of task is relatively low: In both conditions, only two simple rules ('if target press; if not target do not press') had to be maintained (Redick et al., 2011). Another reason for the go/no-go task was that it could be easily adapted for both perceptual modalities.

Measures

As measures for interaction performance (primary task), task completion time and percentage of successful first trials were used. As measures of performance in the secondary task, the reaction times and accuracy (the percentage of correct responses among all responses) were assessed. Reaction times were logged within the program.

Modality choice in the multimodal block was recorded online by the experimenter. For each task the modality (or combination of modalities) chosen by the participant was annotated. As in the previous studies, mental effort was measured using the SEA-scale (Eilers, Nachreiner, & Hänecke, 1986). To assess retrospective judgments of the perceived quality of the interaction, the Attrakdiff Mini (Hassenzahl & Monk, 2010) had to be filled in. This questionnaire is a short version of the AttrakDiff (Hassenzahl, Burmester, & Koller, 2003) which has already been described in the previous studies of this thesis. Ratings were collected on all scales of the AttrakDiff Mini, which are Attractiveness, Hedonic Qualities-Stimulation, Hedonic Qualities-Identity and Pragmatic Qualities. The two hedonic scales were merged into one, as no hypotheses for the sub-aspects of hedonic qualities were formulated.

Procedure

One session took between 90-120 minutes and was conducted in a laboratory setting. At first, the participants were requested to answer a short questionnaire to collect demographic information and previous experience with the input modalities.

After that, the device, the application, the different input modalities, and the secondary task were explained. Both tasks should be performed as well as possible, but the interaction task was prioritized. The general explanations were followed by a training phase for the secondary task, and a successive training phase for the input modality and the first set of interaction tasks. The specific instructions for the training phases were presented verbally and in writing. After each training phase, the participant was asked if he/she had understood the instructions and was feeling able to follow them. If the participants answered those questions positively, the first set of interaction tasks had to be carried out, while simultaneously performing the go/no-go task. When the first set of interaction tasks was finished, the SEA-scale and the AttrakDiff Mini had to be filled in. Then, the next set of interaction tasks were trained and after that performed by the participants along with the secondary task. Again, the questionnaires were presented. Next, was the training for the last set of interaction tasks. Once more, the interaction tasks and the secondary task had to be carried out and the questionnaires had to be filled in. This procedure was the same for all input modalities. An example procedure for a participant with auditory secondary task, is displayed in Figure 6.7. In the multimodal block, participants were told that they could freely choose or switch the input modality at any time. For each interaction task, participants had three attempts. After the third unsuccessful attempt, the task was cancelled and the next task was carried out.

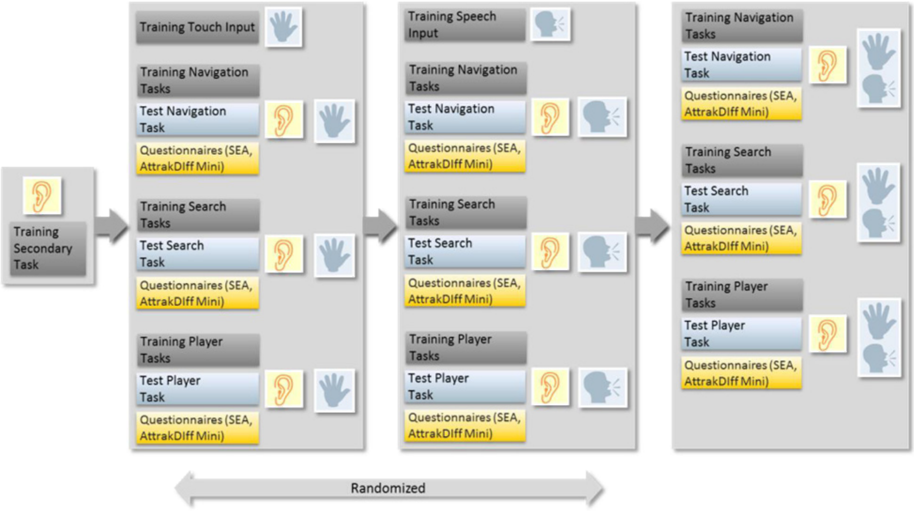

Training Secondary Task
Training Touch Input
Training Navigation Tasks
Test Navigation Task
Questionnaires (SEA, AttrakDiff Mini)
Training Search Tasks
Test Search Task
Questionnaires (SEA, AttrakDiff Mini)
Training Player Tasks
Test Player Task
Questionnaires (SEA, AttrakDiff Mini)
Training Speech Input
Training Navigation Tasks
Test Navigation Task
Questionnaires (SEA, AttrakDiff Mini)
Training Search Tasks
Test Search Task
Questionnaires (SEA, AttrakDiff Mini)
Training Player Tasks
Test Player Task
Questionnaires (SEA, AttrakDiff Mini)
Training Navigation Tasks
Test Navigation Task
Questionnaires (SEA, AttrakDiff Mini)
Training Search Tasks
Test Search Task
Questionnaires (SEA, AttrakDiff Mini)
Training Player Tasks
Test Player Task
Questionnaires (SEA, AttrakDiff Mini)
Randomized

Fig. 6.7. Procedure for one participant with auditory secondary task and the order touch, speech and multimodal input.

Results

In this section, results are reported. If a specific assumption was formulated a-priori (see. Section 6.2.1) one-tailed *p*-values are reported. For all other results, two-tailed p-values are presented.

Modality Choice

A mixed model Analysis Of Variance (ANOVA), with secondary task (visual vs. auditory) as between factor and with input modality (touch vs. speech vs. multimodal) as within factor, showed a main effect for input modality. In the multimodal condition, touch was the most frequently used modality, followed by a combination of the modalities. Speech was least frequently chosen, $M_{\text{speech}} = 1.58$, $SD_{\text{speech}} = 2.19$,

$M_{\text{touch}} = 6.95$, $SD_{\text{touch}} = 2.32$, $M_{\text{multimodal}} = 4.47$, $SD_{\text{multimodal}} = 1.22$, $F(2,34) = 26.60$, $p_{\text{two-tailed}} < .01$, *part. eta²* = .610. A Sidak-corrected post-hoc test showed significant differences between all three possible input modalities ($p_{\text{two-tailed}} <. 01$). Additionally, an interaction effect in the expected direction was observed, $F(2,34) = 2.76$, $p_{\text{one-tailed}} = .034$, *part. eta²* = .140. Speech was less often chosen with auditory secondary task than with visual secondary task. Results for touch are vice versa (cf. Fig. 6.8).

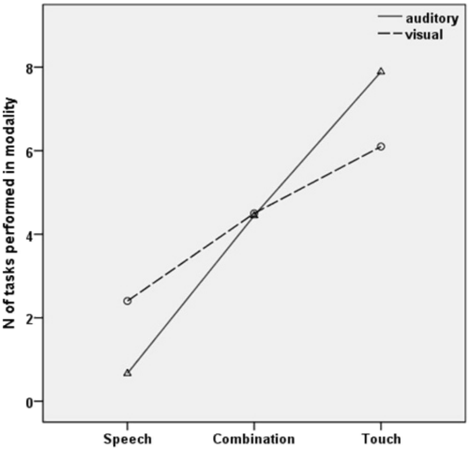


Fig. 6.8. Modality choice in multimodal block by secondary task.

Primary Task - Task Duration

The durations for all tasks were averaged for each input modality. Again a mixed model Analysis Of Variance (ANOVA) was calculated; the results showed a main effect for input modality, $M_{\text{speech}} = 75.00$, $SD_{\text{speech}} = 12.81$, $M_{\text{touch}} = 67.93$, $SD_{\text{touch}} = 17.92$, $M_{\text{multi}} = 50.89$, $SD_{\text{multi}} = 17.85$, $F(2,34) = 11.95$, $p_{\text{two-tailed}} < .01$, *part. eta²*= .413. Post-hoc tests with Sidak correction revealed the task duration in the multimodal block to be significantly shorter than in the unimodal blocks (touch: $p_{\text{two-tailed}} = .02$, speech: $p_{\text{two-tailed}} < .01$). Neither a main effect for secondary task type nor an interaction effect between input modality and secondary task type was observed for task duration.

Primary Task - Task Success

The number of successful first trials was counted for each test block. The results showed a main effect for input modality, $M_{\text{speech}} = 10.52$, $SD_{\text{speech}} = 1.74$, $M_{\text{touch}} = 10.74$, $SD_{\text{touch}} = 1.82$, $M_{\text{multi}} = 11.63$, $SD_{\text{multi}} = 1.26$, $F(2,34) = 3.54$, $p_{\text{two-tailed}} = .040$, *part. eta²*= .172. A Sidak-corrected post-hoc test indicated that participants were marginally more successful in the multimodal condition than in the speech condition ($p_{\text{two-tailed}} = .069$). Moreover, a marginally significant interaction effect between input modality and secondary task was observed, $F(2,34)=2.35$, $p_{\text{one-tailed}} = .055$, *part. eta*^{2}= .122. In line with our assumptions, performance in the speech block was worse with the auditory secondary task than with the visual secondary tasks. For touch input, the visual secondary task decreased task success compared to the auditory secondary task (cf. Figure 6.9). A main effect for type of secondary task was not shown.

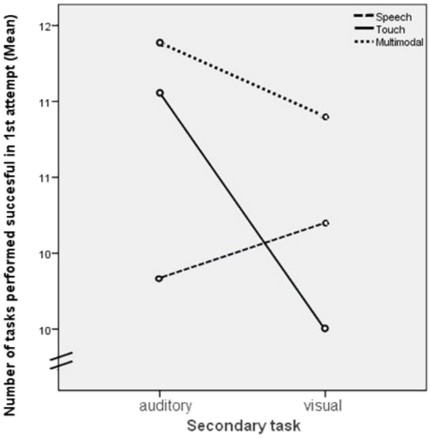


Fig. 6.9. Number of successful performed primary tasks with first attempt by secondary task and input modality

Secondary Task - Reaction Times

Regarding the reaction times, a main effect for input modality was observed, M_{speech} = .80, SD_{speech} = .17, M_{touch} = .93, SD_{touch} = .26, M_{multi} = .85, SD_{multi} = .26, $F(2,32) = 3.29$, $p_{\text{two-tailed}}$ = .050, *part. eta*^{2} = .171. Reaction times were significantly shorter in the speech condition than in the touch condition ($p_{\text{two-tailed}}$ = .014).

Also, the type of secondary task had an effect with faster reactions for the visual secondary task compared to the auditory task, M_{auditory}= .97, SD_{auditory}= .27, M_{visual} =.77, SD_{visual}= .14, $F(1,16)=7.27$, $p_{\text{two-tailed}}$ = .016, *part. eta*^{2}= .312.

Contrary to our expectation, an interaction effect between input modality and type of secondary task could not be observed.

Secondary Task - Accuracy

For accuracy, a main effect for input modality was observed, M_{speech} = 81.48, SD_{speech} = 12.94, M_{touch} = 83.69, SD_{touch} = 10.62, M_{multi} =89.19, SD_{multi} = 11.26, $F(2,32) = 4.35$, $p_{\text{two-tailed}}$ =.021, *part. eta*^{2}= .214. A post-hoc test with Sidak correction showed a significantly higher accuracy in the multimodal condition compared to the speech condition ($p_{\text{two-tailed}}$= .013).

Here, the expected interaction effect between input modality and secondary task could be observed ($F(2,32)$= 2.48, $p_{\text{one-tailed}}$=.050, *part. eta*^{2}= .129). The auditory secondary task was completed with higher accuracy in the touch condition compared to the speech condition. For the visual secondary task, the results were vice versa (cf. Fig. 6.10). A main effect for type of secondary task was not observed.

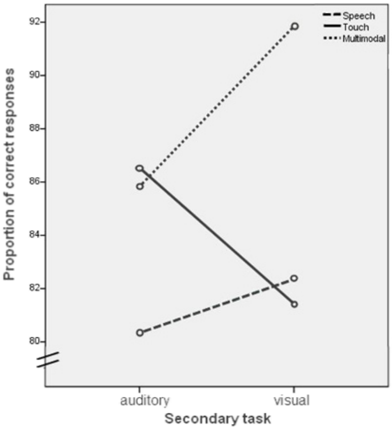


Fig. 6.10. Percentages of correct responses in secondary task by secondary task and input modality

Perceived Mental Effort

Perceived mental effort differed within the test blocks, $M_{\text{speech}} = 68.74$, $SD_{\text{speech}} = 40.50$, $M_{\text{touch}} = 67.23$, $SD_{\text{touch}} = 37.76$, $M_{\text{multi}} = 37.67$, $SD_{\text{multi}} = 35.94$, $F(2,34) = 10.42$, $p_{\text{two-tailed}} < .01$, *part. eta*^{2}= .380. The Sidak corrected post hoc test revealed speech ($p_{\text{two-tailed}} < .01$) and touch ($p_{\text{two-tailed}} < .01$) to be significantly more demanding than the multimodal block. Moreover, the predicted interaction between input modality and type of secondary task could be shown, $F(2,32)= 3.30$, $p_{\text{one-tailed}}=.025$, *part. eta*^{2}= .163. The speech condition was perceived as less straining with the visual secondary task, whereas the touch condition was less straining with the auditory secondary task (cf. Fig. 6.11). A main effect for type of secondary task was not observed.

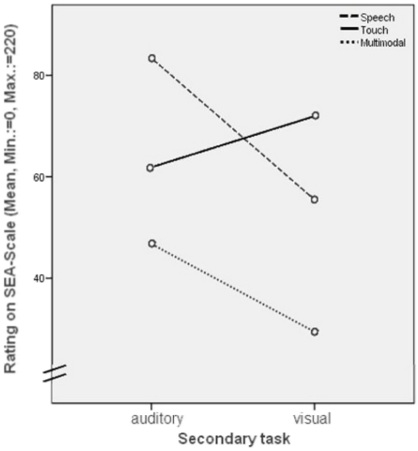


Fig. 6.11. Perceived mental effort by secondary task and input modality

Perceived Quality

Regarding the overall quality, the Attractiveness scale of the AttrakDiff, no main effects and only a marginal interaction effect was observed (cf. Table 6.4). As predicted, no effect was shown for the scale Hedonic Qualities. Regarding the Pragmatic Qualities, a main effect for input modality and an interaction effect between input modality and secondary task was found (cf. Table 6.4). Means and standard deviations are presented in Table 6.5.

Table 6.4 Analysis of Variance for AttrakDiff ratings

	Attractiveness			Pragmatic Qualities			Hedonic Qualities		
Source	*F (df)*	*p*	*part. eta²*	*F (df)*	*p*	*part. eta²*	*F (df)*	*p*	*part. eta²*
Input modality	1.60 (2,34)	.216†	.086	10.90 (2,34)	<.001**	.392	1.63 (2,34)	.105	.087
Secondary task	.17 (1,17)	.689†	.010	.07 (1,17)	.792†	.004	.09 (1,17)	.764†	.005
Secondary task x input modality	2.33 (2, 34)	.113†	.121	5.03 (2,34)	.012*	.228	1.23 (2,34)	.153	.067

Note. † Signifies two-tailed p-values, all other p-values are one-tailed.*p<.05, **p<.01

Table 6.5. Ratings on AttrakDiff (Min. -3/Max. 3) by secondary task and input modality

Input modality	Secondary task	Attractiveness		Pragmatic Qualities		Hedonic Qualities	
		M	*SD*	*M*	*SD*	*M*	*SD*
Speech	auditory	.59	.87	.57	.59	.19	1.19
	visual	.80	.84	1.10	.92	.18	.51
Touch	auditory	.85	.72	1.39	.62	-.13	1.16
	visual	.68	.63	.96	.67	-.08	.73
Multimodal	auditory	.74	.76	1.48	.66	-.05	.85
	visual	1.08	.61	1.59	.60	.17	.72

Intermediate Discussion

Based on the Multiple Resource Theory (Wickens, 1984) it was assumed that touch input via GUI, a spatially-visually demanding modality, interferes more with a visual secondary task than with an auditory secondary task. Vice versa results were expected for speech input, an auditory-verbally demanding input modality. The interference should be present for modality choice and performance measures as well as for questionnaire ratings of mental effort and perceived Pragmatic Qualities. Cross-modal conditions (speech input paired with visual secondary task and touch input paired with auditory secondary task) should be beneficial compared to the respective same modality conditions. In terms of modality selection strategies, an auditory secondary task was assumed to foster usage of the GUI while a visual secondary task was assumed to lead to a higher proportion of speech input.

The results are largely in line with our expectations. For modality choice, task success, accuracy, mental effort and pragmatic qualities, the expected effects could be shown. However, for task success the results were only marginally significant. Only for reaction times (in the secondary task) and task duration (in the primary task) the expected effects could not be shown. Thus, predictions based on Multiple Resource Theory seem to be as reliable to multimodal input as they are for multimodal output.

The results regarding the perceived qualities show that, as observed by Hassenzahl et al. (2002), perceptions of hedonic qualities are relatively stable. Apparently, users can distinguish quite well between the pragmatic value or usefulness of available input modalities and the hedonic qualities of a device. That means, participants seem to rate the hedonic properties of an interface independently from a product's suitability for a specific situation. Judgments of pragmatic qualities are less stable: If an interface's properties do not comply with the situational demands, the perceived pragmatic qualities will decrease.

However, it is reasonable to assume that not only the speech condition required processing of verbal code, but that also the touch condition did as the GUI was labelled verbally. Still, performance in the touch condition was more affected by the visual secondary task than by the auditory one. This observation provides evidence for Wickens (1984), assuming that perceptual modalities indeed refer to different resources, although they might be transformed into identical processing codes (i.e. the internal representation of the written word 'apple' is the same as for the spoken version of that word). Moreover, the multimodal condition, where users could adapt the input modality to the situational demands, was rated as less straining and led to a better performance. While this might be due to the test design, it might also indicate that one of the core benefits of multimodal interaction is its flexibility. In previous lab studies without such situational demands, such a superiority of multimodality was rarely or not observed (one of the example is Wechsung, Naumann, & Hurtienne, 2009).

However, although the type of distraction determines modality choice, the proportion of touch input was higher regardless of the type of distraction. Thus, people might not always switch automatically to a less interfering modality even if it is offered. This indicates that other factors like individual user characteristics are likely to influence modality selections strategies also. Therefore, another study focusing on such individual user characteristics was conducted.

6.3 User Characteristics

Although user characteristics are often assumed to influence quality perceptions and interaction behaviour (cf. Section 3.1.1), characteristics beyond gender, prior experience and age are seldom taken into account. Due to the demographic trend, especially age-related differences have received more attention during the past years (e.g. Wolters et al., 2010). However, it is often assumed that not the chronological age "per se" causes these differences but rather characteristics like a smaller degree of previous experience (Chalmers, 2003), the age-related decrease of cognitive abilities (Wolters et al., 2010), and motor impairments (Carmichael, 1999). Moreover, most of the human-computer-interaction research regarding individual differences is linked to ability related performance measures. Also, the vast majority of studies including individual differences were conducted with graphical user interfaces, with notable exceptions in the area of 3D environments (Chen, Czerwinski, & Macredie, 2000). To the best of the author`s knowledge, the influence of individual differences (such as personality or attitudes) on modality selection has only rarely been investigated systematically.

However, previous research indicates that users show individual preferences of one modality over another (Oviatt, Coulston, & Lunsford, 2004). In addition, the results previously reported in this chapter imply that user characteristics are a relevant factor in explaining modality choice; for instance, it was shown that some users stick to one modality even if the other modality is more efficient (cf. Sections 6.1, 6.2).

Therefore, the following study aims to investigate whether such individual properties of users can explain differences in modality selection and additionally in performance and quality ratings.

6.3.1 Study 6.4

In this section, the hypotheses of the current study are explained based on previous research.

- **Modality Choice.** Concerning modality selection, an influence of personality is assumed, as generally extroverts are reported to be more talkative. This was shown for a Thinking Aloud-test, where extroverts provided more feedback than introverts (Burnett & Ditsikas, 2006). Thus, it is plausible to assume that extraversion will show a positive correlation with the usage of speech.

  Moreover, positive attitudes towards technologies make adoption of new technologies more likely (Matilla, Karjaluoto, & Pento 2003). Likewise it was expected that users will tend to choose speech input – the more innovative technology – more often if they have positive attitudes towards technology. Negative attitudes will inhibit speech usage.
- **Performance.** Regarding performance and interaction, Wolters and colleagues (2009) found recall of the interaction with a spoken dialogue system to be influenced by information processing speed. Users with a lower processing speed were less likely to remember relevant information provided by the system. Additionally, working memory span was shown to influence recall and transfer performance, with a low span being especially disadvantageous in a mobile instructional environment (Doolittle, Terry, & Mariano, 2009). Based on the findings presented above, a positive correlation between cognitive abilities and performance was expected. Apart from cognitive abilities, also attitudes were shown to affect performance: Positive attitudes are related to better performance (Jawahar & Elango, 2001).
- **Quality Perceptions.** For quality perceptions, an influence of personality was shown by Burnett and Ditsikas (2006). They compared introverted and extroverted user and reported that extroverted users found more usability problems and gave slightly worse ratings on a post-hoc usability questionnaire. Thus, extraversion is expected to negatively correlate with perceived quality. Another factor assumed to influence evaluative ratings is mood (Bless et al. 1996). Memory recall is mood congruent: good moods make recall of positive experiences more likely than bad moods and vice versa (Kahneman, 1999). Consequently, a positive mood should result in better ratings.

In summary, it was expected that modality selection is determined by personality and attitudes, that performance measure are influenced by cognitive abilities and attitudes and that quality perceptions are affected by personality and mood.

Method

Participants

Thirty-three German-speaking individuals participated in the experiment. Three of them were excluded from the analysis as they showed zero or near zero variance in their ratings. The remaining subjects (14 m., 16 f.) were aged 18 and 62 (*M*= 45 y., *SD*=12 years). All participants were owners and users of the IPTV service for which the tested application was designed. However, none of them had any previous experience with the application itself.

Device and Tasks

The tested application, a multimodal remote control for an IPTV service, was implemented on an iPhone (cf. 6.12). Available modalities were touch and speech. Feedback was, depending on the task, either given via the TV or via the iPhone screen. Auditory feedback was provided, in case of speech input, for ‘speech control active’, ‘match’ and ‘no match’. For the touch and multimodal condition, 17 tasks had to be performed, compared to 16 tasks in the speech condition. The unequal number of tasks was due to one task not being solvable with speech. This task was excluded from further analysis. The tasks included in the analysis were zapping, switching the channel, increasing and decreasing the volume, setting and resetting time shift, opening the tele-text, retrieving a special site from the tele-text, opening and closing the electronic program guide starting and ending recording, retrieving information for the current broadcast, muting and unmuting the audio playback.

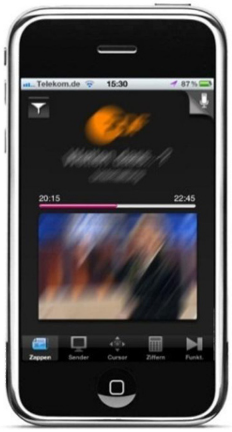


Fig. 6.12. Device used in Study 5.3.

Measures

- **Personality.** To assess users' personality traits the BFI-S was used (Schupp & Gerlitz, 2008). It is a German short version of the Big Five Inventory measuring the personality traits Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism. Openness descibes "the active seeking and appreciation of experiences for their own sake" (Borkenau & Ostendorf, 1994). Neuroticism measures the extent to which individuals "are prone to psychological distress" (Borkenau & Ostendorf, 1994). Extraversion is the "quantity and intensity of energy directed outwards into the social world" (Borkenau & Ostendorf, 1994). Agreeableness refers to the "kinds of interactions an individual prefers from compassion to tough mindedness" (Borkenau & Ostendorf, 1994). Conscientiousness is the "degree of organization, persistence, control and motivation in goal directed behaviour" an individual possesses (Borkenau & Ostendorf, 1994). Each trait is measured with three 5-point Likert scale items.
- **Attitudes.** Participants' attitudes were collected with the TA-EG questionnaire (Karrer et al., 2009) measuring general positive and negative attitudes towards information and communication technology (ICT) as well as competence and enthusiasm towards ICT. The questionnaire comprises 19 items using 5-point Likert-scales. Note that negative and positive attitudes are measured on separate scales. Thus, positive and negative attitudes are understood as two different constructs.
- **Mood.** To assess the participants' mood, an adapted version of the faces scale (Andrews & Withey, 1976) was used. The scale shows seven faces ranging from very sad to very happy.
- **Quality Perceptions.** To assess retrospective quality judgments of the interaction with the system, the AttrakDiff Mini (Hassenzahl & Monk, 2010) had to be filled in. Additionally, the SEA-scale was used to measure perceived mental effort (Eilers, Nachreiner, & Hänecke, 1986).
- **Cognitive Abilities.** The digit span test, a sub test of the Hamburg Wechsler Intelligence test (Tewes, 1991), was used as a measure for working memory span.
- **Modality Choice.** For the multimodal test block, where participants were free to choose the input modality, modality choice was measured. For every task, the modality used to perform the tasks was annotated for further analysis. Based on this annotation, the percentage of modality usage was computed for each participant.
- **Performance.** Performance measures were recorded online during the interaction. Task duration was measured for each block except for the explorative block, as its duration was fixed to five minutes. Task success was calculated as the percentage of tasks successfully completed in the first attempt.

Procedure

Each session lasted for a total of two hours. Participants were seated in front of a TV with access to the IPTV-service. In order to create a setting as natural as possible, the lab room was furnished along the lines of a standard living room. At first, the participants had to fill in a consent form and received a general instruction. After that, they were asked to complete a questionnaire collecting demographic information, the BFI-S questionnaire, and the questionnaire assessing technical affinity and the mood faces scale. This was followed by an explorative test phase. The participants were asked to explore the application on their own to get acquainted with it. The first exploration phase took 5 minutes. After that, the first evaluative judgments were assessed. Then, the specific instructions for the different input modalities were given. Now participants had to perform several tasks read to them by the experimenter. If a task was not accomplished after three trials, the task was aborted and the next task was presented. The modality (speech, touch or multimodal) which was to be tested next, was explained beforehand. If the participant was able to perform a sample task with this modality and confirmed that she understood everything, the test block was started. After all tasks were completed, the post interaction questionnaires had to be filled in. This was repeated for every modality (speech, touch or multimodal), with the modalities being randomized. Finally, to assess the working memory span, the digit span test had to be performed. An example procedure is presented in Figure 6.13.

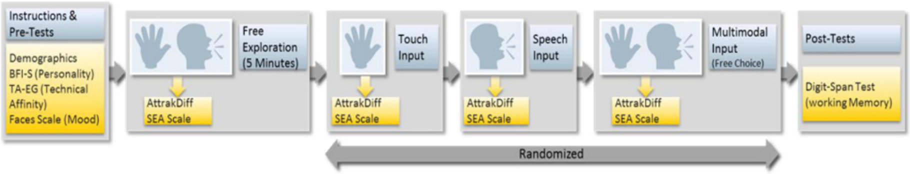


Fig. 6.13. Procedure for one participant with the order touch, speech and multimodal input.

Results

As the primary interest was to investigate the influence of individual characteristics on quality perceptions, and not the differences in quality perceptions between the modalities, all interaction ratings were summed up for each scale over all test blocks. For completeness, the analyses of differences in quality perceptions between the modalities are reported in Appendix A.4.

If not stated otherwise, task duration refers to the summed up task duration of the speech, touch and multimodal block, and task success refers to the mean task success over all trials. Further, if a directional effect (e.g., extraversion increases speech usage) was assumed, the one-tailed p-value was used. If no such assumption was made, the two-tailed p-value was used.

User Characteristics and Modality Selection

To test previously explained assumptions, extraversion and attitudes were correlated with modality usage. Significant correlations were shown for attitudes only. As predicted, negative attitudes towards new technologies lead to a lower percentage of speech interactions, Pearson's r=-.58, N=30, $p_{\text{one-tailed}}$<.01. A positive attitude resulted in a higher percentage of speech usage, Pearson's r=-.33, N=29, $p_{\text{one-tailed}}$=.043.

To examine if this effect was moderated by performance, task success and duration of the speech condition were taken as control variables in a multiple partial correlation (*pr*). The effect remained significant for negative attitudes, *pr* = -.62, $p_{\text{one-tailed}}$ <.01. For positive attitudes the effect vanished, *pr* = .31, $p_{\text{one-tailed}}$ =.065.

In the next step, an explorative analysis was performed trying to figure out which additional user characteristics are of relevance for modality choice. However, none of them showed a significant correlation.

Finally, stepwise multiple linear regression including all assessed characteristics was employed to predict modality choice. The model included negative attitudes, extraversion and technical competence (cf. Table 6.9), thus it complies with our expectations.

User Characteristics and Performance

For performance, working memory span (cognitive abilities) and attitudes were correlated with task success and task duration. No correlations were observed for task duration as well as for working memory span. However, attitudes were related to performance. Positive attitudes were positively correlated with task success, Pearson's r=.43, N=29, $p_{\text{one-tailed}}$=.010. Negative attitudes were negatively correlated with task success, Pearson's r=-.58, N=30, $p_{\text{one-tailed}}$<.01.

Again, an explorative correlation analysis was conducted, including all characteristics. It was shown that age and neuroticism decrease performance while agreeableness, technical competence and technical enthusiasm enhance performance (cf. Table 6.6).

As a direct effect of age seems implausible (cf. Introduction of Section 6.3) multiple partial correlation for age and task success was performed with all other user characteristics as control variables. Nevertheless, the previously significant correlation for task success stayed significant, *pr* = -.55, $p_{\text{two-tailed}}$ =.013.

Stepwise multiple linear regressions using all characteristics as predictors was employed. The models included neuroticism for task duration and negative attitudes, openness, and age for task success (cf. Table 6.9).

Table 6.6. Results of explorative correlations for task success and task duration, N = 30

		Task Success	Task Duration
Age	*r*	-.53	.36
	p	.003**	.054
Neuroticism	*r*	-.54	.58
	p	.002**	.001**
Agreeableness	*r*	.48	-.55
	p	.007**	.002**
Competence	*r*	.50	-.49
	p	.007**	.009**
Enthusiasm	*r*	.43	-.37
	p	.017*	.046*

Note.*$p_{\text{two-tailed}}$<.05, **$p_{\text{two-tailed}}$<.01

User Characteristics and Quality Perception

The measures assessing extraversion, attitudes, and mood were correlated with the interaction ratings. Contrary to the expectations, extraversion showed no correlation with quality perceptions. However, in line with the assumptions, it was shown that positive attitudes towards ICT and mood correlated significantly with the quality ratings. In addition, negative attitudes had a significant influence on both the AttrakDiff's hedonic quality scales. No correlation was observed for the perceived mental effort (SEA-scale). Results of the correlation analyses are presented in Table 6.7.

Table 6.7. Correlation (Pearson's *r*) between attitudes and quality perceptions, and mood and quality perceptions

		Negative Attitudes	Positive Attitudes	Mood
Hedonic Qualities - Stimulation	*r*	-.32	.44	.36
	p	.043*	.008**	.027*
	N	30	29	29
Hedonic Qualities - Identity	*r*	-.42	.43	.40
	p	.011*	.010*	.017*
	N	30	29	29
Attractiveness	*r*	-.23	.41	.32
	p	.114	.013*	.046*
	N	30	29	29
Pragmatic Qualities	*r*	-.29	.56	.38
	p	.059	.001	.021*
	N	30	29	29

Note.*$p_{\text{one-tailed}}$ <.05, **$p_{\text{one-tailed}}$<.01

To control the influence of performance, partial correlations were calculated with overall task duration and overall task success as control variables (cf. Table 6.8). All correlation stayed significant, except for Pragmatic Qualities and mood. Hence, the results indicate that the influence of mood on Pragmatic Qualities is partly moderated by performance. This seems plausible, as Pragmatic Qualities is the scale assessing perceptions of efficiency and effectiveness.

Table 6.8. Partial correlations between attitudes and quality perceptions and mood and quality perceptions controlled for task success and duration

		Negative Attitudes	Positive Attitudes	Mood
Hedonic Qualities-	*pr*	-.34	.46	.40
Stimulation	*p*	.037*	.008**	.020*
Hedonic Qualities-	*pr*	-.48	.44	.39
Identity	*p*	.005**	.010*	.021*
Attractiveness	*pr*	-.31	.48	.35
	p	.052	.006**	.039*
Pragmatic	*pr*	-.29	.57	.32
Qualities-	*p*	.071*	.001**	.054

Note.*$p_{\text{one-tailed}}$ <.05, **$p_{\text{one-tailed}}$<.01

In the explorative analysis including all user characteristics, the only correlation observed was between age and Attractiveness, the Attrakdiff's global scale, Pearson's r = .406, N= 30, $p_{\text{two-tailed}}$= .026. As an influence of the chronological age was not expected, the other user characteristics and the performance measures were used as control variables in a multiple partial correlation. However, the correlation became even higher, which indicates that the relationship of age and rating was not moderated by any of these variables, *pr* = .624, $p_{\text{two-tailed}}$ = .018. Thus, age was related to quality perceptions to a certain extent.

Table 6.9. Results for stepwise multiple regression for perceived quality, modality choice and performance

Measure	Predictors	β	Adj. R^2	*RMSE*
Hedonic Qualities-Stimulation	Positive Attitudes	0.47	0.18	1
Hedonic Qualities-Identity	Negative Attitudes	-0.40	0.37	0.72
	Age	0.47		
	Mood	0.38		
Attractiveness	Positive Attitudes	0.33	0.41	0.71
	Age	0.52		
	Mood	0.36		
Pragmatic Qualities	Positive Attitudes	0.55	0.28	0.75
SEA	Age	-0.59	0.47	12
	Agreeableness	-0.56		
Modality Choice (%Speech)	Negative Attitudes	-0.81	0.58	20.12
	Extraversion	0.31		
	Competence	-0.29		
Task Duration (min:ss)	Neuroticism	0.64	0.39	06:22
Task Success	Negative Attitudes	-0.58	0.56	8.45
	Openness	0.32		
	Age	-0.30		

Intermediate Discussion

The above study investigated if user characteristics can be related to modality selection, quality ratings and performance measures. In particular, it was assumed that modality selection is linked to attitudes and personality, that performance measure are correlated with attitudes and cognitive abilities, and that quality perceptions are related to personality and mood.

The results showed that especially attitudes towards technology affect modality choice, performance, and quality perceptions. This is in line with previous research (Jawahar & Elango 2001). An implication for the strong relation between speech usage (modality selection) and negative attitudes could be, to offer both modalities in parallel if the usage scenario includes a wide variety of users, comprising somewhat technophobic users. Moreover, if the negative attitudes towards technology could be reduced, speech input could become more popular.

Age being associated with performance and quality perceptions, and working memory span not being related to those measures, leads to the conclusion that the assessed parameters of "aging" were not the right ones. This was confirmed by a post-hoc analysis, where no correlation was found between age and the raw score (not corrected for age) of the digit span. Thus, further studies should include other parameters to assess the effect of "cognitive aging".

Personality traits were not relevant for quality judgments but were related to modality selection and interaction behaviour. While the influence of extraversion on speech usage can be attributed to extraverts being more talkative, results for the other traits, neuroticism and openness, are more difficult to interpret. For neuroticism an explanation could be that they are more likely to be affected by the pressure of being tested. Consequently, their performance might decrease. For openness no such explanation can be found.

Also, further studies need to confirm our results and so far our analyses have been largely correlative and are thus not to be interpreted as causal dependencies. However, a general implication of the study is to control for such individual differences in user tests, with attitudes and mood being important for evaluative judgments, and personality traits being important for performance assessment.

6.4 Chapter Discussion and Chapter Summary

The above chapter presents four studies addressing factors influencing modality choice. Results of the studies identified efficiency, situational demands (in terms of allocation of cognitive resources), and user characteristics as relevant factors for modality selection strategies. More specifically, the studies indicate that offering speech as an input modality is especially useful when providing shortcuts or when the visual channel is busy. Or, the other way around, if speech does not offer shortcuts or reduction of mental load, users will probably not interact multimodally. However, even if speech offers shortcuts and reduces mental load, users do not always use it. The latter may be due to user characteristics at least partly; in particular, attitudes towards technology and personality were shown to be related to modality selection. High affinity towards technology and an extraverted, talkative personality increase speech usage.

Furthermore, the assumption of multimodal systems being advantageous over unimodal systems due to their robustness and easier error-recovery by offering alternative input modes was only partly supported: While users do eventually change the modality if they fail, they tend to stay in the same modality up to a certain point and do not employ flexible interaction strategies. This does not mean that effectiveness is of no influence at all. If users actually do switch the modality, this is often related to insufficient effectiveness or error-proneness of the current modality. Moreover, so far these results are limited to the two systems tested. The used speech recognizers had fairly good recognition accuracy (Study 6.1 - 6.3: ~77% /Study 6.4: ~70%). Thus, with less reliable speech recognition error avoidance could have become more important to the participants, leading to a higher influence of effectiveness and robustness on modality selection.

However, multimodal interfaces allow the user to adapt their interaction and modality selection strategies to situational demands: For most of the measures in Study 6.3, the multimodal interface outperformed the unimodal version. Still, this might be due to the design of the study, with the multimodal condition always presented last. Moreover, maybe this effect only became apparent because of the situational demands. This assumption is backed up by previous studies (without secondary tasks, where a general benefit of multimodal interaction could rarely been observed (e.g.: Wechsung, Naumann, & Hurtienne, 2009). Further evidence is provided by Study 6.4, in which speech input led to a better performance, probably due to the TV screen competing with the touch display for visual attentional resources.

Regarding the design of mobile interfaces, the results imply that in visually demanding situations purely GUI-based systems, even if the interaction elements are coded verbally, should be avoided. Sharing attention between controlling a mobile device and responding to external stimuli in the same modality decreases performance and increases mental effort. Furthermore, for tasks that are likely to require longer input, like song names or album titles, a subtle reminder of the speech option (e.g. a microphone icon) might encourage untrained users to switch to this modality.

One important implication of the presented studies concerns evaluations of such systems: Our results implicate, that for (multimodal) systems including speech control, task duration might not be the appropriate parameter for assessing efficiency: An influence on task duration on modality selection was not observed. A second implication is, that laboratory usability tests neglecting contextual factors might not provide reliable results, as at least regarding the perceived pragmatic qualities.

A limitation of all the above studies is that, except for the last study, the users were relatively young. Moreover, the current results should be examined with other tasks (e.g time constraint tasks or less well-defined tasks) and other systems.

Regarding the prediction of modality selection, the studies above are in so far helpful that if all of the identified factors can be controlled or specified, the prediction accuracy, and by this the accuracy in predicting quality ratings too, may increase. However, in practice this will only seldom be the cases. A user study aiming to control for all these aspects will probably be off huge complexity. However, according to the taxonomy presented in Chapter 3, all those factors influence the quality ratings and interaction parameters. Thus, prediction of modality selection may be possible solely based on those two data sources. This assumption is investigated in the next chapter.

7 Is Modality Selection Predictable? - Using Quality Ratings to Predict Modality Selection in Multimodal Systems

According to the studies presented in Chapter 5, the quality ratings for a multimodal system are equal to the weighted sum of the quality ratings of its individual modalities, with the modality that is more frequently used having a stronger influence. These findings suggest that, if the choice of modality can be predicted, an estimation of the quality of the multimodal systems is possible based solely on an evaluation of its component modalities. In Chapter 6 the relative efficiency in terms of interaction steps, situational demands, and user characteristics were identified as factors influencing modality selection. Considering all those factors would result in a highly complex user study.

According to the taxonomy of quality aspects of multimodal human-machine interaction presented in Chapter 3, all of the identified factors also have an influence on the interaction with the system as well as the perceived quality of the system. Hence, it might be possible to predict modality choice based solely on the interaction parameters and quality ratings of the constituent components. To test these assumptions quality ratings and interaction data of three systems were included first in multiple regression analyses and in a second step in path analyses, in order to predict modality choice.

7.1 Study 7.1

7.1.1 Method

Systems and Tasks

Data of three different studies was used. The three different multimodal applications were each installed on a different smartphone. For all systems the available input modalities were speech and touch. Only the input modalities were varied; output was always multimodal and included feedback to the GUI, as well as task-specific auditory output. Apart from the “standard” output given for both modalities, special auditory feedback was provided for speech input for the following system states: ‘recognition active’, ‘match’, ‘no match’.

For the first study, the *Sprachbox* “mailbox” application was used. The *Sprachbox* is a multimodal mailbox system capable of handling speech-, email- and fax-messages, as well as of forwarding calls and mailbox message notification. The application was installed on a HTC Touch Diamond. The speech module used was IBM Embedded Via Voice. Speech recognition was activated via a push-to-talk button on the left hand side of the device. The experimental tasks were: accessing voice messages; retrieving a specific voice message; deleting this message; accessing the email

DOI: 10.1007/978-3-319-03810-0_7, © Springer International Publishing Switzerland 2014

inbox; opening an email; opening the fax inbox, opening a fax; opening the voice messages; sorting them from A to Z; redirecting all calls and confirming this change returning to the menu; and closing the application. For this system, motion control was also implemented but was rarely used and thus excluded from the analysis. More information on the study is provided in Appendix 2.2.

The second study was conducted with a mobile "jukebox" application; detailed information on this study is presented in Section 6.2. The jukebox application was installed on an HTC G1, an Android-based smartphone. The available input modalities were speech and touch. Speech recognition (Nuance Vocon) was activated via a push-to-activate button installed on the back of the device. Participants had to perform the following tasks: opening playlists, opening favorites, opening artist list, opening album list, searching for the song 'first', for the song 'second', for the song 'third', for the song 'fourth', and for the song 'fifth', starting playback, skipping the current song, starting shuffle mode and stopping playback. Apart from the tasks explained above, the participants had to respond to either visually or auditorily demanding stimuli which were presented in a randomized order in time intervals between 3 and 5 seconds.

The third system was a multimodal "remote control" application for an IPTV service, in-depth information on this study is presented in Section 6.3. The application was implemented on an iPhone 3GS. The speech recognition, Nuance Vocon, was activated with a diagonal swiping gesture on the touch screen. Feedback was, depending on the task, either given via the TV or via the iPhone screen. The tasks included in the analysis were zapping, switching channels, increasing and decreasing the volume, setting and resetting time shift, opening tele-text, retrieving a tele-text site, opening and closing the EPG, starting and ending recording, retrieving information for the current program, activating and deactivating mute.

Participants

In all studies, none of the participants had any prior experience with the application and all participants were rewarded with either shopping vouchers or money in cash.

In the "mailbox" study, 23 German-speaking participants aged between 24 and 71 years (*M* = 43 y., *SD* =18 y.) took part.

In the "jukebox" study, 24 German-speaking subjects, aged between 22 and 33 years (*M* = 26 y., *SD* =2 y.), participated.

For the multimodal "remote control" study, 32 German-speaking participants, aged between 18 and 62 years (*M* = 45 y., *SD*=12 y.), were invited. All were owners of the IPTV service tested.

Only complete cases were included in the further analyses. Accordingly the above descriptions do not include data of incomplete cases.

Measures

Quality ratings were collected with the AttrakDiff questionnaire (Hassenzahl, Burmester, & Koller, 2003). In the first study ("mailbox"), the complete 28 items version of the AttrakDiff was used. In the studies "jukebox" and "remote control" the short version, containing 10 items, (Hassenzahl & Monk, 2010) was employed. The perceived mental effort was assessed with the SEA scale.

Regarding interaction parameters, the task aborts (three unsuccessful attempts) and task duration were logged. To assess modality choice, the modality chosen first to solve the task was logged for each task in the multimodal condition. Then the proportion of speech and touch was calculated. As participants could either choose touch or speech, the resulting usage rates were complementary, adding up to 100%. Consequently, for further analysis the proportions of speech only were used. Regression analyses using the touch usage rate show the same results.

Procedure

For all studies, participants had to sign a consent form and were asked to fill in demographic questionnaires. Next, the applications were explained to them. In the unimodal condition, all tasks had to be performed with either touch or speech. In the multimodal condition, participants could choose the input modality freely. The multimodal condition was presented after the unimodal conditions for the studies "mailbox" and "jukebox". In order to avoid learning effects the sequence of the unimodal conditions (speech-touch vs. touch-speech) was altered for each participant in each study. In the study "remote control", a free exploration phase was conducted before the task-based conditions. Moreover, in this study all conditions (except for the exploration phase) were randomized (cf. Figure 6.13).

For each task the participants had three trials, after three trials the task was aborted and the next task had to be carried out. Interaction parameters were logged during the interaction. Quality ratings were assessed after each condition.

7.1.2 Results

In the presentation of results, sub-indices are used to indicate parameters collected in each condition, e.g. HQI_S for the rating on the scale Hedonic Qualities-Identity in the *speech* condition, and HQI_T and HQI_{Mm} for the same ratings in the *touch* and *multimodal* conditions respectively. The scale Hedonic Qualities-Stimulation is abbreviated with *HQS*, the scale Pragmatic Qualities with *PQ* and the Attractiveness scale with *ATT*.

System-Wise Prediction of Modality Choice

Stepwise linear regression analyses were conducted for each system with modality choice as dependent variable. The quality ratings and interaction parameters of the unimodal conditions were used as predictor variables.

For the "jukebox", HQI_S and SEA_S were selected as predictors by the stepwise algorithm. Speech was used more frequently, if the ratings for speech on the scale Hedonic Qualities-Identity were high and the perceived effort of speech was low. The inclusion of the *SEA* scale might be due to the concurrent task the users had to perform. For the "remote control" HQI_S was included in the model (Table 7.1).

For the system "mailbox", no significant predictor was found. In this experiment, the maximum age of the participants was higher compared to the other studies. Since previous research reported that prediction is difficult for older adults (Engelbrecht et al., 2008; cf. Section 5.3), possibly due to age-related decrease in memory capacity, participants older than 55 years were excluded. With only the younger users (N=64), also for the "mailbox" a significant predictor was found. However, in contrast to the other two systems, not HQI_S was included, but ratings on the global scale in the speech condition (ATT_S). Better ratings on ATT_S are related to increased usage of speech.

For the "jukebox" system, the exclusion of older users had no effect on the model, as all participants were younger than 56 years. For the "remote control", the predictors remained the same (only HQI_S), but the accuracy increased considerably. The detailed results of this section are given in Table 7.1.

Table 7.1. Results of stepwise multiple linear regression analyses for each system.

System	Older Users	Predictor	*β*	*F* (*df*)	*p*	*Adj. R²*	*RMSEA*
Mailbox	w	-	-	-	-	-	-
	w/o	ATT_S	.56	5.82 (1,13)	.031	.26	.24
Jukebox	w	SEA_S	-.63	8.17 (2,21)	.002	.38	.15
		HQI_S	.35				
Remote control	w	HQI_S	.47	8.34 (1,30)	.007	.19	.28
	w/o	HQI_S	.63	15.37 (1,23)	.001	.38	.21

Global Prediction of Modality Choice

In the next step, multiple linear regression analysis was performed on the data of all three systems together. If older and younger users were included, HQI_S and HQS_T were significant predictors for modality choice. The better the ratings for speech on the *HQI* scale and the worse the ratings for touch on the *HQS* scale the more likely was the usage of speech in the multimodal condition. If older participants were excluded, *HQI_S* remained in the model, while *HQS_T* was removed. In addition to *HQI_S*, ratings on *PQ_S* and the abort rates in the speech condition (*$Abort_S$*) were chosen by the algorithm. Again, prediction accuracy was higher without the older users (Table 7.2).

Table 7.2. Results of stepwise multiple linear regression analyses over all systems.

Older Users	Predictor	*β*	*F* *(df)*	*p*	Adj. R²	*RMSE*
w	*HQI_S*	.49	7.43 (2,76)	.001	.14	.28
	HQS_T	-.25				
w/o	*PQ_S*	.22	14.68 (3,60)	<.01	.40	.20
	HQI_S	.34				
	$Abort_S$	-.26				

Prediction of multimodal quality ratings based on modality choice predictions

As the prediction performance of all models above increased if older participants were excluded, the following analyses were performed with younger users only. To see if the predicted proportion of modality usage can be used to predict quality ratings in the multimodal condition, the predicted values (*Pr_Use_S* and *Pr_Use_T*) were used as coefficients to the ratings in the single modality conditions (*Q_T* and *Q_S*), as suggested by previous work presented in Chapter 5, leading to the following model:

$$Pr\_Q_{Mm} = Pr\_Use_S \cdot Q_S + Pr\_Use_T \cdot Q_T \tag{7.1}$$

with

Pr_Q_{Mm}: predicted quality rating in the multimodal condition
Pr_Use_S: predicted proportion of speech usage
Q_S: actual quality rating for speech
Pr_Use_T:. predicted proportion of touch usage
Q_T: actual quality rating for touch

Note that *Q* may represent any of the four scales of the *AttrakDiff*.

Based on the regression analysis above (cf. Table 7.2), the usage rates for speech (*Pr_Use_S*) were predicted with the following equation:

$$Pr\_Use_S = PQ_S \cdot .22 + HQI_S \cdot .34 + Abort_S \cdot -.26 \tag{7.2}$$

As touch and speech usage rates were complementary, adding up to 100%, the predicted usage rates for touch (*Pr_Use_T*) were obtained as follows:

$$Pr\_Use_T = 100 - Pr\_Use_S \tag{7.3}$$

The correlation between the actual quality ratings in the multimodal condition (Q_{Mm}) and the quality ratings predicted with Equation 7.1 (Pr_Q_{Mm}) was quite high for all scales (Table 7.3). To check how much information was actually added by the predicted modality usage proportions, another model was built, assuming equal distribution of modality usage. The resulting equation is as follows:

$$Pr\_Q_{Mm} = 0.5 \cdot Q_S + 0.5 \cdot Q_T \tag{7.4}$$

Additionally the assumption of 40-60 speech-touch distribution was tested, as in many of the presented studies, touch was preferred over speech. The resulting equation is as follows:

$$Pr\_Q_{Mm} = 0.4 \cdot Q_S + 0.6 \cdot Q_T \tag{7.5}$$

Table 7.3 shows that those models perform well, too. However, except for *HQS* the correlations were higher for all scales if the predicted modality proportions were used as coefficients, compared to the baseline models.

Paired t-test confirmed that the absolute prediction error was significantly smaller for the scales *HQI* and *PQ* when using the predicted modality proportions (Equation 7.1) compared the baselines (Equation 7.4 & Equation 7.5). For the *ATT* scale, the error was smaller using Equation 7.1, but the difference was not significant.

Table 7.3. Correlation between predicted quality ratings and actual ratings and t-test for absolute errors

			HQI	*HQS*	*ATT*	*PQ*
$Pr\_Q_{Mm} = Pr\_Use_S \cdot Q_S + Pr\_Use_T \cdot Q_T$	Pearson`s *r* (between Pr_Q_{Mm} and Q_{Mm})		.828**	.894**	.857**	.742**
	Abs. Error Pr_Q_{Mm}	*M*	.395	.402	.431	.544
	[Abs(Q_{Mm}-Pr_Q_{Mn})]	(*SD*)	(.41)	(.33)	(.35)	(.51)
$Pr\_Q_{Mm} = 0.5 \cdot Q_S + 0.5 \cdot Q_T$	Pearson`s *r* (between Pr_Q_{Mm} and Q_{Mm})		.727**	.859**	.834**	.716**
	Abs. Error Pr_Q_{Mm}	*M*	.455	.404	.472	.607
	[Abs(Q_{Mm}-Pr_Q_{Mn})]	(*SD*)	(.45)	(.36)	(.45)	(.55)
	t-test	*t*	2.19	.07	.968	1.75
		(*p*)	(.016*)	(.473)	(.169)	(.043*)
$Pr\_Q_{Mm} = 0.4 \cdot Q_S + 0.6 \cdot Q_T$	Pearson`s *r* (between Pr_Q_{Mm} and Q_{Mm})		.802**	.896**	.845**	.734**
	Abs. Error Pr_Q_{Mm}	*M*	.464	.400	.458	.595
	[Abs(Q_{Mm}-Pr_Q_{Mn})]	(*SD*)	(.44)	(.35)	(.40)	(.53)
	t-test	*t*	3.10	.12	.74	1.74
		(*p*)	(.002**)	(.453)	(.232)	(.043*)

Note.*p<.05, **p<.01

7.1.3 Discussion

The study tried to predict modality choice based on quality ratings and performance metrics in order to achieve accurate quality predictions for multimodal systems from data gathered in assessment studies for the component modalities. It was shown that such predictions of modality choice are possible and that with those predictions the prediction accuracy of the quality of multimodal systems is improved. In practical terms, this means that an estimation of the perceived quality of a multimodal system can be obtained from the ratings of their individual components without carrying out expensive multimodal evaluation experiments.

Ratings for touch were seldom included in the models. A possible explanation might be that touch is the “default” modality, and only if speech is perceived as possessing high hedonic qualities the users actually use speech.

It was further observed that the interaction parameter used could not explain modality choice; which is in line with the results of previous studies (Hornbæk & Law, 2007). It has been reported that interaction parameters assessing efficiency and effectiveness do not necessarily reflect the perceived efficiency or effectiveness (Hornbæk & Law, 2007). Since judgments and decisions are made based on the individual perceptions and evaluations, it is reasonable to assume that the perceived efficiency and effectiveness are better predictors for modality choice. However, this is also not consistently the case for the current data: The scale Pragmatic Qualities (*PQ*) measuring quality attributes related to efficiency and effectiveness was less often included in the models than the scale Hedonic Qualities-Identity (*HQI*). This means that modality choice is, aside from the factors explained in Chapter 6, not determined primarily by the modalities’ functional, pragmatic qualities, but by its non-functional, hedonic qualities.

Nevertheless, interaction parameters may have an indirect influence on modality choice moderated by the perceived quality ratings. Such indirect influences cannot be assessed with linear regression; hence, these assumptions will be tested in a next step using path modelling.

7.2 Study 7.2

In the previous study, interaction parameters did not show a significant influence in most of the models. However, linear regression only allows for modelling of direct influences (Miles & Shevlin, 2001). According to the taxonomy explained in Chapter 3, interaction parameters influence quality perceptions and may therefore also indirectly influence modality choice. To test this assumption, path analysis is employed in this section. Path analysis is based on a series of multiple regressions. In contrast to simple multiple regression analysis, path analysis allows to test for multi-level models, which is the case if moderating variables and indirect effects are assumed

(Schendera, 2004). Path analysis is a special case of structural equation modelling, a method which was used in Chapter 4. Unlike structural equation modelling, path analyses only include manifest variables and not latent variables. Manifest variables are variables which can be directly observed and directly measured. Latent variables are hypothesized constructs, which cannot be measured or observed directly. They are operationalized with so-called indicators (Backhaus et al., 2011). Indicators are directly measured and hence manifest variables. Thus path analyses are structural equation models with only directly observable variables. Path analyses were chosen over structural equation models because the sample size did not allow for the latter. While perceived quality may be interpreted as a latent construct, the solutions did not converge for a global model including an additional latent variable "Perceived Quality". Negative variances (Heywood cases) occurred, probably because the low sample size compared to the high number of indicators. Hence, path analyses were calculated for each of the AttrakDiff's sub-scales separately.

7.2.1 Method

The same data as in Study 7.1 was used (cf. Section 7.1.1.). As the models for the older users did perform poorly in Study 7.1 only the 64 younger participants were included.

As the variables differed in terms of the type of scaling used, and as path analysis assumes standardised data (Bortz & Schuster, 2010), Fisher's *Z*-transformation was applied.

7.2.2 Results

Path analyses were conducted using the data from all three systems. The hypothesized structure is displayed in Figure 7.1. A correlation between task duration and abort rates was assumed, as more trials to complete a task automatically lead to longer task duration. Moreover, a correlation between the ratings for the touch and the speech condition was expected, as inter-individual differences regarding response biases are well documented (Van Vaerenbergh & Thomas, 2013); e.g. some persons have a tendency to use extreme ratings while others tend to use the midpoint of a scale. Due to the resulting systematic error, the ratings should show correlations.

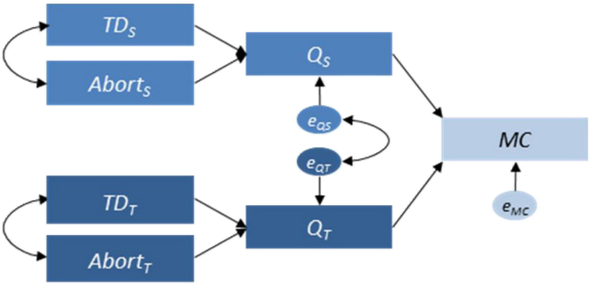
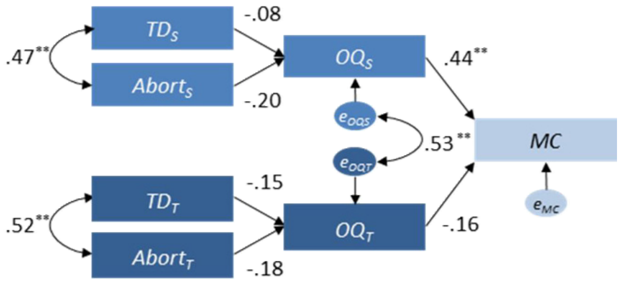


Fig. 7.1 Path diagram of the assumed influences of the interaction parameters *task duration (TD)* and *task abort (Abort)* on *the perceived quality* (*Q*) of speech and touch trials, which influence *modality choice* (*MC*) in the multimodal condition. Ellipses represent the error variables (e.g. measurement error).

In the taxonomy in Chapter 3, the differentiation of quality factors proposed by Hassenzahl (2003a) is adopted. Here, quality factors are divided into hedonic, non-functional aspects and pragmatic, functional aspects. A strong relationship with quality judgments can mostly be expected for the pragmatic, task-relevant aspects, which are also associated with the classical usability concepts efficiency and effectiveness. Hedonic aspects like the innovativeness or aesthetics of a device are not necessarily related to efficiency and effectiveness.

In a first step, the overall score based on the mean of all items of the AttrakDiff (*OQ*) was used. The results imply that modality choice is primarily influenced by the perceived quality of the individual modalities (see Figure 7.2). The interaction parameters showed no significant influence on the perceived overall quality.

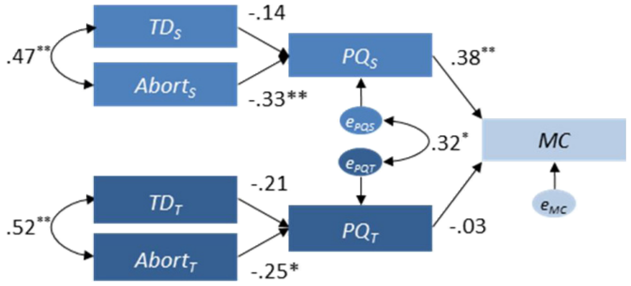


Notes: $\chi^2(12)$= 14.18, *p*=.289, CFI=.970, *RMSEA*=.054, R^2=.15, significant paths/relations are marked as follows: **p*<.05, ***p*<.01

Fig. 7.2. Path diagram of the assumed influences of the interaction parameters *task duration (TD)* and *task abort (Abort)* on *the perceived Overall Quality* (*OQ*) of speech and touch trials, which influence *modality choice* (*MC*) in the multimodal condition. Ellipses represent the error variables (e.g. measurement error).

In the analyses described above, the overall *AttrakDiff (OQ)* rating was used. Although this global score is not affected by the interaction parameters *task duration* and *task aborts*, they may still matter for the pragmatic qualities. Hence, in a further step, the relation between the single quality aspects and the interaction parameters and their influence on modality choice was investigated.

Here, the *task aborts* in the speech condition had a highly significant effect on the perceived *Pragmatic Qualities (PQ)* of speech and thus an indirect effect also on modality choice (see Figure 7.3). Also, the perceived *Pragmatic Qualities (PQ)* of touch was influenced by the abort rates in the touch condition. However, the ratings in the touch condition did not significantly influence modality choice. Prediction accuracy (R^2=.34) was considerably higher compared to the model for overall quality (R^2=.15).

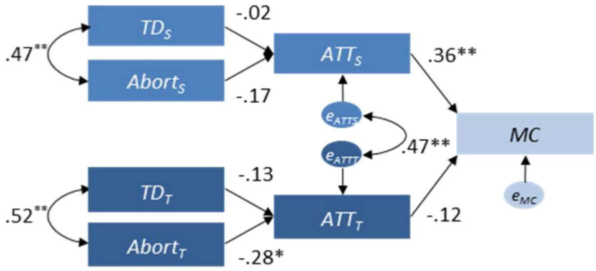


Notes: $\chi^2(12)$= 15.35, *p*=.223, *CFI*=.951, *RMSEA*=.067, R^2=.34, significant path/relations are marked as follows: **p*<.05, ***p*<.01

Fig. 7.3. Path diagram of the assumed influences of the interaction parameters *task duration (TD)* and *task abort (Abort)* on *the perceived Pragmatic Qualities* (*PQ*) of speech and touch trials, which influence *modality choice* (*MC*) in the multimodal condition. Ellipses represent the error variables (e.g. measurement error).

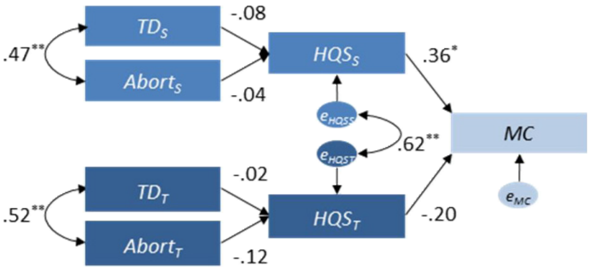
For the scale *Attractiveness (ATT),* which measures pragmatic and hedonic qualities again, the *task aborts* had an influence on the ratings of the touch condition. However, those ratings did not significantly influence modality choice. The interaction parameters in the speech condition did not affect the ratings (see Figure 7.4).

For both of the hedonic scales *(HQS, HQI)* no significant relation between interaction parameters and perceived quality was observed (see Figures 7.5 and 7.6).



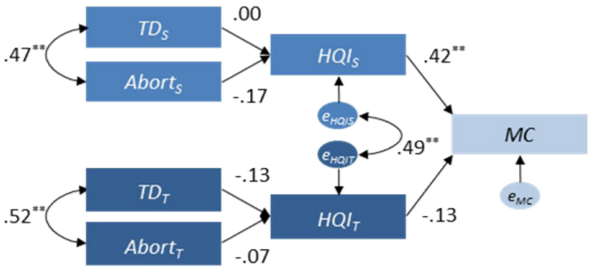
Notes: $\chi^2(12)= 15.04$, $p=.239$, $CFI=.954$, $RMSEA=.063$, $R^2=.11$, significant path/relations are marked as follows: *$p<.05$, **$p<.01$

Fig. 7.4. Path diagram of the assumed influences of the interaction parameters *task duration (TD)* and *task abort (Abort)* on *the perceived Attractiveness* (*ATT*) of speech and touch trials, which influence *modality choice* (*MC*) in the multimodal condition. Ellipses represent the error variables (e.g. measurement error).



Notes: $\chi^2(12)= 16.47$, $p=.171$, $CFI=.939$, $RMSEA=.077$, $R^2=.08$, significant path/relations are marked as follows: *$p<.05$, **$p<.01$

Fig. 7.5. Path diagram of the assumed influences of the interaction parameters *task duration (TD)* and *task abort (Abort)* on *the perceived Hedonic Qualities-Stimulation* (*HQS* of speech and touch trials, which influence *modality choice* (*MC*) in the multimodal condition. Ellipses represent the error variables (e.g. measurement error).



Notes: $\chi^2(12)= 14.41$, $p=.275$, $CFI=.961$, $RMSEA=.056$, $R^2=.14$, significant path/relations are marked as follows: *$p<.05$, **$p<.01$

Fig. 7.6. Path diagram of the assumed influences of the interaction parameters *task duration (TD)* and *task abort (Abort)* on *the perceived Hedonic Qualities-Identity* (*HQI*) of speech and touch trials, which influence *modality choice* (*MC*) in the multimodal condition. Ellipses represent the error variables (e.g. measurement error).

In all analyses, the quality ratings for speech had a stronger influence on modality choice compared to the ratings for touch. Actually, the ratings for touch showed no significant influence on modality choice.

In line with the prior assumptions, highest prediction accuracy was observed for the model for Pragmatic Qualities. This is also the only model where the interaction parameters had a significant indirect influence on modality choice.

7.2.3 Discussion

The above study employs path modelling to investigate direct and indirect effects of interaction parameters on perceived quality and modality choice. It was shown that interaction parameters are mainly related to perceptions of pragmatic qualities, while the perceived hedonic qualities seems to remain unaffected by them. These findings are in line with the underlying conceptualization of these different quality attributes.

It was observed that an increased task abort rate in the speech condition decreases the perceived instrumental qualities (*PQ*) of speech, which decreases the usage rate of speech in the multimodal condition.

Task duration did not influence quality perceptions. Previous results report similar observations. For example, Möller (2006) observed that actual and user judgments of perceived task duration are not correlated.

A further finding is that the correlation between the ratings for touch and the ratings of speech were found to be quite high in all analyses. Such findings indicate that participants whose quality ratings were high for the speech condition also gave high quality ratings for the touch condition and vice versa. These findings indicate that the quality ratings are heavily influenced by individual response styles, apart from the system's actual quality. Hence, care must be taken in when selecting participants (cf. Sec. 6.3).

It should be noted that the sample size in the above study is rather low. Although the number of cases remains large enough for well performing models (according to Iacobucci, 2010), the results still need to be validated with larger samples.

7.3 Chapter Discussion and Chapter Summary

In this chapter, two studies were reported which apply two different methods, stepwise multiple linear regression and path analysis, in order to predict modality choice. First multiple linear regression analyses were employed. Results showed that modality choice is directly influenced by perceptions of a system's hedonic, non-instrumental qualities whereas measures of pragmatic, instrumental qualities were included only in one model. However, the data the study is based on was collected in lab experiments, and it is reasonable to assume, that in the real world (e.g. in a work context), a system's instrumental qualities can be more important. Accordingly, these results have to be verified in more natural usage contexts.

The interaction parameters, which were abort rates as an inverse measure of task success and task duration, were in the most cases not included in the models. This indicates that they do not directly influence modality selection. Thus, path analyses were conducted in a next step.

An indirect effect on modality selection was found for the task abort rates in the speech condition. This effect was mediated through the perceived instrumental qualities (Pragmatic Qualities) of speech. However, the earlier multiple regression analyses showed a similar result: There the only interaction parameter included in the models as a significant predictor, were the abort rates for speech.

As in the previous study, for task duration neither a direct nor an indirect effect was observed. As only the two most salient interaction parameters task duration and abort rate were used, it could be possible that including more detailed parameters like recognition rate (for a detailed listing see Kühnel, 2012) would explain additional variance in the perceived pragmatic qualities and the variance in modality choice.

Furthermore, the systems investigated offered to a large extent sequential input only. It needs to be determined, if predictions as described above are possible also for systems offering extensive parallel input

In summary, the studies showed, that for modality selection the perceived, hedonic qualities are crucial, and that the interaction parameters used are mainly related to perceptions of pragmatic qualities. However, the sample size was, although sufficient, still rather low. Hence, the models may offer a starting point for future investigation of the relationship between interaction parameters and questionnaire data. But they should not necessarily be considered as authoritative.

8 Summary and Outlook

8.1 Summary

This thesis started with an introduction of key concepts of multimodal interaction and a review of currently available evaluation methods (Chapter 2). It was concluded that although multimodal systems have entered the mass market, evaluation methods specifically tailored to multimodal systems are rather rare and that a widely accepted standard method is not available. Hence, the constructs assessed when evaluating multimodal systems are quite diverse and therefore difficult to compare.

As a first step towards a unified evaluation approach, a unified framework, a taxonomy, of quality aspects of multimodal interaction is presented in Chapter 3. Although this framework provides a holistic view integrating experience-related and performance-related aspects of multimodal interaction, only the assessment of the latter aspects have so far been validated for multimodal systems (Kühnel, 2012). For the assessment of the experience-related aspects, mainly methods are available which have been developed for unimodal (predominantly GUI-based) systems.

The most widespread evaluation method regarding experience-related factors are questionnaires, consequently four well-known and popular questionnaires, which were initially developed for unimodal systems, were investigated concerning their appropriateness for the evaluation of multimodal systems (Chapter 4). Of the four questionnaires (SUS, SUMI, SASSI, AttrakDiff) included in this study, the AttrakDiff showed the most promising results. Therefore, it was chosen as a starting point for the development of a new questionnaire, the MMQQ, which is specifically tailored to multimodal systems. The theoretical ground of the MMQQ is the taxonomy of quality aspects of multimodal systems suggested by Möller and colleagues (2009). In parallel to the questionnaire development, Möller's taxonomy was empirically validated and altered accordingly. This validation was achieved with the employment of confirmatory modelling approaches in addition to the exploratory approaches, which are usually employed in questionnaire development.

While the MMQQ is especially useful for the evaluation of complete multimodal systems, such global evaluations tell little about the relationship of a system's individual modalities. Consequently, in Chapter 5 it was investigated how quality ratings of single modalities relate to the global evaluation of multimodal systems. The main intention was to examine if the quality perceptions of a multimodal systems can be predicted based on the quality perceptions of its constituent modalities. It was shown, that especially for overall scales, measuring both pragmatic and hedonic qualities, a rough estimation of the quality of multimodal systems is possible (based on the quality ratings of the single modalities). Moreover, modality usage rates were observed to be central for such predictions. The more frequently a modality was used the higher was the modality's influence on the quality perceptions of the multimodal system.

Ina Wechsung, *An Evaluation Framework for Multimodal Interaction*,
T-Labs Series in Telecommunication Services,
DOI: 10.1007/978-3-319-03810-0_8, © Springer International Publishing Switzerland 2014

Due to this observed importance of modality usage rates, the factors, which influence those rates, were addressed in Chapter 6. In four empirical studies, efficiency, situational demands related to the allocation of cognitive resources, and user characteristics were identified as factors, which are relevant for modality choice.

An experiment in which all of these factors are controlled is, due to the resulting high complexity, difficult to realize. Therefore, Chapter 7 investigates if modality selection is predictable based on quality ratings and interaction parameters. According to the taxonomy presented in Chapter 3 this should be possible: In the taxonomy it is assumed that all the factors, identified as relevant for modality selection, also influence quality ratings and interaction parameters. The results reported in Chapter 7 are partly in line with this assumption: Modality choice was found to be directly influenced by perceptions of a system's hedonic qualities. While the influence of pragmatic qualities was not as prominent, it could further be shown that interaction parameters influence a system's perceived pragmatic quality. Moreover, predictions of the quality of multimodal systems based on the ratings of its individual modalities are more accurate, if the predicted modality usage rates are used as weights in the regression equation, compared to baselines assuming 50/50 or 60/40 usage distributions.

In summary, this thesis presents (1) an exhaustive and empirically validated taxonomy of quality aspects of multimodal interaction as well as respective measurements methods, (2) a validated questionnaire specifically tailored to the evaluation of multimodal systems and covering most of the taxonomy's quality aspects, (3) insights on how the quality perceptions of multimodal systems relate to the quality perceptions of its individual components, (4) a set of empirically tested factors which influence modality choice, and (5) models regarding the relationship of the perceived quality of a modality and the actual usage of a modality.

8.2 Discussion and Future Work

In this section, an outlook on future research directions is presented. While very specific implications of the presented research have already been discussed in the previous chapters, the current section takes a broader perspective aiming to integrate the thesis' main results with recent and emerging trends.

8.2.1 From the Lab to the Field

One avenue for future work concerns the degree to which the results presented in this thesis are generalizable. All studies presented were conducted in a laboratory setting and were for the most part task-oriented. In addition to this, the systems predominantly offered sequential input (where the input modalities were speech and touch) while the output was mostly multimodal.

Although laboratory settings provide a controlled environment, which makes it possible to clearly identify the effects of specific factors, such settings are (by definition) rather artificial. In some of the studies steps were taken to mitigate against this, by providing more naturalistic test environments (as in Studies 6.3, 6.4) and by using commercial devices throughout (although the applications being tested were specially designed prototypes). Hence, aspects such as response time or the touch screen's capabilities (e.g. its recognition accuracy) can be said to be realistic. However, none of the studies was conducted in a truly natural setting. The tasks to be performed by the participants were pre-defined. In all studies, only tasks which were supported by the system were selected. This is of course different to how users interact with systems in a real-world setting, as discovering the capabilities of a system is an essential aspect of the experience of being confronted with a new device (Cordes, 2001). Moreover, the "confounding" variables, which are usually controlled in a lab setting may have a significant impact on a system's usability. Therefore, it may be the case that a system or a modality is received well in a lab test but will fail in the real world. In terms of multimodal interaction, for instance, the speech modality may not be robust enough in noisy environments such as bars and restaurants, or perhaps the touch modality may not work as intended if the user is wearing gloves.

The issues, reported above, show, that although laboratory studies, which aim to eliminate contextual factors strictly, may be appropriate for performance evaluation and have yielded good results from which meaningful insights may be inferred, additional insights into and knowledge of the multi-faceted concept of 'experience' may be gained by extending the studies into wider environments. Therefore, a general next step should be to move from lab studies to field studies, in order to investigate if the results are generalizable and can be confirmed in a wider setting.

8.2.2 From Tasks to Challenges

For systems, which are "per se" not task-oriented but used rather solely for entertainment, additional quality aspects, which are not mentioned in the taxonomy, may be relevant, or the relations between the quality aspects may differ to those reported in this thesis. Dyck et al. (2003) even claimed that games "were 'separated at birth' from most of the accepted paradigms for designing usable interactive software". In contrast to other interactive system, the focus in game development was mainly on novelty and innovation (Dyck et al., 2003). However, the purpose of such new interaction techniques is to be able to play "in more efficient and more interesting ways" (Dyck et al., 2003). Hence, it is reasonable to assume that such systems also need to possess pragmatic qualities apart from hedonic qualities, like discoverability or novelty. For example, if a user wants to perform a certain action in a computer game (such as moving their avatar from A to B) it ought to be possible to carry out this action in an efficient and effective manner. Similarly, loading times should be short.

However, for higher level goals in a game (perhaps finding a treasure, for example), it may be desirable to have diverting detours, involving riddles, puzzles or challenges. Such detours could be considered “inefficient” and “ineffective” in the classical sense of ease-of-use, but are nonetheless integral to the game. On this level, the optimal task duration as a measure of efficiency may follow a U-shape rather than follow the common “the shorter the better” formula.

An example of a product which goes against traditional interface norms is Facebook. According to Hart and colleagues (2008), Facebook barely complies with the classical ease-of-use-focused usability heuristics suggested by Nielsen (1994). Nevertheless, Facebook is a major success. Accordingly, while both “serious” systems and "entertainment" systems should, in the best case, demonstrate both, high hedonic as well as high pragmatic qualities, the importance of the respective quality dimensions may differ.

On the other hand, it should be noted that systems used mainly in a work context should also be motivating and pleasurable, something which can be achieved through so-called gamification. Gamification adapts highly motivating elements of games into non-gaming contexts, e.g. serious software (Detering et al., 2011). The aim is to improve the motivation of the user to interact with the system (Hassenzahl, 2003b). While the term gamification is relatively new, the idea itself is not. An early example presented by Laschke and Hassenzahl (2003), is the PSDoom process manager, where the ego-shooter Doom is adapted and processes are “killed”. Today big companies such as SAP employ similar, game-based strategies in order to motivate their employees (Schacht & Schacht, 2012). In addition, such gamification strategies may be helpful in encouraging sustainable behaviour (e.g. with an in-car game) and health promoting behaviour. Another popular application area is marketing. Many companies use gamifictation approaches, e.g. loyal customers can achieve certain levels, which come along with certain discounts. Here the aim is of course to sell products by employing so–called extrinsic motivators. Unfortunately, extrinsic motivators have been shown to reduce the “free-choice” intrinsic motivation (Deci, Koestner, & Ryan, 2001). Hence, care has to be taken, when employing such strategies. Moreover, the effects of the (extrinsic) rewards are not sustainable. This means, as soon as the rewards are taken away, their effects on the user behaviour will disappear. For example, once the user has all “badges” he may stop using the system. In summary, bringing the alienated siblings “games” and “serious software” together seems like a worthwhile approach to make non-gaming interactive systems more engaging; but further research is necessary, in order to determine which aspects can provide (sustainable) motivation in the context of serious systems.

8.2.3 From Tool to Partner

A further area for future work concerns how systems can be endowed with a personality. This is particularly important for "serious" spoken dialogue systems and avatars, where the system's persona can be a crucial factor, which determines its success. To return to the example provided in the introduction to this thesis, the sassy tone of Apple's Siri system is considered to be one of the reasons for its success. However, such an informal tone would most likely be inappropriate for a telephone banking system. Thus, attempts to incorporate personality into a system raises several questions: Which personality type is suitable for a particular system or user group? How best to design and develop a certain personality using synthesised speech? What are the features of speech (including prosodic elements such as pitch, tone) which are relevant for personality perceptions? What are the appropriate lexical and semantic choices for dialogue responses? Some research in this direction has already been conducted (e.g. Polzehl, Möller, & Metze, 2011; Mairesse & Walker, 2011) but it is still a rather new direction.

For multimodal systems, the question then arises whether all modalities ought to be aligned with the targeted personality, in order for the system to have a consistent personality throughout. In the case of Siri, this was probably not a focus of the iPhone/iPad developers. Consequently, Siri and the iPhone may be perceived as different entities and not as one system. It should be noted however, that many apps and indeed PC programs are also perceived as separate entities that are distinct from their associated device or operating system. What is noteworthy in the case of Siri is that it is a native feature of the iPhone and comes automatically bundled with the device when it is sold, and yet the perception of it as being somehow separate persists, despite its tight integration with other native iPhone apps (such as email or the calendar). Consequently, an additional area for future work to examine is how systems, whose constituent parts are perceived as different "entities" but yet are tightly integrated with each other, can be evaluated.

Moreover, possibly intensified by its anthropomorphic features, a certain intelligence is attributed to Siri. This was also the case for an app called AskWiki, which was developed in at Deutsche Telekom Innovation Laboratories (Burkhardt & Zhou, 2012; Wechsung et al. 2012b). AskWiki allowed users to query Wikipedia using speech. During its development, user research was conducted, which revealed that users often blamed the app for failures, which were in fact due to Wikipedia and not due to the app itself. It may be that for such "intelligent" systems additional quality aspects may become relevant.

8.2.4 From Explicit to Implicit Modalities

Most systems investigated in this thesis offered sequential touch and speech input, a set-up which is arguably the de-facto standard among today's mobile devices. However, information and communication technology is a highly dynamic field, and motion-controlled devices are already widely available. Moreover, modern sensors permit body movements to be tracked (e.g. lip-reading, eye/gaze–tracking, head tracking), as well as making it relatively easy for biometric data to be assessed (e.g. face recognition/detection, fingerprint recognition, motion profiles and even for neuronal activity). Such "sensor"-based input differs from speech and touch input (although touch input is sensor-based itself) as it is more implicit, which means that users are not necessarily aware of the sensors being part of the system. For example, a system might first use face recognition before activating speech recognition. For a smooth and well performing system, a novice user may not recognize this underlying process. Often such sensors can be used to adapt the system to the user, without the user realizing this. One consequence of this is, that it is reasonable to assume that such sensor-based input modalities are only noticed if they fail. This is a fundamental difference between such implicit modalities when compared to the rather explicit modalities investigated in this thesis. The evaluation of such implicit modalities seems to be particularly difficult; for example, how can an entity be judged, if the judge is in the best case not aware of the entity? Whether the taxonomy presented in Chapter 3 and the Multimodal Quality Questionnaire (Chapter 5) can be applied to such systems needs to be determined. For example, aspects such as the level of control a user has concerning the adaptation may be important.

Furthermore, the relationship between the implicit and explicit modalities or indeed the relationship between different implicit modalities regarding the systems overall quality may be different from what was observed in the studies presented in this thesis. Moreover, modality choice is, for such systems, not necessarily a conscious decision made by the user, as such sensors are often not "actively" addressed by the user. In the above example, if face recognition is used as an activation trigger for speech recognition, a novice user might assume that the speech recognition is not working and might not necessarily "blame" the face recognition.

Apart from the modalities being explicit, there was also an inherent and necessary redundancy present. This means with few exceptions all tasks could be performed with both modalities. In addition, only sequential input was offered. Investigating systems that offer complementary, parallel input such as in Bolt's Put-that-there-demonstrator (1980) were beyond the scope of this thesis. Verifying that the results presented also hold for such interfaces should also be examined in future work.

8.2.5 From Short-Term to Long-Term Studies

A final issue which should be considered in future research, is that of long-term usage. All studies were, by necessity, short-term studies. Evaluations of the experienced quality were always assessed directly after the interaction. Hence it was not possible for a study to assess either the experienced quality during the interaction, or indeed the long-term experienced quality as recalled at a later stage (i.e. after some days). In additional research not reported this thesis (Wechsung et al. 2012b), ratings collected "online" during a field trial were compared with ratings assessed after the usage period. While quantitative ratings of overall quality were similar, the qualitative data differed: comments collected during usage were more specific with respect to certain negative or positive aspects of the apps performance. Participants often only reported problems, not judgments. Comments, collected after usage, were often rather general; however, they also contained an affective appraisal of the experience. Thus, the remembered experience does not necessarily represent a one-to-one reflection of the actual experience. These preliminary findings suggest that investigations sustained over longer time-frames are an area worthy of further study.

In addition to the concerns regarding quality perceptions, it is probably also the case that modality usage patterns are dynamic and changing over time. It is imaginable that an exciting new modality may become boring after some time, or that usage of it increases only after a period of familiarisation.

In summary, to tackle the questions raised above, a next step could be to move from the lab to the field and extent the time-frame of the user studies. While the systems used in this thesis, are representative of the mass market of such systems, they will certainly be superseded by devices with extended capabilities (e.g. systems offering "true" conversational competence and parallel input). Such new technologies will not only engage the future users, but hopefully also the future researchers. On the other hand, the "Put-that-there-demonstrator" is over 30 years old (Bolt, 1980), but still such "truly" multimodal devices are rather rare. Accordingly, while the reported findings may partly be limited to current devices, they will remain valid and useful for a while. Together with the results of Kühnel (2012), a holistic approach for the evaluation of current and future multimodal devices is provided by the presented work.

References

Ajzen, I., & Fishbein, M. (1980). *Understanding Attitudes and Predicting Social Behavior*. Englewood Cliffs, NJ: Prentice Hall.

Althoff, F., McGlaun, G., Lang, M., Rigoll, G. (2003). Evaluating multimodal interaction patterns in various application scenarios, In *Proceedings of the 5th International Gesture Workshop (GW 2003)*, 421–435.

Amelang, M., Bartussek, D., Stemmler, G. & Hagemann, D. (2006). *Differentielle Psychologie und Persönlichkeitsforschung.* [Differential psychology and personality research]. Stuttgart: Kohlhammer.

Anderson, J. R., & Lebiere, C. (1998). *Atomic Components of Thought.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Andrews, F.M., & Withey, S.B. (1976). *Social indicators of well-being*. New York: Plenum.

Angel, A., Hartmann, J., & Sutcliffe ,A. (2009). The effect of brand on the evaluation of websites. *Proceedings of 12th IFIP TC13 Conference on Human-Computer Interaction (INTERACT '09)*, 638-651

Armitage, C. J., Conner, M., & Norman, P. (1999). Differential effects of mood on information processing: evidence from the theories of reasoned action and planned behaviour. E*uropean Journal of Social Psychology, 29*, 419-433.

Armitage, C. J., & Deeprose, C. (2004). Changing student evaluations by means of the numeric values of rating scales. *Psychology Learning and Teaching 3*(2), 122-125

Backhaus, K., Erichson, B., Plinke, W. & Weiber, R. (2011). *Multivariate Analysemethoden*. [Multivariate analysis methods]. Berlin: Springer.

Baddeley, A.D. (1992). Working Memory. *Science 255*(5044), 556-559.

Baddeley, A.D. (2003). Working memory: Looking back and looking forward. *Nature Reviews:Neuroscience 4*(10), 829-83

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation*. New York: Academic Press, pp. 47-90.

Baillie, L., Schatz, R., Simon, R., Wegscheider, F., Anegg, H., Pucher, M., Niklfeld, G., & Gassner, A. (2005): Designing Mona: User Interactions with a Multimodal Mobile Game. Paper presented at the *11th International Conference on Human-Computer Interaction (HCII 2005)*. Retrieved from http://mona.ftw.at/papers/designingmona_final.pdf

Balbo, S., Coutaz, J., & Salber, D. (1993). Towards automatic evaluation of multimodal user interfaces. *Proceedings of the 1st International Conference of Intelligent User Interfaces (IUI 1993)*, 201-208.

Baber, C. (2001). Computing in a Multimodal World. In *Proceedings of the 9th International Conference on Human-Computer Interaction (HCI International 2001),* 232-236.

Bargas-Avila J.A., & Hornbæk K. (2011). Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2011)*, 2689-2698.

Barkhuus, L. & Rode, J.A. (2007). From mice to men – 24 years of evaluation in CHI. In *Extended Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2007).*

Barrett, J. & Jiang, Y. (2012). *Apple iPhone Siri Users.* Dallas, TX: Parks Associates.

Ina Wechsung, *An Evaluation Framework for Multimodal Interaction*,
T-Labs Series in Telecommunication Services,
DOI: 10.1007/978-3-319-03810-0, © Springer International Publishing Switzerland 2014

Bartneck, C., & Hu, J. (2009). Scientometric Analysis Of The CHI Proceedings. *Proceedings of the Conference on Human Factors in Computing Systems (CHI 2009)*, 699-708.

Bauckhage, C., Fritsch, J., Rohlfing, K., Wachsmuth, S., & Sagerer, G. (2002). Evaluating Integrated Speech- and Image Understanding. In *Proceedings of the 4th IEEE International Conference on Multimodal interfaces (ICMI 2002)*, 9-14.

Beringer N., Kartal, U., Libossek M., & Steininger, S. (2002). *Gestaltung der End-to- End Evaluation in SmartKom 2.0 [Design of the end-to-end-evaluation in SmartKom 2.0]*, Technical Report NR-19. Retrieved from http://www.phonetik.uni-muenchen.de/forschung/SmartKom/TechDok-NR-19.ps

Bernsen, N. O. (1997). Defining a taxonomy of output modalities from an HCI perspective. *Computer Standards and Interfaces, 18,* 537-553.

Bernsen, N.O. (2008). Multimodality theory. In D. Tzovaras (Ed.), *Multimodal user interfaces. From signals to interaction, Signals and communication technology.* Berlin: Springer, pp. 5-29.

Bernsen, N.O., & Dybkjaer, L. (2004). Evaluation of Spoken Multimodal Conversation. In *Proceedings of the Sixth International Conference on Multimodal Interfaces (ICMI 2004)*, 38-45.

Bevan, N. (2009). What is the difference between the purpose of usability and user experience evaluation methods. Paper presented at the *Workshop User Experience Evaluation Methods in Product Development (UXEM'09)*. Retrieved from http://www.nigelbevan.com/papers/What_is_the_difference_between_usability_and_user_experience_evaluation_methods.pdf

Bilici, V., Krahmer, E., Riele, S., & Veldhuis, R. (2000). Preferred modalities in dialogue systems. In *Proceedings of 6th International Conference on Spoken Language Processing (ICSLP 2000)*, 727–730.

Billi, R., Castagneri, G., & Danieli, M. (1997). Field trial evaluations of two different information inquiry systems. *Speech Communication*, *23*(1–2), 83-93.

Bless, H., Schwarz, N., Clore, G.L., Golisano, V., Rabe, C., & Wölk, M. (1996). Mood and the use of scripts: Does being in a happy mood really lead to mindlessness? *Journal of Personality and Social Psychology, 71,* 665-679.

Bolt, R.A. (1980). "Put-that-there": Voice and gesture at the graphics interface. In *Proceedings of the 7th annual conference on Computer graphics and interactive techniques (SIGGRAPH '80)*, 262-270.

Borg, G. (1982). Psychophysical Bases of Perceived Exertion, *Medicine and Science in Sports and Exercise 14*, 377–381.

Borkenau, P., & Ostendorf, F. (1994). *Das NEO Fünf-Faktoren-Inventar (NEO-FFI): Handanweisung* [The NEO Five-Factor-Inventory: Manual]. Göttingen: Hogrefe.

Bornträger, C., Cheverst, K., Davies, N., Dix, A., Friday, A., & Seitz, J.: Experiments with multi-modal interfaces in a context-aware city guide. In *Proceedings of the 5th International Symposium on Human Computer Interaction with Mobile Devices and Services Mobile (HCI 2003)*, 116-130.

Boros, M., Eckert, W., Gallwitz, F., Gorz, G., Hanrieder, G., & Niemann, H. (1996). Towards understanding spontaneous speech: word accuracy vs. concept accuracy. In *Proceedings of the Fourth International Conference on Spoken Language (ICSLP 1996)*, Vol. 2, 1009-1012.

Bortz, J. & Döring, N. (2007). *Forschungsmethoden und Evaluation* [Research methods and evaluation]. Berlin, New York: Springer.

Bortz, J., & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler.* [Statistics for human and social scienists]. Heidelberg: Springer.

Bradley, M.M., & Lang, P.J. (1994). Measuring emotion: the self-assessment Manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry 25*(1), 49-59.

Brave, S., & Nass, C. (2007). Emotion in human-computer interaction. In Sears, A. & Jacko, J. (Eds.), *The Human-Computer Interaction Handbook.* Mahwah, NJ: Lawrence Erlbaum Associates, pp. 77-92

Brooke, J. (1996). SUS: A "quick and dirty" usability scale. In P. Jordan, B. Thomas, B. Weerdmeester, & I. McClelland (Eds.), *Usability Evaluation in Industry.* London: Taylor & Francis, pp. 189-194.

Brown, R.M., Hall, L.R., Holtzer, R., Brown, S.L., & Brown, N.L. (1997). Gender and video game performance. *Sex Roles, 36* (11-12), 793-812.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* [Introduction to test and questionnaire construction]. München: Pearson.

Burke, J. L., Prewett, M. S., Gray, A. A., Yang, L., Stilson, .F. R., Coovert, M. D., Elliot L. R., & Redden, E. (2006). Comparing the effects of visual-auditory and visual-tactile feedback on user performance: a meta-analysis. In *Proceedings of the 8th international Conference on Multimodal interfaces (ICMI '06)*, 108-117.

Burkhardt, F., & Zhou, J. (2012). "AskWiki": Shallow Semantic Processing to Query Wikipedia. In *Proceedings of the 20th European Signal Processing Conference (EUSIPCO 2012)*, 350-354.

Burmester, M., Mast, M., Jäger, K., & Homans, H. (2010).Valence method for formative evaluation of user experience, In *Proceedings of Designing Interactive Systems Conference (DIS '10)*, 364–367, .

Burnett, G.E., & Ditsikas, D. (2006). Personality as a criterion for selecting usability testing participants. *Proceedings of the IEEE 4th International conference on Information and Communications Technologie (ICICT),* 487-498.

Buttle, F. (1996). SERVQUAL: review, critique, research agenda. *European Journal of Marketing 30*(1), 8–31.

Card, S., Moran, T., & Newell, A. (1983). *The Psychology of Human-Computer Interaction.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Carmichael, A. (1999). *Style guide for the design of interactive television services for elderly viewers.*Winchester, UK: Independent Television Commission.

Chalmers, P.A. (2003). The role of cognitive theory in human-computer interface. *Computers in Human Behavior 19*(5), 593-607.

Charwat, H. J. (1992). *Lexikon der Mensch-Maschine-Kommunikation.* [Encyclopedia of Human.-Machine-Communication]. München: Oldenbourg.

Chen, C., Czerwinski, M., & Macredie, R. (2000). Individual differences in virtual environments. *Journal of the American Society for Information Science and Technology 51*(6), 499-507.

Chen, X., & Tremaine, M. (2006). Patterns of multimodal input usage in non-visual information navigation. In *Proceedings of the 39th Hawaii International Conference on Systems Sciences (HICSS' 2006)*, 123c-133c.

Cohen, P., McGee, D., & Clow, J. (2000). The efficiency of multimodal interaction for a map-based task, In *Proceedings of the 6th Conference on Applied Natural Language Processing (ACL' 2000)*, 331–338.

Cordes, E.R. (2001). Task-selection bias: a case for user-defined tasks. *International Journal of Human–Computer Interaction 13*(4), 411–420

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI): Professional Manual.* Odessa, FL: Psychological Assessment Resources.

Cox, A.L., Cairns, P.A., Walton, A., & Lee, S. (2008). Tlk or txt? Using voice input for SMS compo-sition. *Personal Ubiquitous Computing 12*(8), 567-588.

Damianos, L., Drury, J., Fanderclai, T., Hirschman, L., Kurtz, J., & Oshika, B. (2000). Evaluating Multi-party Multi-modal Systems. In Proceedings of the 2nd International Language Resources and Evaluation Conference (LREC 2000), Athens, June 2000.

Davis, F.D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly 13*(3), 318-340.

DeCarlo, L.T. (1997). On the Meaning and Use of Kurtosis. *Psychological Methods 2*(3), 292-307.

DeAngeli, A., Gerbino, W., Petrelli, D. & Cassano, G. (1998). Visual Display, Pointing and Natural Language: The Power of Multimodal Interaction. In *Proceedings of the Visual Conference on Advanced Visual Interfaces (AVI' 98)*, 164 - 173

Deci, E. L., Koestner, R., & Ryan, R. M. (2001). Extrinsic rewards and intrinsic motivation in education: Reconsidered once again. *Review of Educational Research 71*(1), 1-27.

De Waard, D. (1996). *The Measurement of Drivers' Mental Workload*, PhD thesis, University of Groningen, Haren.

Desmet, P. (2004). Measuring emotions: development and application of an instrument to measure emotional responses to products. In M. Blythe, C. Overbeeke, A. F. Monk, & P. C. Wright (Eds.), *Funology: From Usability to Enjoyment.*Dordrecht: Kluwer Academic Publishers, pp. 111–123.

Desmet, P., & Hekkert, P. (2007). Framework of product experience, International. *Journal of Design 1*(1). 57-66.

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. E. (2011). From game design elements to gamefulness: defining "gamification". In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments (MindTrek '11*, 9-15.

Dillon, A., & Watson, C. (1996). User analysis in HCI -the historical lessons from individual differences research. I*nternational Journal of Human-Computer Studies 45*(6), 619-637.

Dix, A., Finlay, J., Abowd, G. and Beale R. (1993). *Human-Computer Interaction*. New York: Prentice Hall.

Dixon J. A., Deets, J. K., & Bangert, A. (2001). The representation of the arithmetic operations include functional relationships. *Memory & Cognition 29*(3), 462-477.

Doolittle, P.E., Terry, K.P., & Mariano, G.J. (2009). The effects of working memory capacity on learning and performance in multimedia learning environments. In R. Zheng (Ed.), *Cognitive effects of multimedia learning.* Hershey, PA: Idea Group Reference, pp. 17-33.

Dyck, J., Pinelle, D., Brown, B., & Gutwin, C. (2003).Learning from games: HCI design innovations in entertainment
software. In *Proceedings of Graphics Interface Conference (GI 2003)*, 105-112.

Eilers, K., Nachreiner, F., & Hänecke, K. 81986) Entwicklung und Überprüfung einer Skala zur Erfassung subjektiv erlebter Antrengung [Development and evaluation of a scale to assess subjectively perceived effort]. *Zeitschrift für Arbeitswissenschaft 40*, 215–224.

Elting, C., Zwickel, J., & Malaka, R. (2002). Device-dependant modality selection for user-interfaces: an empirical study. In *Proceedings of the 7th International Conference on Intelligent User Interfaces (IUI 2002)*, 55-62,

Engelbrecht, K.-P. (2012). *Estimating Spoken Dialog System Quality with User Models*. Berlin: Springer.

Engelbrecht, K.-P., Möller, S., Schleicher, R. & Wechsung, I. (2008). Analysis of PARADISE Models for Individual Users of a Spoken Dialog System. In *Proceedings of 19. Konferenz Elektronische Sprachsignalverarbeitung (ESSV 2008)*, 86-93.

Engelbrecht, K.-P., Quade, M. & Möller, S. (2009). Analysis of a New Simulation Approach to Dialogue System Evaluation. *Speech Communication 51*(12), 1234-1252.

Epstein, S. (1994). Integration of the cognitive and psychodynamic unconscious. *American Psychologist*, *49*, 709–724.

Ferris, T., & Sarter, N. (2010). When content matters: The role of processing code in tactile display design. *IEEE: Transactions on Haptics, 3*(3), 199-210.

Fisher, A.G. (2004). *Assessment of Motor and Process Skills (AMPS). Vol. 2: User Manual.*Fort Collins, CO: Three Star Press.

Forlizzi, J., & Battarbee, K. (2004). Understanding experience in interactive systems. In *Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques (DIS '04)*, 261-268.

Fornell, C., Larcker, D.F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research 18*(1), 39-50.

Frøkjær, E., Hertzum, M., & Hornbæk, K. (2000). Measuring usability: are effectiveness, efficiency, and satisfaction really correlated?. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems (CHI '00)*, 345-352.

Gannon, K.M., & Ostrom, T.M. (1996). How meaning is given to rating scales: The effects of response language on category activation. *Journal of Experimental Social Psychology 32*(4), 337-360

Gediga, G., Hamborg, K., & Düntsch, I. (1999). The IsoMetrics Usability Inventory: An operationalisation of ISO 9241-10. *Behaviour and Information Technology 18*(3), 151 - 164.

Geller, T. (2012). Talking to machines. *Communications of the ACM 55*(4), 14-16.

Gerbino, E., Baggia, P., Ciaramella, A., & Rullent, C. (1993) Test and evaluation of a spoken dialogue system. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-93)*,Vol. 2, pp. 135-138.

Gerhard, M. (2003). *A Hybrid Avatar/Agent Model for Educational Collaborative Virtual Environments.* Ph.D. Thesis, Leeds Metropolitan University, UK.

Gibbon, D., Moore, R.K., & Winski, R. (1998). *Handbook of Standards and Resources for Spoken Language Systems. Vol. IV. Spoken Language Reference Materials.* Berlin, New York: Mouton de Gruyter.

Glass, J., Polifroni, J., Seneft, S., & Zue, V. (2000). Data collection and performance evaluation of spoken dialogue systems: The MIT experience. In *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP 200)*, 1-4.

Gong, L. (2003). Multimodal interactions on mobile devices and users' behavioral and attitudinal preferences. In *Proceedings of the 10th International Conference on Human-Computer Interaction (HCII 2003)*,Vol. 4, 1402-1406.

Greenberg, S., & Buxton, B. (2008). Usability Evaluation Considered Harmful (Some of the Time). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2008)*, 111-120.

Grudin, J. (1992). Utility and Usability: Research Issues and Development Contexts. I*nteracting with Computers 4*(2), 209-217.

Hackman, G. S., & Biers, D. W. (1992). Team usability testing: Are two heads better than one?. In Proceedings of the 36 the Annual Meeting of the Human Factors Society, 1205-1209.

Hajdinjak, M., & Mihelic F. (2006). The PARADISE Evaluation Framework: Issues and Findings. *Computational Linguistics 32*(2), 263-272.

Hart, J., Ridley, C., Taher, F., Sas, C., & Dix, A. (2008). Exploring the Facebook experience. A new approach to usability. In *Proceedings of the 5th Nordic conference on Human-computer interaction: building bridges (NordiCHI '08)*, 471-474.

Hart, S.G., & Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research, In P. Hancock and N. Meshkati (Eds.), *Human Mental Workload*, Amsterdam: North Holland, pp. 139–183.

Hassenzahl, M. (2003a). The thing and I: Understanding the relationship between user and product. In M. Blythe, C. Overbeeke, A. F. Monk, & P. C. Wright (Eds.), *Funology: From Usability to Enjoyment.*Dordrecht: Kluwer Academic Publishers, pp. 287-302.

Hassenzahl, M. (2003b). Attraktive Software. Was Gestalter von Computerspielen lernen können. [Attractive software. What designers can learn from computer games]. In Machate, J., & Burmester, M. (Eds.), User Interface Tuning.Benutzungsschnittstellen menschlich gestalten. Frankfurt: Software&Support, pp. 27-45.

Hassenzahl, M. (2008). User experience (UX): towards an experiential perspective on product quality. In *Proceedings of the 20th International Conference of the Association Francophone d'Interaction Homme-Machine (IHM '08)*, 11-15.

Hassenzahl, M. (2010). *Experience Design: Technology for All the Right Reasons*. Princeton, NJ: Morgan & Claypool.

Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität [AttracDiff: A questionnaire to measure perceived hedonic and pragmatic quality]. In J. Ziegler, & G. Szwillus (Eds.), *Mensch & Computer 2003. Interaktion in Bewegung*. Stuttgart, Leipzig: B.G. Teubner, pp. 187-196.

Hassenzahl, M., Diefenbach, S., & Göritz, A.S. (2010). Needs, affect, and interactive products - Facets of user experience, *Interacting with Computers, 22*(5), 353-362.

Hassenzahl, M., Kekez, R., & Burmester, M. (2002). The importance of a softwares pragmatic quality depends on usage modes. In *Proceedings of the 6th international conference on Work With Display Units (WWDU)*, 275-276.

Hassenzahl, M. & Monk, A. (2010). The inference of perceived usability from beauty. *Human-Computer Interaction 25*(3), 235-260.

Hassenzahl, M., Platz, A., Burmester, M. & Lehner, K. (2000). Hedonic and ergonomic quality aspects determine a software's appeal. In *Proceedings of the Conference on Human Factors in Computing (CHI 2000)*, 201-208.

Hassenzahl, M. and Roto, V. (2007). Being and doing: A perspective on User Experience and its measurement. *Interfaces, 72*: 10-12.

Hassenzahl, M., & Sandweg, N. (2004). From mental effort to perceived usability: transforming experiences into summary assessments. In *Proceedings of the Conference on Human Factors in Computing. Extended abstracts (CHI 2004)*, 1283-1286.

Hassenzahl, M., & Tractinsky, N. (2006). User Experience - a research agenda. *Behavior & Information Technology 25*(2), 91-97.

Hassenzahl, M., & Trautmann, T. (2001). Analysis of web sites with the repertory grid technique. In *Proceedings of Conference on Human Factors in Computing Systems. Extended Abstracts (CHI 2001*, 167-168.

Hassenzahl, M., & Ullrich, D. (2007). To do or not to do: Differences in user experience and retrospective judgments depending on the presence or absence of instrumental goals. *Interacting with Computers 19*(4), 429-437.

Hassenzahl, M., & Wessler, R. (2000). Capturing design space from a user perspective: The repertory grid technique revisited. *International Journal of Human-Computer Interaction 12*(3&4), 441-459.

Hearst, M.A. (2011). 'Natural' Search User Interfaces. *Communications of the ACM 54*(11), 60-67.

Hedicke, V. (2000). Multimodalität in Mensch-Maschine-Schnittstellen [Multimodality in Human-Machine-Interfaces]. In K.-P. Timpe, T. Jürgensohn, & H. Kolrep (Eds.), Mensch-Maschine-Systemtechnik. Düsseldorf: Symposion publishing pp. 203-230.

Hegner, M. (2003). *Methoden zur Evaluation von Software,* IZ-Arbeitsbericht Nr. 29.Bonn: Informationszentrum Sozialwissenschaften.

Hekkert, P. (2006). Design Aesthetics: Principles of Pleasure in Product Design. *Psychology Science 48*(2), 157-172.

Hemsen, H. (2004). Evaluation of a Multimodal Dialogue System for Small-screen Devices. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, 1703-1706.

Herzberg, F. (1968). One more time: how do you motivate employees?. *Harvard Business Review 46*(1), 53-62.

Herzog, G., & Reithinger, N. (2006). The SmartKom Architecture: A Framework for Multimodal Dialogue Systems. In: W. Wahlster (Ed.), *SmartKom: Foundations of Multimodal Dialogue Systems,* Berlin: Springer, 2006, pp. 55-70.

Hjalmarsson, A. (2002). *Evaluating AdApt, a Multi-Modal Conversational, Dialogue System Using PARADISE.* Master's thesis, KTH Royal Institute of Technology, Stockholm, Sweden.

Höllerer, S. (2002). Challenges And Important Aspects In Planning And Performing Evaluation Studies For Multimodal Dialogue Systems. In *Proccedings of the the Third International Conference on Language Resources and Evaluation (LREC 2002)*, pp. 1-6.

Holzinger, A. (2005). Usability engineering methodsfor software developers. *Communications of the ACM, 48*(1) , 71-74.

Homburg, Ch., & Giering, A. (1996). Konzeptualisierung und Operationalisierung komplexer Konstrukte - Ein Leitfaden für die Marketingforschung [Conceptualisation and operationalisation of complex constructs - a guideline for market research.], *Marketing – Zeitschrift für Forschung und Praxis 18*(1), 5-24.

Hone, K.S. , & Graham,R. (2000). Towards a Tool for the Subjective Assessment of Speech System Interfaces (SASSI), *Natural Language Engineering 6*(3/4), 287–303.

Hornbæk, K., & Law, E.L.(2007). Meta-analysis of correlations among usability measures. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems (CHI '07)*, 617-626.

Hovy, E. & Arens, Y. (1990). When is a picture worth a thousand words? Allocation of modalities in multimedia communication. Paper presented at the *AAAI Symposium on Human-Computer Interfaces*, Stanford University.

Hu, L.T., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Coventional criteria versus new alternatives. *Structural Equation Modeling 6*(1), 1-55.

Huisman G, Van Hout M (2008) The development of a graphical emotion measurement instrument using caricatured expressions: the LEMtool. In *Proceedings of the Workshop Emotion in HCI - Designing for People,* 5–7.

Iacobucci, D. (2010). Structural equations modeling: Fit indices, sample size, and advanced topics. *Journal of Consumer Psychology 20,* 90–98.

ISO 9241-11 (1998). *Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs). Part 11: Guidance on Usability*. Geneva: International Organization for Standardization (ISO).

ISO 9241-210 (2010). *Ergonomics of human system interaction—part 210: human-centred design for interactive systems (formerly known as 13407).*Geneva: International Organization for Standardization (ISO).

ITU-T Rec. P.851 (2003). *Subjective Quality Evaluation of Telephone Services Based on Spoken Dialogue Systems*, Geneva: International Telecommunication Union (ITU).

Jahn, G., Oehme, A., Krems, J.F., & Gelau, C. (2005). Peripheral detection as a workload measure in driving: Effects of traffic complexity and route guidance system use in a driving study. *Transportation Research Part F 8,* 255- 275.

Jawahar, I.M., & Elango, B. (2001). The effect of attitudes, goal setting and self-efficacy on end user performance. *Journal of End User Computing 13*(2), 40-45.

Jekosch, U. (2000). *Sprache hören und beurteilen: Sprachqualitätsbeurteilung als Forschungs- und Dienstleistungsaufgabe* [Speech listening and judging: Speech quality evaluation as a research and service task]. Habilitation thesis, University Essen.

John, B.E., & Kieras, D.E. (1996). Using GOMS for user interface design and evaluation: which technique?. *ACM Transactions on Computer-Human Interaction 3*(4), 287–319.

John, O.P., Donahue, E. M., & Kentle, R. L. (1991). *The Big Five Inventory-Versions 4a and 54.* Berkeley, CA: University of California, Berkeley, Institute of Personality and Social Research.

Jokinen, K. (2008). User interaction in mobile navigation applications. In L. Meng, A. Zipf and S. Winter (Eds.), *Map-Based Mobile Services: Design, Interaction and Usability*. Berlin: Springer, pp. 168-197.

Jokinen, K. (2009). Natural Language and Dialogue Interfaces. In C. Stephanidis (Ed.), *The Universal Access Handbook.* Boca-Raton, FL: CRC Press Taylor & Francis Group. pp. 495-506.

Jokinen, K., & Hurtig, T. (2006). User expectations and real experience on a multimodal interactive system. In *Proceedings of the Ninth International Conference on Spoken Language Processing (Interspeech 2006)*, 1049-1052 .

Jokinen, K., & Raike, A. (2003). Multimodality – Technology, Visions and Demands for the Future. Proceedings of the 1st Nordic Symposium on Multimodal Interfaces, 239–251.

Jordan, P., (2000). *Designing Pleasurable Products. An Introduction to the New Human Factors.* London, New York: Taylor & Francis.

Kahneman, D. (1999). Objective happiness. In D. Kahneman, E. Diener and N. Schwarz (Eds.), *Well-being:*
Foundations of hedonic psychology. New York: Russell Sage Foundation Press, 3-25.

Kahneman, D. (2003). Maps of Bounded Rationality: Psychology for Behavioral Economics. *American Economic Review, 93*(5), 1449-1475,

Kahneman, D., & Frederick, S.(2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, and D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment.* New York : Cambridge University Press, pp. 49-81.

Kahneman, D., Fredrickson, B.L., Schreiber, C.A., & Redelmeier, D.A.(1993). When more pain is preferred to less: Adding a better end. *Psychological Science*, *4*(6), 401-405.

Kahneman, D., & Miller, D.T. (1986).Norm theory: comparing reality to its alternatives. *Psychological Review (93)*, 136-53.

Kamii, C., Lewis, B.A., & Kirkland, L.D. (2001). Fluency in subtraction compared with addition. J*ournal of Mathematical Behavior 20*(1), 33-42.

Kamm, C., Litman, D., & Walker, M. A. (1998) From novice to expert: The effect of tutorials on user expertise with spoken dialogue systems. *Proc. International Conference on Spoken Language Processing (ICSLP98)*, 1211-1214.

Kamvar, M., & Beeferman, D. (2010). Say what? why users choose to speak their web queries, In *Proceedings of 11th Annual Conference of the International Speech Communication Association (Interspeech 2010)*, 1966–1969.

Kaplan, K. J. (1972). On the ambivalence–indifference problem in attitude theory and measurement: A suggested modification of the semantic differential technique. *Psychological Bulletin 77*(5), 361–372.

Kaptelinin, V., Nardi, B., Bødker, S., Carroll, J., Hollan, J., Hutchins, E., &Winograd, T. (2003). Post-cognitivist HCI: second-wave theories. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI '03)*, 692-693.

Karapanos, E., Martens, J.-B., & Hassenzahl, M. (2012). Reconstructing experiences with iScale. *International Journal of Human-Computer Studies, 70*(11), 849–865.

Karrer, K., Glaser, C., Clemens, C., & Bruder, C. (2009). Technikaffinität erfassen – der Fragebogen TA-EG [Assessing technical affinity - the questionnaire TA-EG]. In A. Lichtenstein, C. Stößel and C. Clemens (Eds.), *Der Mensch im Mittelpunkt technischer Systeme.8. Berliner Werkstatt Mensch-Maschine-Systeme.* Düsseldorf: VDI Verlag GmbH. pp. 196-201.

Kelly, G.A. (1955). *The Psychology of Personal Constructs.* New York: Norton.

Kieras, D.E. (2003). Model-based evaluation. In Sears, A. & Jacko, J. (Eds.), *The Human-Computer Interaction Handbook*. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 1139-1151.

Kieras, D. & Polson, P. (1985). An approach to the formal analysis of user complexity. *International Journal of Man-Machine Studies 22*(4), 365-394.

Kieras, D.E., & Meyer, D.E. (1997). An overview of the EPIC architecture for cognition and performance with application to human–computer interaction. *Human-Computer Interaction 12*(4), 391–438

Kincaid, J. (2010). Google: 25% Of Queries From Android 2.0 Devices Use Voice Search. *TechCrunch*. Retrieved from http://techcrunch.com/2010/08/12/googles-hugo-barra-25-of-android-queries-are-voice-based/.

Kirakowski, J. Corbett M. (1993). SUMI: The Software Usability Measurement Inventory. *British Journal of Educational Technology 24*(3), 210–212.

Kolenikov, S., & Bollen, K.A. (2012). Testing Negative Error Variances: Is a Heywood Case a Symptom of Misspecification?. *Sociological Methods and Research 41*(1), 124-167.

Krämer, N.C. & Nitschke, J. (2002). Ausgabemodalitäten im Vergleich: Verändern sie das Eingabeverhalten der Benutzer? [Output modalities by comparison: Do they change the input behaviour of users?] In R. Marzi, V. Karavezyris, H.-H. Erbe & K.-P. Timpe (Eds..), *Bedienen und Verstehen. 4. Berliner Werkstatt Mensch-Maschine-Systeme.* Düsseldorf: VDI-Verlag, pp. 231-248.

Krosnick, J.A., & Fabrigar, L.R. (1997). Designing rating scales for effective measurement in surveys. In L. Lyberg, P. Biemer, M. Collins, E. deLeeuw, C. Dippo, N. Schwarz, & D.Trewin (Eds.), *Survey Measurement and Process Quality*. New York: Wiley, pp. 141-164.

Kühnel, C. (2012). *Quantifying Quality Aspects of Multimodal Interactive Systems* (T-Labs Series in Telecommunication Services), Berlin: Springer.

Kujala, S., Roto, V., Vaananen-Vainio-Mattila, K., Karapanos, E., Sinnela, A. (2011). UX Curve: A Method for Evaluating Long-Term User Experience. *Interacting with Computers 23*(5), 473-483.

Laakkonen, M. (2007). *Learnability makes things click – A grounded theory approach to the software product evaluation*. Rovaniemi: Lapland University Press.

Lai, J., Mitchell, S., Pavlovski ,C. (2007). Examining modality usage in a conversational multimodal application for mobile e-mail access. *International Journal of Speech Technology 10*(1), 17–30.

Lamel, L., Bennacef, S., Gauvain, J.L., Dartigues, H., & Temem, J.N. (1998). User evaluation of the MASK kiosk. In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP'98)*, 2875–2878

Landauer, T.K. (1995). *The trouble with computers: Usefulness, usability and productivity*. Cambridge, MA: MIT Press.

Larsen, J.T., McGraw, A.P., & Cacioppo, J.T. (2001). Can people feel happy and sad at the same time?. *Journal of Personality and Social Psychology 81*(4), 684-696.

Larsen, L. B. (2003a). Assessment of spoken dialogue system usability - what are we really measuring? In *Proceedings of the 8th European conference on speech communication and technology (Eurospeech 2003)*, 1945-1948.

Larsen, L.B. (2003b). On the Usability of Spoken Dialogue Systems. Ph.D. Thesis, Aalborg University, Denmark.

Laschke, M., & Hassenzahl, M. (2011). Mayor or patron? The difference between a badge and a meaningful story. Paper presented at the *CHI 2011 Workshop "Gamification: Using Game Design Elements in Non-Gaming Contexts"*. Retrieved from http://gamification-research.org/wp-content/uploads/2011/04/18-Laschke.pdf

Laszlo, J.I., & Bairstow, P.J. (1983). Kinaesthesis: Its measurement, training and relationship to motor control. *Quarterly Journal of Experimental Psychology 35A (2)*, 411-421.

Lavie, T., & Tractinsky, N.(2004). Assessing dimensions of perceived visual aesthetics of web sites. *International Journal of Human-Computer Studies 60*(3), pp. 269–298.

Law, E. L., Roto, V., Hassenzahl, M., Vermeeren, A. P., & Kort, J. (2009). Understanding, scoping and defining user experience: a survey approach. In *Proceedings of the 27th international Conference on Human Factors in Computing Systems (CHI 2009)*, 719-728.

Law, E., Roto, V., Vermeeren, A.P.O.S., Kort, J., & Hassenzahl, M. (2008). Towards a shared definition of user experience. In *Proceedings of Conference on Human Factors in Computing Systems. Extented Abstracts (CHI 2008)*, 2395-2398.

Lazonder, A. W., Biemans, H. J. A., & Wopereis, I. G. J. H. (2000). Differences between novice and experienced users in
searching information on the World Wide Web. J*ournal of the American Society for Information Science 51*(6), 576-581.

Lee, L., Amir, O, & Ariely, D. (2009). In Search of Homo Economicus: Cognitive Noise and the Role of Emotion in Preference Consistency. *Journal of Consumer Research, 36,* 173-187.

Lemmelä, S., Vetek, A., Mäkelä, K., & Trendafilov, D. (2008). Designing and evaluating multimodal interaction for mobile contexts, In *Proceedings of the 10th international conference on multimodal interfaces (ICMI 08)*, pp 265–272.

Lewis, J.R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human–Computer Interaction 7*(1), 57–78.

Lewis, T., Langdon, P.M., & Clarkson, P.J. (2008). Prior experience of domestic microwave cooker interfaces: A user study. In P.M. Langdon, P.J. Clarkson, & P. Robinson (Eds.), *Designing Inclusive Futures*, London: Springer, pp. 95-106.

Lin, H. X., Choong, Y., & Salvendy, G. (1997). A proposed index of usability: a method for comparing the relative usability of different software systems. *Behavior and Information Technology 16*(4/5), 267-278.

Lindgaard, G. (1994). *Usability Testing and System Evaluation: A Guide for Designing Useful Computer Systems*. London: Chapman and Hall.

Loewenstein, G. & O'Donoghue, T. (2004). *Animal Spirits: Affective and Deliberative Processes in Economic Behavior.* Working Papers 04-14. Cornell University.

Lopez Cozar, R. & Araki, M. (2005). *Spoken, multilingual and multimodal dialogue systems.* New York:Wiley.

Mahlke, S. & Lindgaard, G. (2007). Emotional experiences and quality perceptions of interactive products. In Proceeding
HCI'07 Proceedings of the 12th International Conference on Human-Computer Interaction: Interaction Design and Usability (HCI'07), 164–173.

Mahlke, S., & Minge, M. (2006). Emotions and EMG measures of facial muscles in interactive contexts. Paper presented at the *CHI 2006 Workshop "HCI and the Face"*. Retrieved from http://www.bartneck.de/workshop/chi2006/papers/mahlke_hcif06.pdf.

Mairesse, F. ,& Walker, M.A. (2011). Controlling User Perceptions of Linguistic Style: Trainable Generation of Personality Traits. *Computational Linguistics 37*(3), 455-488.

Matilla, M., Karjaluoto, H. , Pento, T. (2003). Internet banking adoption amongmature customers: early majority or laggards? *Journal of Services Marketing, 17(5)*, 514-528.

Mayer, J.D., & Gaschke, Y.N. (1988). The experience and meta-experience of mood. *Journal of Personality and Social Psychology, 55,* 102-111.

Mayer, R.E., & Moreno, R. (1998). A split-attention effect inmultimedia learning: evidence for dual processing systems in working memory. *Journal of Educational Psychology 90*(2), 312-320.

Mayhew, D. J. (1999). *The usability engineering lifecycle*. San Francisco: Morgan Kaufmann.

McCarthy, J., & Wright, P. (2004). Technology as experience. *interactions 11*(5), 42-43.

Metze, F., Wechsung, I., Schaffer, S., Seebode, J., & Möller, S. (2009). Reliable Evaluation of Multimodal Dialogue Systems. In *Proceedings of the 13th International Conference on Human-Computer Interaction. Part II: Novel Interaction Methods and Techniques (HCII 2009)*, 75-83

Miles, J.N.V. & Shevlin, M.E. (2001). *Applying regression and correlation: a guide for students and researchers.* London: Sage Publications.

Möller, S. (2005). *Quality of Telephone-Based Spoken Dialogue Systems.* New York: Springer.

Möller, S. (2010). *Quality Engineering*. Heidelberg: Springer.

Möller, S. (2006). Messung und Vorhersage der Effizienz bei der Interaktion mit Sprachdialogdiensten [Measurement and prediction of interaction efficiency with spoken dialogue systems]. In S. Langer & W. Scholl (Eds.), *Fortschritte der Akustik - DAGA 2006.* pp. 463-464.

Möller, S., Engelbrecht, K.-P., Kühnel, C., Wechsung, I., & Weiss, B. (2009). *A Taxonomy of Quality of Service and Quality of Experience of Multimodal Human-Machine Interaction.* In Proceedings of the First International Workshop on Quality of Multimedia Experience (QoMEX'09), 7-12 .

Möller, S., Engelbrecht, K-P., Schleicher, R. (2008). Predicting the Quality and Usability of Spoken Dialogue Services. *Speech Communication 50*(8-9), 730-744.

Mohs, C., Hurtienne, J. , Kindsmüller, M.C. , Israel, J.H. , Meyer, H.A., & IUUI Research Group (2006). IUUI – Intuitive Use of User Interfaces: Auf dem Weg zu einer wissenschaftlichen Basis für das Schlagwort "Intuitivität" [IUUI – Intuitive Use of User Interfaces: On the way towards a scientific basis for the buzzword "intuitivity"]. *MMI-Interaktiv* 11, 75-84.

Moosbrugger, H., & Kelava, A. (2007). *Testtheorie und Fragebogenkonstruktion* [Test theory and questionnaire construction]. Heidelberg: Springer.

Moreno, R. & Mayer, R.E. (1999). Cognitive principles of multimedia learning: The role of modality and contiguity. *Journal of Educational Psychology 91*(2), 358-368.

Morgan, K., Morris, R., Macleod, H., Gibbs, S. (1994). The Possible Roles of Gender and Cognitive Style in the Design of Human-Centred Computer Interfaces, In F. Schmid, S. Evans, A.W.S. Ainger and R.J. Grieve (Eds.), *Computer Integrated Production Systems and Organizations*, Berlin: Springer, pp.110-112.

Morris, W.N. (1989). *Mood. The frame of mind*. New York: Springer.

Naumann, A., Hurtienne, J., Israel, J. H., Mohs, C., Kindsmüller, M. C., Meyer, H. A., & Hußlein, S. (2007). Intuitive Use of User Interfaces: Defining a vague concept. In *Proceeding of the 7th international conference on Engineering psychology and cognitive ergonomics* (EPCE'07), 128-136.

Naumann, A.B., & Wechsung, I. (2008). 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)*, 8-12.

Naumann, A., Wechsung, I., & Hurtienne, J. (2010). Multimodal Interaction: A Suitable Strategy for Including older Users? *Interacting with Computers 22*(6), pp. 465-474.

Naumann, A., Wechsung, I. & Möller, S. (2008). Factors Influencing Modality Choice in Multimodal Applications. In *Proccedings of the 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems (PIT 2008)*. 37–43.

Nielsen, J. (1994). *Usability Engineering.* San Diego: Academic Press.

Nielsen, J., & Levy, J. (1994). Measuring usability: Preference vs. performance. *Communications of the ACM,37*(4), 66-75.

Nigay, L. & Coutaz, J. (1993). A design space for multimodal systems – concurrent processing and data fusion. In *Proceedings of Conference on Human Factors in Computing Systems (INTERCHI `93)*, 172-178.

Nigay, L., & Coutaz, J. (1995). Multifeature systems: The CARE properties and their impact on software design. In J. Lee (Ed.), *Intelligence and multimodality in multimedia interfaces: Research and applications.* Menlo Park: AAAI Press.

Norman, D.A. (1986). Cognitive engineering. In: D.A. Norman and S.W. Draper (Eds.), *User Centered System Design: New Perspectives on Human-Computer Interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 31-61.

Norman, D.A. (1988). *The psychology of everyday things.* New York, NY: Basic Books.

Norman, D.A. (2004). *Emotional design: Why we love (or hate) everyday things*. New York: Basic Books.

Norman, D.A., Miller, J., & Henderson, A. (1995). What you see, some of what's in the future, and how we go about doing it: HI at Apple Computer. In *Proceedings of the Conference Companion on Human Factors in Computing Systems (CHI 1995)*, 155.

Nyeck, S., Morales, M., Ladhari, R., & Pons, F. (2002). 10 years of service quality measurement: reviewing the use of the SERVQUAL instrument. *Cuadernos de Diffusion 7*(13), 101-107.

Oviatt, S. (1996). Multimodal interfaces for dynamic interactive maps. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Common Ground (CHI '96)*, 95-102.

Oviatt, S. (1999). Ten myths of multimodal interaction. *Communications of the ACM, 42*(11), 74-81.

Oviatt, S. (2002). Multimodal Interfaces. In J.A. Jacko and A. Sears (Eds.), *The Human-Computer Interaction Handbook.* Mahwah, NJ: Lawrence Erlbaum Associates, pp. 286-304.

Oviatt, S., Cohen, P., Wu, L., Vergo, J., Duncan, L., Suhm, B., Bers, J., Holzman, T., Winograd, T., Landay, J., Larson, J., & Ferro, D. (2000). Designing the user interface for multimodal speech and pen-based gesture applications: state-of-the-art systems and future research directions. *Human-Computer Interaction, 15*(4), 263-322.

Oviatt, S., Coulston, R., & Lunsford, R. (2004). When do we interact multimodally? Cognitive load and multimodal communication patterns. In *Proceedings of the 6th international Conference on Multimodal interfaces (ICMI '04)*, 129-136.

Paivio, A, (1986). *Mental representations: a dual coding approach.* Oxford: Oxford University Press.

Parasuraman, A., Zeithaml, V.A., &Berry, L.L. (1988). SERVQUAL: a multi-item scale for measuring consumer perceptions of the service quality. *Journal of Retailing 64*(1), 2-40.

Peissner, M., Sell, D., & Steimel, B. (2006). *Akzeptanz von Sprachapplikationen in Deutschland 2006.* Stuttgart Fraunhofer IAO / Initiative Voice Business.

Perakakis, M., & Potamianos, A. (2008). Multimodal system evaluation using modality efficiency and synergy metrics. In *Proceedings of the 10th international conference on Multimodal interfaces* (ICMI '08), pp. 9-16.

Petersen, M.G.(1998). Towards Usability Evaluation of Multimedia Applications. *XRDS: Crossroads 4*(4), 3–7.

Petrelli, D., De Angeli, A., Gerbino, W., & Cassano, G. (1997). Referring in multimodal systems: the importance of user expertise and system features. In *Proceedings of the Workshop on Referring Phenomena in a Multimedia Context and Their Computational Treatment*, 14-19.

Picard, R.W. (1997). *Affective Computing*. Cambridge, MA: MIT Press.

Polzehl, T., Möller, S., & Metze, F. (2011). Modeling Speaker Personality Using Voice. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011),* 2369-2372.

Preece, J., Rogers, Y., Sharp, H., Benyon, D., Holland, S. & Carey, T. (1994). *Human-Computer Interaction: Concepts And Design.* Wokingham, UK: Addison-Wesley.

Prewett, M.S., Yang, L., Stilson, F.R., Gray, A.A., Coovert, M.D., Burke, J., Redden, E., & Elliot, L.R. (2006). The benefits of multimodal information: a meta-analysis comparing visual and visual-tactile feedback, In *Proceedings of the 8th International Conference on Multimodal interfaces (ICMI 06)*, 333–338.

Prümper, J.(1997). Der Benutzungsfragebogen ISONORM 9241/10: Ergebnisse zur Reliabilität und Validität. [The Usability Questionnaire ISONORM 9241/10: Results of Reliability and Validity] In R. Liskowsky, B. Velichkovsky, & W. Wünschmann (Eds.), *Software-Ergonomie '97: Usability Engineering - Integration von Mensch-Computer-Interaktion und Software-Entwicklung.* Stuttgart: Teubner, pp. 254-262

Qvarfordt, P. (2004). *Eyes on Multimodal Interaction.* Ph.D. thesis, Linköping University, Sweden.

Qvarfordt, P., Jönsson, A., & Dahlbäck, N. (2003). The role of spoken feedback in experiencing multimodal interfaces as human-like. In *Proceedings of the Fifth International Conference on Multimodal Interaction, (ICMI'03)*, pp. 250-257.

Raisamo, R. (1999). *Multimodal Human-Computer Interaction: a constructive and empirical study*, Ph.D. Thesis, University of Tampere, Tampere, Finland.

Rammstedt, B. & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality, 41,* 203-212.

Redick, T.S., Calvo, A., Gay, C.E., & Engle, R.W. (2011). Working memory capacity and go/no-go task performance: Selective effects of updating, maintenance, and inhibition. *Journal of Experimental Psychology: Learning, Memory, and Cognition 37*(2), 308-324.

Reinecke, J. (2005). *Strukturgleichungsmodelle in den Sozialwissenschaften* [Structural equation models in the social sciences.] München: Oldenbourg.

Richter, T., Naumann, J., & Groeben, N. (2000). Attitudes toward the computer: Construct validation of an instrument with scales differentiated by content, *Computers in Human Behavior,* 16, 473–491.

Roto, V., Obrist, M., &Väänänen-Vainio-Mattila, K. (2009). User Experience Evaluation Methods in Academic and Industrial Contexts. Paper presented at the *Workshop User Experience Evaluation Methods in Product Development (UXEM'09)*. Retrieved from http://www.cs.tut.fi/~kaisavvm/UXEM09-Interact_ObristRotoVVM.pdf

Rudnicky, A.I. (1993). Mode preference in a simple data-retrieval task. In *Proceedings of the Workshop on Human Language Technology (HLT 1993)*, 364–369.

Ruge, M. (2011). *Stimmungen und Erwartungen im System der Märkte : eine Analyse mit DPLS-Modellen* [Sentiments and Expectations in the systems of the markets: an analysis with DPLS-models]. Potsdam : Universitätsverlag Potsdam.

Russell, J.A., & Carroll, J.M. (1999a). On the bipolarity of positive and negative affect. *Psychological Bulletin 125*(1), 3-30.

Russell, J.A., & Carroll, J.M. (1999b). The Phoenix of Bipolarity: Reply to Watson and Tellegen. *Psychological Bulletin 125(5)*, 611-617.

Sarter, N.B. (2006). Multimodal information presentation: Design guidance and research challenges.*International Journal of Industrial Ergonomics 36*(5), 439-445.

Sarter, N.B. (2007). Multiple-resource theory as a basis for multimodal interface design: Success stories, qualifications, and research needs. In A. F. Kramer, D. A. Wiegmann, & A. Kirlik (Eds.), *Attention: From theory to practic*. Oxford: Oxford University Press, pp. 187-195.

Sauro, J. & Dumas, J. (2009). Comparison of Three One-Question, Post-Task Usability Questionnaires. In *Proceedings of Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI 2009)*, 1599-1608

Sauro, J. and Kindlund, E. (2005). A method to standardize usability metrics into a single score. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '05)*, 401-409.

Schacht, M., & Schacht, S. (2012). Start the Game: Increasing User Experience of Enterprise Systems Following a Gamification Mechanism. In A. Maedche, A. Botzenhardt, L. Neer (Eds.), *Software for People*. Berlin: Springer, pp. 181-199.

Schacter, D.L. (2001). *The seven sins of memory: How the mind forgets and remembers*. Boston, MA: Houghton Mifflin.

Schendera, C.F.G. (2004). *Datenmanagement und Datenanalyse mit dem SAS System: Vom Einsteiger zum Profi (SAS 8.2)* [Data management and data analysis with the SAS system: from novice to expert]. München: Oldenbourg.

Schimmack, U., Boeckenholt, U., & Reisenzein, R. (2002).
Response styles in affect ratings: Making a mountain out of a molehill. *Journal of Personality Assessment 78*(3), 461–483

Schleicher, R. (2009). *Emotionen und Peripherphysiologie* [Emotions and Peripherpsychology]. Lengerich: Pabst Science Publishers, 2009.

Schleicher, R. & Trösterer, S. (2009). The 'Joy-of-Use'-Button: Recording Pleasant Moments While Using a PC. In *Proceeding Proceedings of the 12th IFIP TC 13 International Conference on Human-Computer Interaction: Part II (INTERACT '09)*. 630-633.

Schnotz, W., Bannert, M., & Seufert, T. (2002). Toward an integrative view of text and picture comprehension: Visual effects on the construction of mental models. In A. Graesser, J. Otero, & J. A. Leon (Eds.), *The Psychology of Science Text Comprehension.* Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 385-416.

Schomaker, L., J. Nijtmans, A. Camurri, F. Lavagetto, P. Morasso, C. Benoit, T. Guiard-Marigny, B. Le Goff, J. Robert-Ribes, A. Adjoudani, I. Defee, S. Munch, K. Hartung, & J. Blauert. (1995). *A Taxonomy of Multimodal Interaction in the Human Information Processing System. Multimodal Integration for Advanced Multimedia Interfaces (MIAMI).* ESPRIT III, Basic Research Project 8579.

Schupp, J., & Gerlitz, J.-Y. (2008). *BFI-S: Big Five Inventory-SOEP. Zusammenstellung sozial-wissenschaftlicher Items und Skalen* [BFI-S: Big Five Inventory-SOEP collection of socio-scientific items and scales]. Bonn: GESIS.
Bonn.

Schwarz, N., & Clore, G.L. (2003). Mood as information: 20 years later. *Psychological Inquiry 14*(3), 296-303.

Seebode, J. (2009). *Evaluation und Weiterentwicklung eines multimodalen Rauminformationssystems.* [Evaluation and further development of a multimodal room information system], Master's thesis. TU Berlin, Germany. Retrieved from
http://www.prometei.de/fileadmin/prometei.de/publikationen/Seebode_Magisterarbeit.pdf

Seebode, J., Schaffer, S., Wechsung, I., & Metze, F. (2009). Influence of training on direct and indirect measures for the evaluation of multimodal systems. In *Proceedings of the 10th Annual Conference of the ISCA (Interspeech 2009)*. 300-303

Shackel, B. (2009). Human-computer interaction - Whence and whither?. *Interacting with Computers 21*(5-6), 353-366.

Sheldon, K.M., Elliot, A.J., Kim, Y., & Kasser T. (2001). What is satisfying about satisfying events? Testing 10 candidate psychological needs. *Journal of Personality and Social Psychology, 80,*325-339.

Shi, H., & Maier, A. (1996). Speech enabled shopping application using Microsoft SAPI. *International Journal of Computer Science and Network Security 6*(9), 33- 37.

Shneiderman, B. (1998). *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Reading: Addison Wesley.

Silvia, P.J., & Warburton, J. B. (2006). Positive and negative affect: Bridging states and traits. In D. L. Segal & J. C. Thomas (Eds.), C*omprehensive handbook of personality and psychopathology*, Vol. 1: Personality and everyday functioning. New York: Wiley. pp. 268-284.

Simpson, A., & Fraser, N.M. (1993). Black box and glass box evaluation of the SUNDIAL system. In *Proceedings of the Third European Conference on Speech Communication and Technology (Eurospeech'93)*, 1423-1426.

Sluckin, W., Hargreaves, D. J., & Colman, A. M. (1983). Novelty and human aesthetic preferences. In J. Archer & L. Birke (Eds.), *Exploration in animals and humans.* Wokingham: Van Nostrand Reinhold, pp. 245-269.

Smith, B., Caputi, P. & Rawstorne, P. (2007). The development of a measure of subjective computer experience. *Computers in Human Behavior 23*(1), 127–145.

Solomon, S. (1978). Measuring dispositional and situational attributions. *Personality and Social Psychology Bulletin 4*(4), 589-594.

Spencer, W.D., & Raz, N. (1995). Differential effects of aging on memory for content and context: a meta-analysis. *Psychology & Aging 10*(4), 527–539.

Spool, J, & Schroeder, W. (2001). Test web sites: five users is nowhere near enough. In *Proceedings of the Conference on Human Factors in Computing Systems. Extended Abstracts (CHI 2001)*, 285-286.

Stegmann, J., Henke, K., & Kirchherr, R. (2008). Multimodal interaction for access to media content.Paper presented at *the 12th International Conference on Intelligence in Next Generation Networks (ICIN2008)*. Retrieved from http://www.icin.biz/files/2008papers/Poster-08.pdf

Steiner, W. J., & Weber, A. (2009). Ökonometrische Modellbildung. In C. Baumgarth, M. Eisend, & H. Evanschitzky (Eds.), *Empirische Mastertechniken.*Wiesbaden: Gabler. pp. 389-429.

Sturm, J.A. (2005). *On the Usability of Multimodal Interaction for Mobile Access to Information Services.* Ph.D. thesis, Universiteit Nijmegen, Nijmegen: PrintPartners Ipskamp.

Sturm, J.A., Bakx, I., Cranen, B., Terken, J., & Wang, F. (2002). Usability evaluation of a Dutch multimodal system for train timetable information. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, 255–261.

Sturm, J., & Boves, L. (2005). Effective error recovery strategies for multimodal form-filling applications. *Speech Communication 45*(3), 289–303.

Suhm, B., Myers, B., & Waibel, A. (2001). Multimodal error correction for speech user interfaces. *ACM Transactions on Computer Human Interaction 8*(1), 60–98.

Szpunar, K.K., Schellenberg, E.G., & Pliner, P. (2004). Liking and memory for musical stimuli as a function of exposure. *Journal of Experimental Psychology: Learning, Memory, and Cognition 30*(2), 370-381.

Tewes, U. (1991). *HAWIE-R. Hamburg-Wechsler-Intelligenztest für Erwachsene* [Hamburg-Wechsler-Intelligence Test for Adults]. Bern: Huber.

Tractinsky, N., Katz, A.S., & Ikar, D. (2000). What is beautiful is usable", *Interacting with Computers 13*(2), 127–145.

Tullis, T. & Stetson, J. (2004). A Comparison of Questionnaires for Assessing Website Usability. In *Proceedings of the Usability Professionals Association Conference (UPA 2004)*, 7-11.

Turunen, M., Hakulinen, J., Melto, A., Heimonen, T., Laivo,T., & Hella, J. (2009). SUXES - user experience evaluation method for spoken and multimodal interaction. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech 2009)*, 2567-2570.

Väätäjä, H., & Roto, V. (2009). Questionnaires in User Experience Evaluation. Paper presented at the *Workshop User Experience Evaluation Methods in Product Development*

(UXEM'09). Retrieved from http://research.nokia.com/files/public/VaatajaRoto-Questionnaires.pdf

Van Vaerenbergh, Y. & Thomas, T.D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research 25*(2), 195-217.

Van Vliet, P.J.A., Kletke, M.G., & Chakraborty, G. (1994). The Measurement of Computer Literacy – A Comparison of Self-Appraisal and Objective Tests. *International Journal of Human-Computer Studies 40*(5), 835-857.

Virzi, R.A. (1992). Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors 34*(4), 457-471.

Vollmeyer, R., Imhof, M. & Beierlein, C. (2006). Gender differences in learning the SPSS software. In R. Sun and N. Miyake (Eds.), *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Societ*. Hillsdale, NJ: Erlbaum. pp. 2323-2328.

Walker, M.A., Litman, D.J., Kamm, C.A. & Abella, A. (1997). *Paradise: A general framework for evaluating spoken dialogue agents.* In Proceedings of the 35th Annual Meeting of the Association of Computational Linguists (ACL/EACL), 271-280.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale - 4th Edition*. San Antonio, TX: Pearson

Wechsung, I., Engelbrecht, K.-P., Naumann, A., Schaffer, S., Seebode, J., Metze, F., & Möller, S. (2009). Predicting the quality of multimodal systems based on judgments of single modalities. In *Proceedings of 10th Annual Conference of the International Speech Communication Association* (Interspeech 2009) 1827-1830.

Wechsung, I., Engelbrecht, K.-P., Kühnel, C., Möller, S., & Weiss, B. (2012a). Measuring the Quality of Service and Quality of Experience of Multimodal Human-Machine Interaction Journal on Multimodal User Interfaces. In *Journal on Multimodal User Interfaces (6)*1, 73–85.

Wechsung, I., Hurtienne, J., & Naumann, A. (2009). Multimodale Interaktion: Intuitiv, robust, bevorzugt und altersgerecht? [Multimodal interaction: intuitive, preferred and senior-friendly?] In H. Wandke, S. Kain, & D. Struve (Eds.), *Mensch & Computer 2009: Grenzenlos frei!?* . München: Oldenbourg, pp. 213-222.

Wechsung, I., Jepsen, K., Burkhardt, F., Köhler, A., & Schleicher, R. (2012b). View from a Distance: Comparing Online and Retrospective UX-Evaluations. In *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services companion (MobileHCI 2012)*, 113-118 .

Wechsung, I., & Naumann, A. (2008). Evaluation Methods for Multimodal Systems: A Comparison of Standardized Usability Questionnaires. In *Proceedings of the 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems* (PIT 2008), 276-284.

Wechsung, I., Naumann, A. & Schleicher, R. (2008). Views on usability and user experience: from theory and practice. Paper presented at the *NordiCHI Wokshop Research Goals and Strategies for Studying User Experience and Emotion*. Retrieved from http://www.cs.uta.fi/~ux-emotion/submissions/Wechsung-etal.pdf

Weiss, B., Kühnel, C., Wechsung, I., Fagel, S., &Möller, S. (2010). Quality of Talking Heads in Different Interaction and Media Contexts. *Speech Communication, 52*(6), 481–492.

Weiss, B., Wechsung, I. & Marquardt, S. (submitted). Multimodal HCI: Effects of first impression and single modality ratings for two case studies.

Wharton, C., Rieman, J., Lewis, C. & Polson, P. (1994). The Cognitive Walkthrough Method: A Practitioner's Guide. In J. Nielsen and R. Mack, (Eds.), *Usability Inspection Methods*.Wiley: New York, pp. 105-140.

Whitley, B.E. (1996). Gender Differences in Computer-Related Attitudes. It Depends on What You Ask. *Computers in Human Behavior, 12*(2)*,* 275-289.

Wickens, C.D. (1984). Processing resources in attention. In R. Parasuraman and R. Davies (Eds.), *Varieties of attention*.New York: Academic Press, pp. 63–101.

Wickens, C.D. (2000). *Engineering Psychology and Human Performance*. Upper Saddle River, NJ: Prentice Hall.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science, 3*(2), 159–177.

Williams, S.M. (1987). Repeated Exposure and the. Attractiveness of Synthetic Speech: An Inverted-U Relationship. *Current Psychology 6*(2), 148-154.

Witt, H. (2001). Forschungsstrategien bei quantitativer und qualitativer Sozialforschung [Research strategies in quantitaive and qualitative social research]. *Forum Qualitative Sozialforschung 2*(1). Retrieved from http://www.qualitative-research.net/index.php/fqs/article/view/969/2114.

Wolters, M., Engelbrecht, K.-P., Gödde, F., Möller, S., Naumann, A., & Schleicher, R. (2010). Making it easier for older people to talk to smart homes: Using help prompts to shape users' speech. *Universal Access in the Information Society, 9*(4), 311-325.

Wolters, M., Georgila, K., Moore, J.D., Logie, R.H., MacPherson, S.E., & Watson, M. (2009). Reducing working memory load in spoken dialogue systems. *Interacting with Computers 21*(4), 276-287.

Woolrych, A., & Cockton, G. (2001). Why and when five test users aren't enough. In *Proceedings of the 15th Annual Conference of the British HCI Group (IHM-HCI 2001)*, Vol. 2, 105-108.

Zajonc, R.B. (1968). The Attitudinal Effects of Mere Exposure. *Journal of Personality and Social Psychology 9*(2), 1-27.

Zijlstra, F.R.H. (1993). *Efficiency in work behavior. A design approach for modern tools.* PhD thesis, Delft University of Technology, Delft: Delft University Press.

Zijlstra, F.R.H., Van Doorn, L. (1985). *The construction ofa scale to measure perceived effort.* Technical Report. Delft: Delft University Press.

Zimbardo, P.G. (1995). *Psychologie*, Berlin: Springer, 1995.

Appendix A - Additional Material Related to Empirical Studies

A.1 Additional Results of Study 4.1

Table A.1.1 Ratings on SUMI Sub-Scales (Min.=10/ Max.=30)

Scale	System	Mean	SD	F (2,40)	*p* (part. eta²)
Efficiency	Tablet PC	19.00	3.39	6.19	.005 (.236)
	PDA	19.90	3.48		
	Desktop PC	16.67	3.15		
Affect	Tablet PC	19.33	2.13	10.02	.000 (.334)
	PDA	19.95	2.13		
	Desktop PC	17.38	2.73		
Helpfulness	Tablet PC	22.05	1.96	1.46	.244 (.068)
	PDA	22.19	2.09		
	Desktop PC	21.48	1.89		
Control	Tablet PC	22.67	2.27	3.33	.046 (.143)
	PDA	22.24	2.21		
	Desktop PC	21.38	2.22		
Learnability	Tablet PC	15.57	3.23	5.98	.005 (.230)
	PDA	14.57	3.47		
	Desktop PC	16.90	4.40		
Global	Tablet PC	40.19	7.87	6.56	.003 (.247)
	PDA	45.29	10.14		
	Desktop PC	38.04	7.06		

Table A.1.2 Ratings on AttrakDiff Sub-Scales (Min.=-3/ Max.=3)

Scale	System	Mean	SD	F (2,38)	*p* (part. eta²)
Pragmatic Qualities	Tablet PC	.91	.84	16.80	.000 (.469)
	PDA	.01	.89		
	Desktop PC	1.34	.61		
Hedonic Qualities-Stimulation	Tablet PC	.63	.84	3.59	.037 (.159)
	PDA	.32	.57		
	Desktop PC	.27	.64		
Hedonic Qualities-Identity	Tablet PC	.81	.81	23.03	.000 (.548)
	PDA	.64	.76		
	Desktop PC	-.33	.89		
Attractiveness	Tablet PC	.99	1.04	4.04	.026 (.175)
	PDA	.34	.88		
	Desktop PC	.74	.66		

Ina Wechsung, *An Evaluation Framework for Multimodal Interaction*,
T-Labs Series in Telecommunication Services,
DOI: 10.1007/978-3-319-03810-0, © Springer International Publishing Switzerland 2014

Table A.1.3. Ratings on SASSI Sub-Scales (Min.=0/Max.=4)

Scale	System	Mean	SD	t(20)	*p*
Global	Tablet PC	2.24	.52	2.00	.059
	PDA	1.96	.48		
Speed	Tablet PC	1.64	.50	.40	.693
	PDA	1.59	.46		
Accuracy	Tablet PC	2.11	.69	1.44	.166
	PDA	1.90	.58		
Likeability	Tablet PC	2.60	.62	5.00	.065
	PDA	2.23	.61		
Habituality	Tablet PC	1.69	.74	1.20	.246
	PDA	1.52	.52		
Cognitive Demand	Tablet PC	2.30	.59	2.25	.036
	PDA	1.86	.68		
Annoyance	Tablet PC	2.49	.59	1.79	.089
	PDA	2.18	.64		

Table A.1.4 Ratings on SUS (Global Scale: Min. = 0/Max. = 100; Items: Min.=0/Max.= 4)

Item	Device	Mean	SD	t(20)	*p*
Global Scale	Tablet PC	53.93	16.59	1.32	.232
	PDA	50.12	13.38		
I think that I would like to use the voice control frequently.	Tablet PC	2.00	1.18	1.45	.162
	PDA	1.74	1.05		
I found the voice control unnecessarily complex.	Tablet PC	2.19	.93	.72	.480
	PDA	1.74	1.05		
I thought the voice control was easy to use.	Tablet PC	2.10	.94	1.24	.229
	PDA	1.80	.98		
I think that I would need the support of a technical person to be able to use the voice control.	Tablet PC	2.33	.86	.70	.493
	PDA	2.42	.93		
I found the various functions in this system were well integrated.	Tablet PC	2.00	.84	.44	.666
	PDA	2.10	.70		
I thought there was too much inconsistency in the voice control.	Tablet PC	2.04	1.02	2.36	.029
	PDA	1.67	.66		
I would imagine that most people would learn to use the voice control very quickly.	Tablet PC	2.48	.98	.75	.463
	PDA	2.29	1.01		
I found the voice control very cumbersome to use.	Tablet PC	2.14	1.01	.55	.590
	PDA	2.00	.89		
I felt very confident using the voice control.	Tablet PC	1.76	1.00	.78	.446
	PDA	1.57	.87		
I needed to learn a lot of things before I could get going with the voice control.	Tablet PC	2.52	.68	1.9	.072
	PDA	2.14	.96		

A.2 Detailed Descriptions of Pilot and Data Collection Studies for Questionnaire Construction

A.2.1 Pilot Study A

Design

The study was carried out in a between-subject design. The participants were randomly assigned to the test conditions. The independent variable was input modality with the conditions touch, speech, motion control, and multimodal. The dependent variables were performance data (task duration, successful task completion, and task aborts) as well as subjective ratings of user satisfaction.

Participants

A total of 62 students of TU Berlin (Berlin Institute of Technology), with a mean age of 24 years (SD=2.7) participated in the study. 44 of them were male, 18 were female. The participants were recruited at a students' dormitory and received small presents for their participation. Gender effects werenot found. None of the participants was familiar with either the application or the technical device used in the study.

Materials and Tasks

The device used was a smart phone (HTC Touch Diamond; IBM embedded ViaVoice) controllable via motion (tilt and twist), speech and touch input. For speech and motion input activation, a respective button had to be pressed before starting to speak or tilt and twist. The push-to-talk trigger is placed on the left hand side and the push-to-move button on the front of the device.

Each participant tested only one modality condition which was either motion, speech, touch or the multimodal condition, in which the participants could choose the input modalities or a combination of modalities themselves. Each participant solved 10 tasks with the given modality. All tasks could be solved in any modality. The following combination of tasks was chosen, which included easy and more difficult tasks

- Access your voice messages.
- Retrieve the voice message of “garage Heinz”.
- Delete the message.
- Access your e-mail inbox.
- View the e-mail by Philip.
- Access your fax inbox.
- View the fax by Felix.

- Access your voice messages and sort them from A to Z.
- Redirect your calls to the default number and confirm this change.
- Return to the menu and close the application.

To assess subjective ratings regarding user satisfaction, the AttrakDiff (Hassenzahl, Burmester, & Koller, 2003) questionnaire was used. Additionally, participants were asked to rated their perceived mental effort on the SEA-scale (Eilers, Nachreiner, Hänecke, 1986). Moreover, a non-final version of the MMQQ was used (cf. Section 4).

Furthermore, interaction performance data was protocolled by the experimenter:

- Successful task completion (i.e. number of trials that were successful on the first attempt)
- Aborts of task execution after two minutes (task execution was aborted by the experimenter when exceeding 2 minutes)
- Task completion time

Procedure

First, the experimenter gave a short introduction to the device and the modality to use. In the multimodal condition, all three input modalities were explained. Then, the participants had to solve the ten tasks with the modality they were assigned to. The time for solving the task and the number of errors made were recorded by the experimenter who was sitting next to the participant observing him or her solving the task (time was measured with a stopwatch) and listed on paper together with the errors made). If the participant was not able to solve the task within two minutes, he or she was asked to move on to the next task. After solving all ten tasks, the participants filled in the questionnaire.

A.2.2 Pilot Study B

Design

The study was carried out in a within-subject design. As in Pilot Study A, the independent variable was the input modality with the conditions touch, speech, motion control, and multimodal. The order of the three single input modality conditions touch, speech, and motion was balanced (latin square). The fourth and last condition was always the multimodal condition, since all modalities had to be trained once before a decision for a modality or modality combination could be made by the participants. The dependent variables again were performance data (task duration, successful task completion, and task aborts) as well as subjective ratings of user satisfaction. Additionally, in the multimodal test block, the chosen input modality was recorded.

Participants

30 participants (15 male, 15 female) aged between 22 and 78 years took part in the study. Half of them were younger than 35 years (M=28.73, SD=3.58) and half of them were older than 55 years (M=65.73, SD=7.26). None of the participants was familiar with either the application or the technical device. Regarding the modalities tested, 48.3% of the participants had prior experience with touch, 27.6% with speech, and 10.3% with motion control.

Material and Tasks

The device used for testing was the same as in Pilot Study A, the HTC Touch Diamond smartphone with touch, speech, and motion input. To assess ratings regarding the users' perceptions, again, the AttrakDiff (Hassenzahl, Burmester, & Koller, 2003) questionnaire,the SEA-scale (Eilers, Nachreiner, & Hänecke, 1985) and a non-final version of the MMQQ were employed. Interaction performance data (successful task completion in the first trial, aborts of task execution after three unsuccessful attempts, task duration measured with a stop watch) were logged as in the first study.

Additionally an early version of the QUESI questionnaire was used (Hurtienne & Naumann, 2010). It measures the subjective consequences of intuitive use. The participants had to execute 4 blocks of tasks with a total of 14 tasks similar to the ones in Pilot Study A (get messages, reply to them, forward, and sort messages as well as changing notification settings).

Procedure

First, the participants filled in a questionnaire about their demographic data and their experience with devices and input modalities. Then, all participants executed the 14 tasks for each of the four modality conditions. If the task goal was not achieved within three trials, task execution was aborted by the experimenter and the next task started. First, participants were asked to solve all tasks with a given modality. Then, participants evaluated the interaction via the questionnaire. This was repeated for all three modalities – touch, speech, and motion. In the final multimodality condition, participants were free to choose the modalities they used for solving the task. Here, they could always switch or combine modalities as they liked. Again, the participants evaluated the interaction after solving all tasks in this condition.

A.2.3 Detailed Descriptions of Data Collection Study

Participants

30 German-speaking individuals (15m, 15f, M = 28 yrs.) took part in the study. All of them were paid for their participation. The majority (70%) was familiar with touch input; voice control was considerably less known (30%).

Material

The tested application,named Mobile Multimodal Information Cockpit,offered the functionality of a remote control, a mobile TV and video player, video on demand services and games. The application was implemented on a Tablet PC, the Samsung Q1 Ultra. The tested system is controllable via a graphical user interface with touch screen and speech input. The output is given via the graphical user interface and audio feedback. For some tasks only one of the modalities was available. To assess ratings for hedonic and pragmatic qualities the AttrakDiff questionnaire (Hassenzahl, Burmester & Koller, 2003) was employed. Furthermore, the SEA-scale (Eilers, Nachreiner, & Hänecker, 1986)was used as a measure of perceived mental effort. In addition, an early version of the MMQQ was employed.

Procedure

The experiment consisted of two blocks: one task-oriented and one explorative. Half of the participants started with the task-oriented block followed by the explorative block. For the other half of participants the order was reversed (within-subject design). They were either instructed to perform 16 given tasks (e.g. logging in to the system, switching the channel, searching for a certain movie, a certain TV show, a certain actor, increasing volume, decreasing volume, playing the quiz, switching between the different categories) or to use the next 15 minutes to do whatever they want to do with the device. The duration was set to 15 minutes since pre-tests showed that this was the average time to accomplish all tasks. In both test blocks the participants were free to choose the input modality. It was at any time possible to switch or combine modalities. In order to rate the previously tested condition, the questionnaires had to be filled in after each test block. To analyse which modality the participants used, for every interaction step, the modality used to perform the step was logged.

A.3 Additional Results of Study 5.1

The results show differences between the three versions of the system for all AttrakDiff scales. For the scale pragmatic qualities the touch-based version was rated bestand the voice control version worst, *F*(2,66)= 93.79, *p*=.000, *eta*^{2}= .740. For bothhedonic scales, the multimodal version was rated best. Regarding Hedonic Qualities-Stimulationthe speech version received the lowest ratings, *F*(2,68)=12.84, *p*=.000, *eta*^{2}= .274. For Hedonic Qualities-Identity the touch-based version was rated worst,*F*(1.65, 55.99) = 15.35, *p* = .000, *eta*^{2}= .311.

The Attractiveness scale, the AttrakDiff scale covering pragmatic as well as hedonicqualities, showed the lowest ratings for the speech-based version and highest ratings for the touch-based version, *F*(1.51,51.22)= 47.53, *p*=.000, *eta*^{2}=.583.

Regarding the overall scale, the scale based on the mean of all items, the speechbasedversion was rated worse than the touch-based version and multimodal systems version, *F* (2, 66) = 38.38, p = .000, *eta*^{2} = .538.The touch-based version and the multimodal version were rated equally good.

A.4 Additional Results of Study 6.4

Regarding Pragmatic Qualities, a repeated measure ANOVA with Sidak-corrected post-hoc tests showed that the explorative condition was rated worse than the multimodal condition. Furthermore, the touch condition was perceived as less stimulating (Hedonic Qualities-Stimulation) compared to the multimodal condition. No difference was found on the scale Hedonic Qualities-Identity, Attractiveness and for mental effort (SEA scale).

For the performance measures post hoc test showed that participants were faster and more successful in the speech condition compared to the touch condition. All results are displayed in Table A.4.1.

Table A.4.1.. Comparison between different modalities for Study 6.4

Measure	M_{Speech} SD_{Speech}	M_{Touch} SD_{Touch}	$M_{\text{Multimodalt}}$ $SD_{\text{Multimodalt}}$	$M_{\text{Explorative}}$ $SD_{\text{Explorative}}$	*F* *(df)*	*p*	part. eta^2
Pragmatic Qualities	1.51 (1.23)	1.08 (.99)	1.46 (1.04)	.82 (1.29)	3.93 (3,87)	.023	.119
Hedonic Qualities - Stimulation	1,47 (1.22)	.92 (1.19)	1.38 (1.22)	1.08 (1.07)	4.36 (3,87)	.007	.131
Hedonic Qualities - Identity	1.08 (1.22)	1.02 (1.00)	1.30 (1.07)	1.23 (.90)	.867 (3,87)	.462	.029
Attractiveness	1.57 (1.26)	1.47 (1.02)	1.70 (.98)	1.25 (1.26)	1.49 (3,87)	.223	.049
Mental effort (SEA scale)	28.43 33.62	26.17 17.43	20.73 13.14	34.50 25.73	2.43 (1.97, 56.98)	.099	.077
Task duration in mm:ss.ms	7:03.37 (02:31.12)	08:51.10 (04:00,68)	07:50,84 (03:12,68)	n.a.	3.93 (2,56)	.025	.123
Task sucess	61.46 (13.74)	49.58 (17.60)	57.29 (20.70)	n.a.	5.00 (2,58)	.010	.147

Appendix B –Additional Material Related to the MMQQValidation

B.1 Complete List of Items

Table B.1.1 Initial List of Items with Corresponding Results of Jarque-Bera-Normality-Test, Item Difficulty Indices, and Item Discrimination Indices

Abbreviation	Original German Wording	English Translation	Jarque-Bera χ^2 (p)	Item difficulty	Item discrimination
PER1	**Das System ist unfreundlich - freundlich.**	**The system is unfriendly - friendly.**	4.13 (0.13)	70	.75
PER2	**Das System ist unsympathisch - sympathisch.**	**The system is unsympathetic - sympathetic.**	1.28 (0.53)	68	.69
PER3	Die Interaktion mit dem System ist unangenehm - angenehm.	The interaction with the system is unpleasant - pleasant.	4.47 (0.11)	64	.72
PER4	Die Interaktion mit dem System ist nervig - spaßig.	The interaction with the system is nerved - amusing.	1.55 (0.46)	64	.82
AEST1	**Die Gestaltung des Systems ist abstoßend - anziehend.**	**The design of the system is off - putting appealing.**	1.17 (0.56)	68	.60
AEST2	**Die Gestaltung des Systems ist hässlich - schön.**	**The design of the system ugly - beautiful.**	1.68 (0.43)	63	.66

Ina Wechsung, *An Evaluation Framework for Multimodal Interaction*,
T-Labs Series in Telecommunication Services,
DOI: 10.1007/978-3-319-03810-0, © Springer International Publishing Switzerland 2014

Abbreviation	Original German Wording	English Translation	Jarque-Bera χ^2 (p)	Item difficulty	Item discrimination
AEST3	Die Gestaltung des Systems ist unansehnlich - ansehnlich.	The design of the system is unsightly - sightly.	0.43 (0.81)	61	.83
DISC1	Die Interaktion mit dem System ist entmutigend - motivierend.	The interaction with the system is discouraging - encouraging.	1.79 (0.41)	64	.77
DISC2	Die Interaktion mit dem System ist langweilig - unterhaltsam.	The interaction with the system is boring - entertaining.	3.56 (0.17)	64	.47
DISC3	**Die Interaktion mit dem System ist eintönig - abwechslungsreich.**	**The interaction with the system is monotonous - varied.**	1.50 (0.47)	64	.69
DISC4	Die Gestaltung des Systems ist konventionell - originell.	The design of the system is conventional - original.	2.07 (0.36)	52	.50
DISC5	**Die Gestaltung des Systems ist reizlos - reizvoll.**	**The design of the system is plain - attractive.**	1.32 (0.52)	64	.75
DISC6	**Die Gestaltung des Systems ist öde - interessant.**	**The design of the system is dull - interesting.**	0.77 (0.68)	55	.63

Abbreviation	Original German Wording	English Translation	Jarque-Bera χ^2 (*p*)	Item difficulty	Item discrimination
EFFIT1	Die Interaktion mit dem System ist umständlich - direkt.	The interaction with the system is cumbersome - direct.	2.16 (0.34)	55	.69
EFFIT2	**Die Interaktion mit dem System ist holprig - flüssig.**	**The interaction with the system is clunky - smooth.**	1.90 (0.39)	53	.79
EFFIT3	Die Interaktion mit dem System ist lahm - flott.	The interaction with the system is sluggish - responsive.	1.50 (0.47)	48	.77
EFFIT4	**Die Interaktion mit dem System ist langsam - schnell.**	**The interaction with the system is slow - fast.**	3.19 (0.20)	63	.84
EFFIM1	**Die Interaktion mit dem System ist beanspruchend - schonend.**	**The interaction with the system is demanding - relieving.**	1.07 (0.59)	60	.84
EFFIM2	**Die Interaktion mit dem System ist belastend - entlastend.**	**The interaction with the system is taxing - disencumbering.**	1.99 (0.37)	51	.74
INTUI1	Das System ist schwierig zu bedienen - einfach zu bedienen.	The system is difficult to operate - easy to operate.	3.12 (0.21)	62	.71

Abbreviation	Original German Wording	English Translation	Jarque-Bera χ^2 (*p*)	Item difficulty	Item discrimination
INTUI2	Die Interaktion mit dem System ist kompliziert - unkompliziert.	The interaction with the system is complicated - uncomplicated.	1.27 (0.53)	71	.80
INTUI3	**Die Gestaltung des Systems ist chaotisch - geordnet.**	**The design of the system is chaotic - well-structured.**	1.39 (0.50)	58	.86
INTUI4	**Die Gestaltung des Systems ist unklar - klar.**	**The design of the system is unclear - clear.**	1.56 (0.46)	69	.81
INTUI5	Die Gestaltung des Systems ist inkonsistent - konsistent.	The design of the system is consistent - inconsistent.	1.12 (0.57)	60	.67
EFFEC1	**Das System ist fehleranfällig - fehlertolerant.**	**The system is error-prone - error-tolerant.**	0.81 (0.67)	47	.68
EFFEC2	**Die Interaktion mit dem System ist fehlerreich - fehlerarm.**	**The interaction with the system is high in errors - low in errors.**	0.21 (0.90)	52	.68
EFFEC3	Die Gestaltung des Systems ist ablenkend - zielführend.	The design of the system is distracting - targeted.	1.20 (0.55)	67	.79

Abbreviation	Original German Wording	English Translation	Jarque-Bera χ^2 (p)	Item difficulty	Item discrimination
LEARN1	**Die Gestaltung des Systems ist störend - helfend.**	**The design of the system is impeding - helpful.**	0.77 (0.68)	68	.80
LEARN2	Die Gestaltung des Systems ist ungeeignet - geeignet.	The design of the system is unsuitable - suitable.	3.75 (0.15)	69	.81
LEARN3	Die Gestaltung des Systems ist unangebracht - angebracht.	The design of the system is inappropriate - appropriate.	2.08 (0.35)	63	.79
IQ1	**Die verschiedenen Eingabemöglichkeiten: sind nachteilig - sind vorteilhaft.**	**The different input modalities are disadvantageous - advantageous.**	1.25 (0.53)	63	.78
IQ2	**Die verschiedenen Eingabemöglichkeiten blockieren sich - ergänzen sich.**	**The different input modalities are blocking each other - are complementing each other.**	1.61 (0.45)	71	.70
IQ3	**Die verschiedenen Eingabemöglichkeiten: behindern sich - unterstützen sich.**	**The different input modalities hinder each other - support each other.**	1.17 (0.56)	68	.72
IQ4	Die verschiedenen Eingabemöglichkeiten sind schlecht miteinander integriert - sind gut miteinander integriert.	The different input modalities are poorly integrated with each other - are well integrated with each other.	1.07 (0.59)	67	.71

Abbreviation	Original German Wording	English Translation	Jarque-Bera χ^2 (*p*)	Item difficulty	Item discrimination
IQ5	Die verschiedenen Eingabemöglichkeiten: sind schlecht aufeinander abgestimmt - sind gut aufeinander abgestimmt.	The different input modalities are poorly aligned with each other - are well aligned with each other.	0.59 (0.74)	68	.69
IQ6	Die verschiedenen Eingabemöglichkeiten sind schlecht zu koordinieren - sind gut zu koordinieren.	The different input modalities are poorly to coordinate - well to coordinate.	5.69 (0.06)	61	.79
IQ7	**Die verschiedenen Eingabemöglichkeiten sind schlecht zu kombinieren - sind gut zu kombinieren.**	**The different input modalities are poorly to combine - are well to combine.**	1.99 (0.37)	66	.66
IQ8	**Die verschiedenen Eingabemöglichkeiten arbeiten schlecht zusammen - arbeiten gut zusammen.**	**The different input modalities are working poorly together - are working well together.**	1.25 (0.53)	65	.75
OQ1	Die Rückmeldungen des Systems sind dumm - klug.	The system's feedback is dumb - smart.	0.38 (0.83)	61	.57
OQ2	Die Rückmeldungen des Systems sind unnötig - notwendig.	The system's feedback is unnecessary - necessary.	2.04 (0.36)	59	.73
OQ3	Die Rückmeldungen des Systems sind hemmend - unterstützend.	The system's feedback is hindering - supporting.	1.14 (0.57)	69	.65

Abbreviation	Original German Wording	English Translation	Jarque-Bera χ² (*p*)	Item difficulty	Item discrimination
OQ4	**Die Rückmeldungen des Systems sind hinderlich - förderlich.**	**The system's feedback is inhibiting - facilitating.**	0.96 (0.62)	60	.78
OQ5	**Die Rückmeldungen des Systems sind sinnlos - sinnvoll.**	**The system's feedback is pointless - meaningful.**	2.16 (0.34)	67	.64
OQ6	**Die Rückmeldungen des Systems sind verwirrend - erklärend.**	**The system's feedback is confusing - explanatory.**	2.51 (0.29)	61	.72
OQ7	Die Rückmeldungen des Systems sind zweckdienlich -zwecklos	The system's feedback is expedient – futile.	0.87 (0.65)	72	.30
OQ8	Die Rückmeldungen des Systems sind konstruktiv-destruktiv	The system's feedback is constructive - destructive	1.80 (0.41)	66	.59
OQ9	Die Rückmeldungen des Systems sind zureichend – unzureichend.	The system's feedback is sufficient - insufficient.	1.59 (0.45)	67	.54
excluded$^{1.2}$	Die Gestaltung des Systems ist unergonomisch - ergonomisch.	The design of the system is unergonomic - ergonomic.	2.37 (0.31)	59	.68

Abbreviation	Original German Wording	English Translation	Jarque-Bera χ^2 (*p*)	Item difficulty	Item discrimination
excluded[3]	Die Interaktion mit dem System ist beschwerlich - unbeschwerlich.	The interaction with the system is arduous - effortless.	1.27 (0.53)	58	.81
excluded[4]	Die Interaktion mit dem System ist anstrengend - bequem.	The interaction with the system is exhausting - convenient	0.86 (0.65)	57	.80
excluded[4]	Die Interaktion mit dem System ist deprimierend - aufheiternd.	The interaction with the system is depressing - exhilarating.	1.03 (0.60)	61	.67
excluded[4]	Die Interaktion mit dem System ist ermüdend - fesselnd.	The interaction with the system tiring - compelling.	1.89 (0.39)	62	.71
excluded[1]	Die Interaktion mit dem System ist problematisch - unproblematisch.	The interaction with the system is problematic - unproblematic.	1.60 (0.45)	56	.80
excluded[3]	Die Interaktion mit dem System ist stressig - unstressig.	The interaction with the system is stressful - unstressful.	1.60 (0.45))	59	.73
excluded[1]	Die Interaktion mit dem System ist unbefriedigend - befriedigend.	The interaction with the system is dissatisfying - satisfying.	2.32 (0.31)	62	.87

Abbreviation	Original German Wording	English Translation	Jarque-Bera χ^2 (p)	Item difficulty	Item discrimination
excluded[2]	Die Interaktion mit dem System ist uneffektiv - effektiv.	The interaction with the system is ineffective - effective.	2.63 (0.27)	57	.74
excluded[3]	Die Interaktion mit dem System ist unlustig - lustig.	The interaction with the system is unfunny - funny.	3.32 (0.19)	55	.61
excluded[1]	Die Interaktion mit dem System ist unvorhersehbar - vorhersehbar.	The interaction with the system is unpredictable - predictable.	1.46 (0.48)	62	.72
excluded[1]	Die Rückmeldungen des Systems sind unangemessen - angemessen.	The system's feedback is inappropriate - appropriate.	3.04 (0.22)	58	.76
excluded[2]	Die Rückmeldungen des Systems sind ungenügend - genügend.	The system's feedback is insufficient - sufficient.	1.63 (0.44)	51	.69
excluded[3†]	Die Rückmeldungen des Systems sind unintelligent - intelligent.	The system's feedback is unintelligent - intelligent.	0.05 (0.98)	62	.67
excluded[1]	Die Rückmeldungen des Systems sind unpassend - passend.	The system's feedback is unsuitable - suitable.	5.01 (0.08)	61	.75

Abbre-viation	Original German Wording	English Translation	Jarque-Bera χ² (*p*)	Item difficulty	Item discrimi-nation
excluded[1]	Die Rückmeldungen des Systems sind unverständlich - verständlich.	The system's feedback is incomprehensible - comprehensible.	2.14 (0.34)	64	.66
excluded[3]	Das System ist agil - schwerfällig.	The system is agile - cumbersome.	2.58 (0.28)	47	.49
excluded[1]	Das System ist instabil - stabil.	The system is instable - stable.	0.66 (0.72)	58	.67
excluded[1‡]	Das System ist nutzlos - nützlich.	The system is useful - useless.	2.59 (0.27)	70	.82
excluded[1]	Das System ist unattraktiv - attraktiv.	The system is unattractive - attractive.	4.50 (0.11)	67	.82
excluded[1.5‡]	Das System ist unbrauchbar - brauchbar.	The system is use impractical - practical.	7.45 (0.02)	71	.74
excluded[1]	Das System ist unflexibel - flexibel.	The system is inflexible - flexible.	1.56 (0.46)	63	.79

Abbre-viation	Original German Wording	English Translation	Jarque-Bera χ² (*p*)	Item difficulty	Item discrimi-nation
exclud-ed[1]	Das System ist unkontrollierbar - kontrollierbar.	The system is uncontrollable - controllable.	3.26 (0.20)	63	.81
exclud-ed[3]	Das System ist unrobust - robust.	The system is unrobust - robust	0.88 (0.64)	55	.67
exclud-ed[3]	Das System ist untauglich - tauglich.	The system is unapt - apt.	3.07 (0.22)	65	.86
exclud-ed[2§]	Das System ist unübersichtlich - übersichtlich.	The system is unclear - clear.	3.42 (0.18)	68	.82
exclud-ed[1]	Das System ist unzuverlässig - zuverlässig.	The system is unreliable - reliable.	0.35 (0.84)	57	.78

Notes. Items in **bold-face** are items of the final questionnaire.

1 = No agreement regarding underlying dimension achieved

2 = Meaning not clear /not understood by participant

3 = Non-standard/unusual German

4 = No antonyms

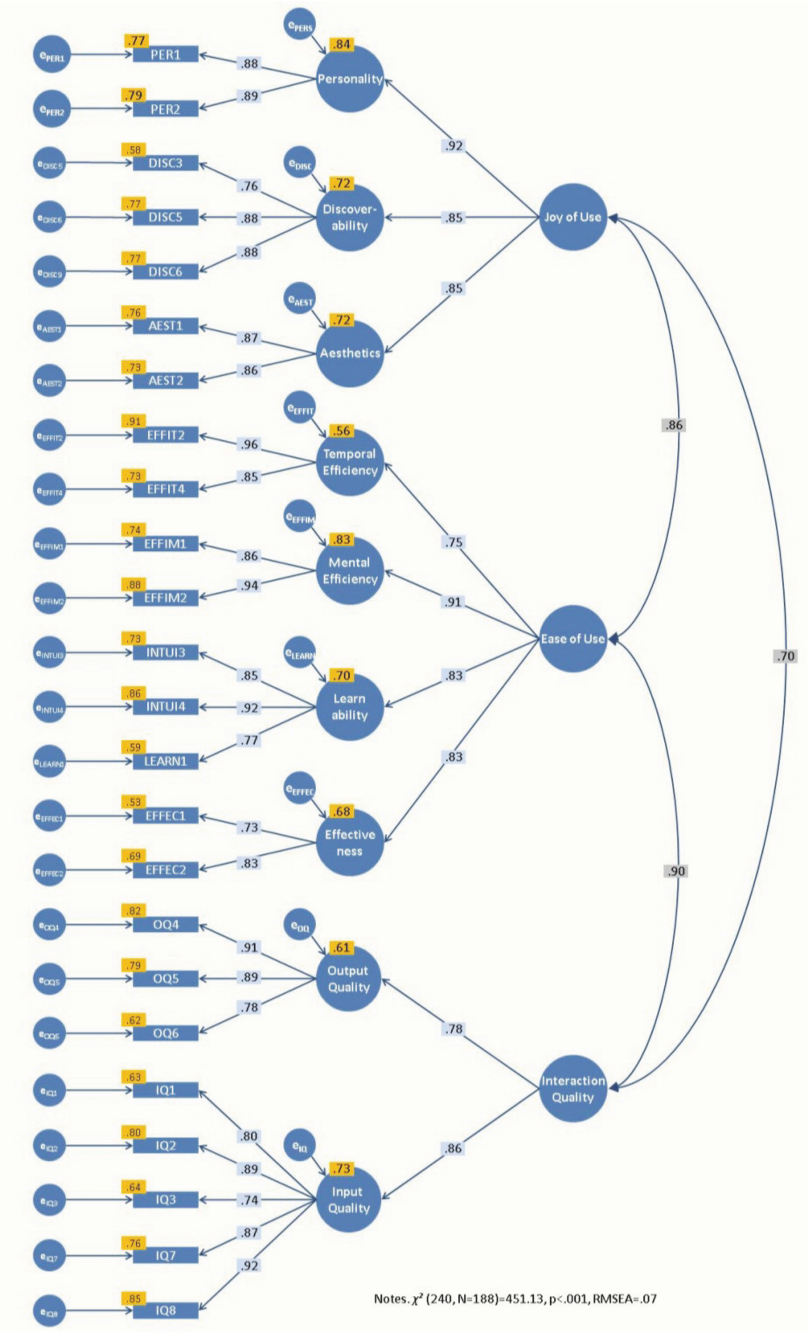
5 = Not normally distributed

† *dumm* (dumb) was expected as the corresponding negative adjective

‡ The assumed underlying dimension was utility

§ The meaning of “unclear - clear” was understood but not in combination with system, experts suggested that it should have been “the design of the system is un clear - clear”

B.2 Validated Model of Whole Questionnaire



Notes. χ^2 (240, N=188)=451.13, p<.001, RMSEA=.07

Fig.B.2. Final global model for validation sample (after removal of items LEARN2, item OQ3 and item IQ4).