Technische Universität Berlin

Measuring the User Experience of Mobile Applications – an Empirical Validation of a Quantitative Method

Masterarbeit zur Erlangung des akademischen Grades
Master of Science (M.Sc.)

im Studiengang
Human Factors

Vorgelegt von: Carina Kuhr, B.Sc. Psychologie
c.v.kuhr@gmail.com

1. Gutachter: Prof. Dr. phil. Manfred Thüring
2. Gutachter: Dr.-Ing. Michael Minge

Fakultät V Verkehrs- und Maschinensysteme
Institut für Psychologie und Arbeitswissenschaft
Fachgebiet Kognitionspsychologie & Kognitive Ergomonie
Technische Universität Berlin

Berlin, Oktober 2013

Abstract

Based on an empirically validated and well-established model of the components of user experience – the *CUE-model* (Thüring & Mahlke, 2007) – a new questionnaire, the *meCUE* (Minge & Riedel, 2013) consisting of 34 items that assess i*nstrumental* and *non-instrumental product perceptions, emotions, consequences and overall judgment* was developed. These subscales form four modules that can also be applied separately. In contrast to existing questionnaires, which lack central aspects of user experience, this tool offers a new method that assesses the major components of user experience in a comprehensive manner.

This study addresses the discriminative, convergent and criterion-related validity of the questionnaire. Moreover it is investigated whether the meCUE qualifies for application in industrial settings where interactive consumer goods are the object of evaluation. An expert review was conducted to identify three public transport apps with different quality of usability and design. These apps served as independent variables in the main study where the meCUE and relevant other questionnaires were applied by 24 participants to evaluate the interaction with the apps. The results show that the data of the meCUE successfully discriminates between the different apps. Moreover, the pattern of results is consistent with other validated questionnaires. In order to assess the convergent validity, correlations of the meCUE's modules with related constructs of other questionnaires were calculated and produced significant results. Finally the results reveal significant correlations with a relevant external criterion.

The thesis shows that the meCUE produces valid results and can be of special interest in product development of interactive systems.

Zusammenfassung

Basierend auf einem empirisch validierten und etablierten Modell über die Komponenten von User Experience – dem sogenannten *CUE-Modell* (Thüring & Mahlke, 2007) – wurde ein neuer Fragebogen, der *meCUE* (Minge & Riedel, 2013), entwickelt. Der meCUE umfasst 34 Items, die *aufgabenbezogene* und *nicht-aufgabenbezogene Produktwahrnehmungen, Emotionen, Konsequenzen* sowie das *Gesamturteil* der Interaktion eines Produktes erfassen. Aus diesen Subskalen ergeben sich insgesamt vier Module, die auch unabhängig voneinander angewandt werden können. Im Gegensatz zu bereits existierenden Fragebögen, die meist nicht alle zentralen Aspekte von User Experience (UX) erfassen, bietet dieses Messinstrument eine neue Methode, um die Hauptkomponenten von User Experience umfassend zu erheben.

Die vorliegende Studie überprüft die diskriminative, konvergente und Kriteriumsvalidität des meCUE-Fragebogens. Außerdem wird untersucht, ob der meCUE sich für die Evaluierung von interaktiven Konsumgütern im industriellen Kontext eignet. Es wurde eine Experten-Evaluation durchgeführt, um drei Apps zur Routenplanung mit öffentlichen Verkehrsmitteln auszuwählen, die sich hinsichtlich ihrer Qualität in Design und Usability unterscheiden. Diese drei Apps dienten dann als unabhängige Variablen in der Hauptstudie, in welcher der meCUE sowie weitere relevante UX-Fragebögen von insgesamt 24 Testteilnehmern genutzt wurden, um die Interaktion mit den Apps zu bewerten. Die Ergebnisse zeigen, dass der meCUE erfolgreich zwischen den Apps unterscheiden kann. Außerdem stimmen die Ergebnisse des meCUE mit den Ergebnissen der anderen UX-Fragebögen überein. Um die konvergente Validität zu erfassen, wurden Korrelationen der Module des meCUE mit verwandten Konstrukten anderer Fragebögen berechnet, die signifikante Ergebnisse zeigten. Zusätzlich konnten signifikante Korrelationen mit einem relevanten externen Kriterium gefunden werden.

Diese These zeigt auf, dass der meCUE valide Ergebnisse produziert und von besonderem Interesse in der Produktentwicklung von interaktiven Systemen sein kann.

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisor Dr. Michael Minge, whose expertise, guidance and support added considerably to the development of my thesis. I am also grateful to Prof. Dr. phil. Manfred Thüring who gave me the opportunity to work on this topic.

Additionally, special thanks go out to the four experts who agreed to participate in my expert review and made a major contribution to the methodological setup of the main study.

Furthermore, I would like to thank Frauke Rogalla and Murali Nair for making a lot of valuable comments and suggestions when reading my thesis.

Finally I would like to thank my family and friends for their support and company. I am especially thankful to Caroline Merkel for important discussions and patient answers to countless questions.

Abbreviations

CUE	Components of User Experience
CVPA	Centrality of Visual Product Aesthetics
df	Degrees of Freedom
HCD	Human-Centered Design
HCI	Human Computer Interaction
KPI	Key Performance Indices
LemTool	Layered Emotion Measurement tool
M	Mean
meCUE	Modular Evaluation of the Central Aspects of User Experience
PANAS	Positive Affect Negative Affect Scale
ROI	Return on Investment
RQ	Research Question
SAM	Self-Assessment Manikin
SD	Standard Deviation
ServUX	Service User eXperience
TA	Affinity for Technology
UEQ	User Experience Questionnaire
UX	User Experience
VisAWI	Visual Aesthetics of Websites Inventory

List of Figures

List of Tables

Table of Contents

1. Introduction

Good human-machine interaction is an ever-increasing factor determining the quality of life. As web technologies become ubiquitous and the *Internet of things* is by far no science fiction anymore, they have become a crucial part of people's daily routines. The numerous situations where people make use of technologies range from simple tasks such as getting directions from a navigational system or reading eBooks on a tablet computer to more complex usage of machines in industrial contexts.

When people operate, utilize or control technological systems there is always some type of interface involved, which functions as a display of information on the current state of the technological system and how to operate it. In an ideal world these interfaces are efficient, easy to use, aesthetically pleasing, intuitive and they evoke positive emotions.

However, these attributes are considered and implemented to a greater or lesser extent in today's systems resulting in more or less severe consequences. In the case of online shopping for instance, malfunction of an interface can result in dissatisfied users and thus decrease sales. When controlling airplanes, a bad interface design can have more serious, even life-threatening consequences.

To reduce errors and make usage of interfaces more satisfying, research about the interaction between a human being and a technological system needs to be conducted. If there is knowledge about the benefits and shortcomings concerning the usage of an interface, designers and engineers can use this knowledge to improve and optimize the means of interaction.

This investigation of human-machine interaction (HCI) benefits to a great extent from psychological research methods, as these methods have been developed to explore and assess people's behaviors, attitudes, emotions and cognition. Within the research field of *Human Factors* these methods are applied in order to design systems that fit humans and their cognitive abilities, thus taking into account human characteristics in the design phase of any technology. Its main traditional goals are to increase health, safety and productivity in any area in which humans operate and control technical systems (Wickens & Hollands, 2000).

With the increasing presence of web technologies in people's lives, Human Factors research does not restrict itself to work-related health, safety and productivity issues, but addresses also the improvement of interaction with consumer goods, where aspects other than purely work-related ones are the focus of research (Jordan, 2002).

In this context, research focuses on people's experience of interactive products such as websites or apps and how a user's skills, knowledge, previous experience and expectations will affect the perception of the product. The major objectives of this research area called *user experience* (UX) are the improvement of products in making them more easy, efficient and convenient to use.

The relevant literature suggests a variety of qualitative and quantitative research methods to assess user experience aspects for all sorts of products. All of the methods bring advantages as well as restrictions with them. Whereas qualitative methods are especially helpful in getting insights about reasons for particular user experience evaluations, quantitative methods allow big samples and provide an easy and efficient way of comparing different products with each other.

The following thesis focuses on a newly developed quantitative research method that can be applied to evaluate technological products with respect to the experience they cause in users. The so-called *meCUE* questionnaire was developed based on a well-established and empirically validated model of user experience – the *CUE-model* (Thüring & Mahlke, 2007). This study investigates the questionnaire with the goal to apply this measurement tool in industrial settings where consumer goods are being evaluated. More precisely, the focus of the following research is the questionnaire's validity, which means that it is empirically determined whether the questionnaire measures what it is supposed to measure.

In the following chapters, first, a comprehensive literature research illustrates the questionnaire's practical relevance (chapter 2). Then the development of the meCUE is reviewed (chapter 3) in order to embed this research in a broader context and build upon it. For the purpose of comparing the meCUE to well-established UX measurement tools and consequently assess the meCUE's validity, other related questionnaires are selected and reviewed in chapter 4. The relevant literature research and the theoretical foundations of the meCUE-model form the basis for the research questions, which are introduced in chapter 5. As the meCUE is supposed to be applied not only in academic

contexts, but also in research and development of interactive consumer goods, this study 'tests' the questionnaires ability to evaluate already existing mobile phone applications. For the purpose of selecting relevant apps as independent variables, an expert review is presented and analyzed in chapter 6. The hypotheses that are derived from the expert review's results are empirically investigated in a laboratory study with 24 participants (chapter 7). Eventually, results and methodology of this thesis are discussed in chapter 8.

2. State of Research

The following chapter gives a brief overview of user experience research and application. It describes how user experience evolved from the well-known concept of usability and how the two concepts differ. The scope and rationale of user experience is defined and it is shown what implications UX has when applied in industry and how practitioners can benefit from quantitative methods that assess UX.

2.1. User Experience vs. Usability

In recent years, research on the interaction of people with technical devices has shifted from a problem-oriented, instrumental focus to a more holistic view of the experience people make before, during and after the interaction (Bargas-Avila & Hornbæk, 2012; Glanznig, 2012; Scapin, Senach, Trousse, & Pallot, 2012; Clemmensen, Hertzum, Yang, & Chen, 2013). Early research on HCI merely focused on the usability of a system, which can be defined 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" (*ISO 9241-11*, 1998). The main research goal was to reveal malfunctions of an interface and improve its ease of use. Usability research methods mainly served as "interface debugging tests" (Nielsen & Landauer, 1993) and attention was devoted to the users' knowledge and skills and the time it takes to learn how to use a system (see e.g. usability design principles in: Norman, 2002). This approach implicitly expected that technology's main function is to help accomplishing work-related or instrumental goals.

On the contrary, the rather new focus on user experience takes a more holistic stance on HCI and considers non-instrumental qualities to be equally important to the interaction. Several attempts have been made to define the effects that compromise these non-instrumental qualities and consequently to define the determinates of a satisfying user experience. Approaches include hedonic qualities (Hassenzahl, Burmester, & Koller, 2003), aesthetics and visual attractiveness (Lavie & Tractinsky, 2004, Moshagen & Thielsch, 2010), enchantment (McCarthy, Wright, Wallace, & Dearden, 2006), emotions (Forlizzi & Battarbee, 2004) or engagement (O'Brien & Toms, 2008) to name but a few. Although many attempts have been made to describe and compose the term, there is no consensus on the definition of UX. The following

chapter offers an overview of the state of the art of UX and defines its rationale and scope.

2.2. The Scope of User Experience

User experience is the umbrella term for the research area that concerns itself with the investigation of peoples' experience with interactive technologies. In academia, it is with the research about the elements and characteristics of this experience and the psychological constructs it incorporates. In industry, UX methods are applied in research and product development with the goal of making products better in one way or another.

In order to define and analyze user experience in academic research, mainly two different approaches can be found: Holistic and reductionist approaches (Riedel, 2013). Both concepts provide conceptual frameworks and classify different components. Holistic models view user experience in its entirety and lay the focus on its uniqueness (see for example Kakar, 2012). However, this often results in the impossibility of empirical assessment or falsification of the models (Minge, 2011). On the contrary, researchers that follow reductionist approaches try to narrow down the concept of user experience to its core components in order to make it measurable. Examples of this approach are the two-component-model (Burmester, Hassenzahl, & Koller, 2002) or the CUE-model (Thüring & Mahlke, 2007) which forms the basis of the research in this thesis.

As the relatively new concept of UX is based on multidisciplinarity, it lacks a common definition and standardized, universal research methods (Glanznig, 2012; Scapin et al., 2012). Apart from numerous qualitative research methods (such as personas or contextual inquiry) which are beyond the scope of this thesis, there are several validated questionnaires that determine different aspects of UX. For example hedonic and pragmatic qualities can be assessed with the AttrakDiff (Hassenzahl, Burmester, & Koller, 2008), visual aesthetics with the VisAWI (Thielsch & Moshagen, 2011) or with the Aesthetics Scale (Lavie & Tractinsky, 2004). Emotions can be assessed with the PANAS (Tellegen, Watson, & Clark, 1988), the SAM (Bradley & Lang, 1994), the Emocards (Desmet, Overbeeke, & Tax, 2001), the PrEmo (Desmet, 2003) or the LEMtool (Huisman, van Hout, van Dijk, van der Geest, & Heylen, 2013). For general UX,

amongst others, the ServUX (Väänänen-Vainio-Mattila & Segerstahl, 2009) or the UEQ (Laugwitz, Schrepp, & Held, 2006) can be applied.

Since user experience strongly relates to product development, these academic models and methods should be adapted for application in industrial contexts. However, the practical implementation of UX lacks unity and a clear definition, too. Consequently, as UX is a multidisciplinary field, its adaptation and application mainly depends on the individual background of the practitioner as well as the product category itself. Although user experience is always applied in order to improve products, its scope can take very different forms. From a psychologist's point of view for example, user experience can be derived from knowledge about people's cognition (Weinschenk, 2010). Coming from interaction design, user experience is something that can be designed for by following best practice user interface design patterns. As Sharp, Rogers and Preece (Sharp, Rogers, & Preece, 2007, p. 15) put it: "[…] one cannot design a user experience […], but only create the design features that can evoke it". In product management, UX can be one of several product features or *KPIs* (key performance indices) that need to be measured (Halalka, 2013).

Despite this disagreement about the scope, stakeholders in industry seem to agree on the fact that user experience helps making products better in one way or another. Moreover, user experience has become a quality aspect of products (Abbasi, Lew, Rafique, & Li, 2012) as it places the user in the center and contributes to making their tasks and goals more easily accomplishable when interacting with the corresponding system. As Mahlke (2008, p.23) states: "User experience takes an entirely user-oriented perspective on human-technology interaction. The user´s perspective on the quality of the interaction is the ultimate criterion."

Following such a user-centered approach in product development can have a variety of advantages in all kinds of fields. Applying UX to industries such as medical technology for example can result in better patient experience and increasing therapy rates (Moynihan, Paul, & Markus, 2013). In e-commerce, user experience methods contribute to more usable online shops and improved functionalities, with the goal to make shopping more easy and pleasurable (Knijnenburg et al. 2012, Yu & Wu, 2010) whereas in gaming industry user experience principles (here called *player experience*) are applied to make tasks more challenging. As games are played for their hedonic

value, here more emphasis is put on emotions and non-instrumental product perceptions (Nacke & Drachen, 2011).

However, implementing UX methods in product development also comes with diverse challenges. The following chapter gives an overview of what problems practitioners face when trying to apply UX methods in fields where interactive systems are designed and built and how an UX-questionnaire can be useful in these settings.

2.3. User Experience in Industry

The above-described development from usability towards user experience did not just have an impact on scientific research, but also implied changes for professionals in the field. One example of this is the recent change of name from the international 'Usability Professionals Association' (UPA) to 'User Experience Professionals Association' (UXPA) in June 2012 (Gunther, 2012).

As in academia, the field of UX as a profession lacks a common identity and stakeholders do not agree upon whether user experience is an enhancement or restriction in comparison to usability (Ardito, Buono, Costabile, & Lanzilotti, 2012). Despite the controversial discussion about the self-image of usability/user experience professionals, the actors of the field have to face new challenges that come with the shift. For once, methods that assess the ease of use or ease of learning of a system are not sufficient enough anymore. User research is no longer about finding bugs and making software intuitive alone. User experience requires a much richer understanding of the user and her interaction with interactive systems. Hence, new methods and metrics have to be developed to comprehend this broader scope and the consequences it has for software and product design (Law, Abrahão, Vermeeren, & Hvannberg, 2012).

Although there are still software and product design companies that do not systematically apply usability and user experience methods in their product development, there is at least a growing interest (Ardito et al., 2012). More and more businesses understand that a positive user experience can be the reason customers choose their product over the competitors' alternatives.

The problem that companies face however is a helplessness about how to efficiently integrate user research methods into their daily business (Lallemand, 2011). The

reasons for these circumstances are manifold. For example Ardito and colleagues (2011) conducted a survey among software developers and found out that Human–Centered Design (HCD) methods in general are perceived as being not suitable and too resource demanding. Summarizing their research, the authors state that the lack of HCD methods in software development processes are due to three reasons: "1) time and costs of the HCD methods; 2) cultural prejudices [and] 3) lack of frameworks guiding the software development team in applying HCD methods" (Ardito et al., 2012).

Especially the final argument – the lack of available guidelines for efficient user research – seems to be a major reason for the absence of UX methods in many product development processes. Based on the results of a research among 35 participants coming from academia and industry, Roto and colleagues (2009) point out the importance of fast, simple-to-use and resource-conserving quantitative approaches that "allow repeatable and comparative studies in an iterative manner". According to the authors this is "especially important in the hectic product development cycle in industry, but also in design research that needs effective evaluation tools for quick iterations" (Roto et al, 2009, p. 4).

Qualitative UX methods (for example user interviews, personas or contextual inquiry) that are often applied in early product development phases (Roto et al., 2009) hardly meet these requirements. Based on their subjective nature they do not qualify well for repetition and comparison. Additionally they often require more time and effort than quantitative methods.

Although quantitative methods like questionnaires cannot and should not replace qualitative user research methods in product development processes, they provide an easy and resource-efficient way to assess the valence of user experience that an interactive product evokes (Väänänen-Vainio-Mattila & Segerstahl, 2009). This is especially valuable when it comes to tracking changes in the interaction over a longer period of usage or when comparing different versions of a product (Minge, Riedel, & Thüring, 2013). As Law and colleagues (2012, p.1) state: "both usability and UX measures should enable professionals to benchmark competitive design options".

Law and colleagues (2012) believe, however, that incorporating UX methods in the early product development is even more difficult than integrating usability methods, because feedback from UX testing is often difficult to interpret. Designers face the

challenge of translating 'experiential problems' into improvement requirements without knowing whether the issues identified in the test are simply due to individual user characteristics such as mood for example. In addition to that, the authors question the authenticity of evaluating experiential phenomena with early low-fidelity prototypes. In order to effectively integrate UX evaluation results in the next development cycle, designers need to be convinced of the urgency to fix the issues. The authors raise the question whether UX evaluation feedback is less convincing to designers and developers than usability feedback due to the above-mentioned reasons.

For companies that apply agile product development it seems to be especially challenging to reconcile the techniques of UX design with the pace of their processes. In contrast to the more bureaucratic traditional software development, agile approaches such as *Scrum* promote rapid and flexible responses to change and place people and "working software over processes and comprehensive documentation" (Fowler & Highsmith, 2001, p. 2). While the 'lightweight' iterations and 'incremental mini-releases' that determine the pace and rhythm of agile product development would provide ideal milestones to integrate user experience feedback, practitioners have difficulties in finding opportunities to obtain that feedback (Ferreira, Sharp, & Robinson, 2012). Main reasons for this are the contradicting claims of agile processes that refrain from documentation on the one hand and traditional user research methods that heavily rely on detailed formal documents which communicate findings to the development teams on the other hand (Sy, 2007). Additionally, as agile iteration cycles take two to three weeks on average, user experience research consumes too much time and resources.

A rather new approach that tries to solve these problems and which provides a framework for effectively intertwining agile product development and UX can be found in the framework of *Lean UX* (Gothelf & Seiden, 2013). The authors encourage UX professionals to "get out of the deliverables business" (p. 12) and provide principles and methods to make UX more flexible and resource-efficient. As Lean UX is a rather new concept, there is no research on the advantages and disadvantages of its application in practice available yet and it is still to be clarified whether or not it offers a successful approach to effectively fuse product development and UX methods.

In conclusion it can be said that a variety of obstacles withhold practitioners from successfully applying UX methods. Although the belief in the necessity of good user experience design increases, the field still has to face objections and prejudices. One of the major criticisms of UX as a profession is that 'one cannot put a number on' user experience, meaning that its return on investment (ROI) is not measurable (Gube, 2010). This is true since the effects of changes in user experience are not directly linkable to KPIs such as conversion rates. However, quantitative methods like the meCUE can be used for quantifying the effects of, for example, design changes and consequently help to make better decisions concerning the user experience of products.

The literature research described above revealed a variety of already existing quantitative methods for assessing UX and this raises the question whether yet another method is needed. Indeed, the mentioned questionnaires provide reliable measurement tools for the different components of user experience. They fail, however, in depicting a complete picture of the central aspects that contribute to users' evaluations of interactive systems. Even questionnaires as AttrakDiff and UEQ, which include several aspects lack the assessment of emotions and consequences that arise from interaction.

A tool, that defines and incorporates the major facets of user experience (hedonic and pragmatic qualities, emotions, aesthetics and general UX) would not only close a research gap, but would also be of special interest for particular areas of application.

The following chapter gives a brief overview of the development of the meCUE and the theoretical basis underlying it.

3. Review on the Development of the meCUE Questionnaire

As the preceding construction of the tool is not within the scope of this study, the theoretical framework underlying the questionnaire and the analysis of its quality criteria will not be discussed in detail. However, in order to understand the content of the questionnaire and to embed the research of this thesis into a broader context, a short review of the preceding studies is given as follows.

Based on an established and empirically validated model concerning the aspects that contribute to experience people make (Thüring & Mahlke, 2007) when interacting with a technological product, Minge and Riedel (2013) developed a questionnaire that assesses these aspects – the *meCUE* questionnaire. According to the so-called *CUE Model* (Components of User Experience) by Thüring & Mahlke (2007), consequences - as for example the intention to use a product - arise from the perception of instrumental as well as non-instrumental qualities of the product. These are in turn mediated by the emotions the product evokes. System and task properties and user characteristics have a major influence on the interplay between the product qualities and emotions. For each of these dimensions (*non-instrumental, instrumental qualities, emotions* and *consequences*) Riedel (2013) generated a set of items, which led to a first version of the questionnaire consisting of 67 items.

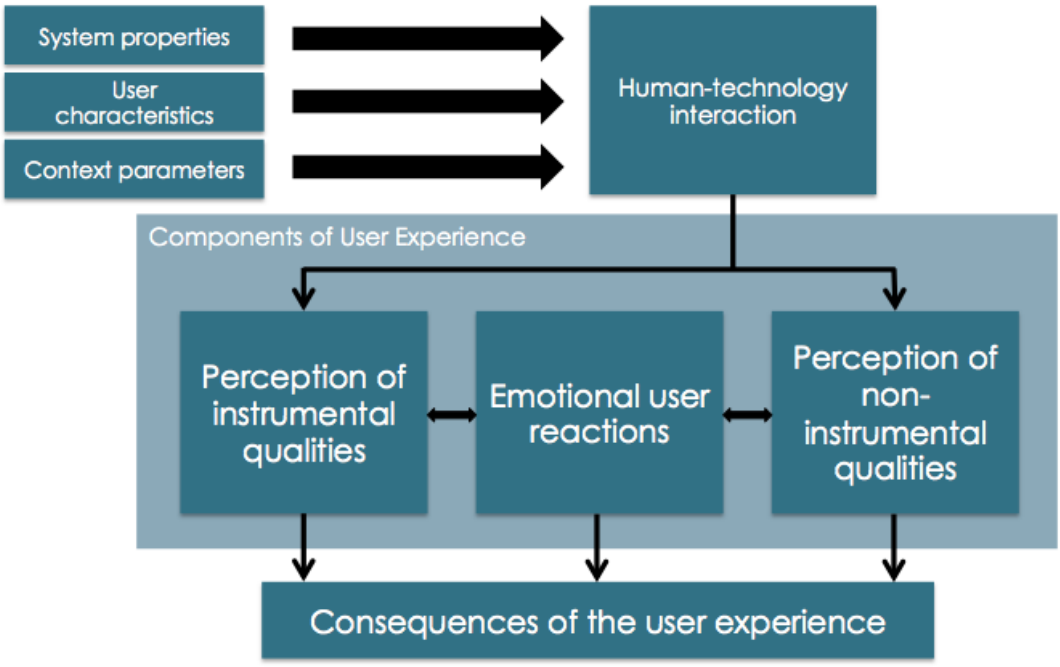

Figure 1: The CUE Model (adapted from Thüring & Mahlke, 2007, p.262)

In two consecutive online studies, each time 238 participants were asked to evaluate a technological product they had in mind by filling out the questionnaire. Using a 7-point-Likert scale, participants could rate their extent of approval and/or rejection of statements such as "the design of the product is attractive". A series of principal component analyses and factor analyses uncovered five independent factors in the category *product perceptions* (effectiveness, efficiency, visual aesthetics, status and commitment), two factors in the category *emotions* (positive emotions and negative emotions) and two factors in the category *consequences* (product loyalty and intention to use). The items with the highest factor loadings were chosen to be included in the questionnaire, which resulted in 34 items in the current version (see appendix A).

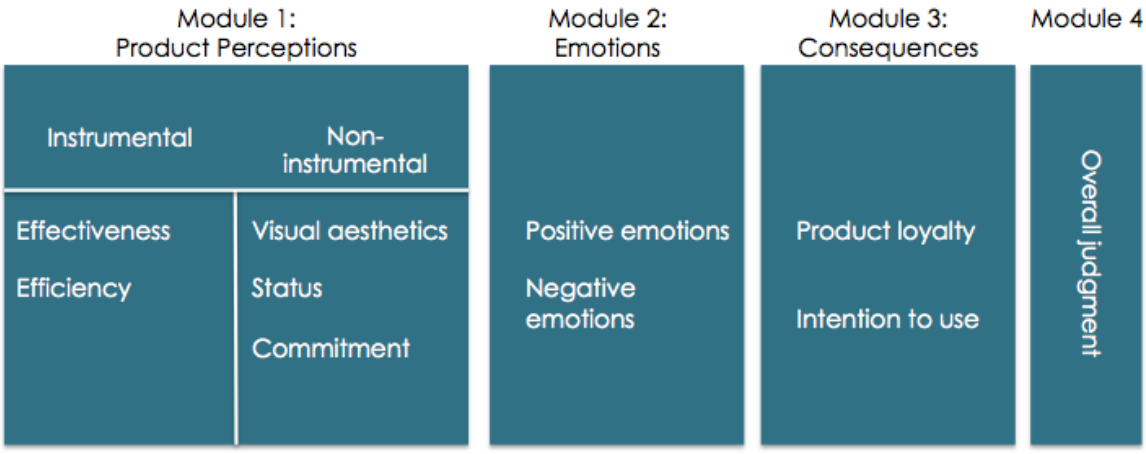

Figure 2: Modules of meCUE questionnaire (adapted from Minge & Riedel, 2013, p.3)

In a further laboratory study, 67 participants evaluated text editing software, a portable mp3-player and a mobile phone. In doing so the study reproduced the former determined factors and confirmed that the meCUE provides reliable and valid results.

All of the scales produced a high internal consistency with Cronbach's Alpha between .76 and .94 (Minge et al., 2013). Additionally, a number of validated questionnaires were used to investigate criterion-related validity and significant correlations with these tools were found. In another lab study, 31 patients evaluated lower limb orthoses (Doria, Minge, & Riedel, 2013), and the meCUE found similar patterns of significant differences as the well-established AttrakDiff questionnaire and the PANAS did.

Based on the underlying theoretical framework and its powerful empirical validated construction, the meCUE offers some major advantages over existing quantitative user research methods. As Riedel (2013) found out, many of the existing user experience

methods lack an empirical basis and further more concentrate only on single aspects of user experience. The meCUE-questionnaire closes this gap by providing a method that assesses many aspects of user experience, including emotions and consequences. Together, the modules provide a comprehensive user experience evaluation, but they can also be applied separately.

To make sure that in future the meCUE can be applied widely, further validation studies need to be conducted. It is necessary to investigate if the questionnaire produces consistent results independently from the to-be-evaluated object, researcher and situation. The following chapter provides a brief overview of the rationale behind the selected questionnaires for comparison with the meCUE.

4. Selection of Related Dimensions and Questionnaires

As described earlier, one major advantage of the meCUE over other well-known questionnaires is the incorporation of manifold dimensions, which are of importance for user experience. In fact the initial wording and selection of items for the assessment of the components of the CUE-model were based on and inspired by the most often applied and best-known measurement tools in the user experience field (Riedel, 2013). Consequently, these questionnaires qualify well for the comparison of scores and the establishment of construct and discriminative validity for the meCUE. In order to understand the affiliation of these questionnaires with the meCUE, a brief overview of the relationship is given below.

The division of product perceptions into instrumental and non-instrumental qualities in the CUE-model follows a similar line of argument as Hassenzahls Two-Component Theory (Hassenzahl, 2007) of pragmatic and hedonic qualities. Hassenzahl, Diefenbach and Göritz argue that usability and utility together can be described as pragmatic qualities of a product as opposed to hedonic qualities that arise from „a product´s potential to support pleasure in use and ownership" (2010, p. 357). This understanding of self-referential product qualities can be found in the model underlying the meCUE as well. The subscales that Riedel (2013) describe as *status* and *commitment* measure related constructs as those in hedonic product qualities with status being a product's possibility to add to the social identity of the user and commitment referring to the relationship between user and product (Riedel, 2013).

In Hassenzahls (2007) opinion the two concepts – hedonic and pragmatic qualities - are uncorrelated. Together they form an overall judgment of a product, which he describes as *attractiveness*. This model is operationalized in the repeatedly validated and well–known questionnaire AttrakDiff (Hassenzahl et al., 2008).

Besides self-referential product qualities the visual aesthetics of a product also contribute to non-instrumental product perceptions in the model of the meCUE. The items for this subscale were based on the rather new questionnaire VisAWI (Thielsch & Moshagen, 2011) that was developed to assess visual aesthetics of websites. Consequently the VisAWI qualifies well as a reference point for the quality of meCUE's visual aesthetics subscale.

The UEQ (Laugwitz et al., 2006) also served as an example for the wording of the items for the dimension *product perceptions*. Like the AttrakDiff it assesses pragmatic and hedonic qualities but puts with three subscales (perspicuity, efficiency, dependability) more emphasis on the pragmatic qualities than the AttrakDiff does. The inclusion of the UEQ in the main study can therefore add towards a more sophisticated analysis of the dimension product perceptions.

The items for the module *emotions* were developed on the basis of the Circumplex-Model (Russell, 1979) and are supposed to assess two dimensions of emotions: arousal and valence. The resulting four manifestations of emotions – high arousal and positive valence, high arousal and negative valence, low arousal and positive valence, low arousal and negative valence – could be successfully verified by factor analysis in in the initial development study (Riedel, 2013). Later studies (Minge & Riedel, 2013), however, were not able to reproduce the four manifestations of emotions, but revealed only two underlying factors with high factor loadings on positive and negative emotions. Consequently Minge and Riedel decided in favor of two subscales (positive and negative emotions) in the module *emotions*. Correlations with the related assessment tools SAM (Bradley & Lang, 1994) and LEMtool (Capota, van Hout, & van der Geest, 2007) in the initial study indicated a sufficient criterion validity of the module *emotions*. Since the SAM and LEMtool are however non-verbal instruments, this study takes the opportunity to compare the items to another verbal emotions assessment tool to further support criterion validity. The PANAS (Tellegen et al., 1988) was chosen for this purpose as it is well-known and well-validated and served as a source of inspiration for the item wording of the meCUE (Riedel, 2013).

Factor analysis of the module overall judgment, which initially contained several items, proved to be difficult to analyze (Minge et al., 2013). Minge, Riedel and Thüring (2013) consequently suggest assessing overall judgment by a single item. According to Riedel (2013) overall judgment should be seen as a global positive or negative product evaluation that rates the product as a whole. Similar concepts can be found in the AttrakDiff and UEQ, in which the subscale *attractiveness* is supposed to assess overall product evaluation. In the current version of the questionnaire, a 21-point- scaled adjustable slider that allows for a more precise analysis of the global product evaluation, operationalizes the module overall judgment (see appendix A).

Currently, no questionnaires and assessment tools can be found that measure similar concepts as the module consequences with its subscales product loyalty and intention to use. Chapter 8 discusses potential external criteria to investigate the validity of this module in future studies. An overview of the above-described questionnaires that were selected for the main study is given in table 1.

5. Objectives and Research Questions

The aim of this thesis is to investigate whether the meCUE accurately measures the constructs it is supposed to measure and how well the questionnaire can be applied in real-life settings. This quality criterion of test instruments is known as *validity* and can be defined as „the adequacy of a scale as a measure of a specific variable“ or the „extent to which a measure reflects the intended phenomenon“ (Dooley, 2001, p. 88). Validity is a rather broad term containing various specific types each describing the quality of different facets of a psychological measurement tool. As described earlier, one of the essential application-scenarios for the meCUE in academia and industry is the assessment of differences in user experience aspects over time or between different products. Having this in mind, the main focus of the study is on the ability to distinguish between products of different quality (discriminative validity, Fawcett, 2008). In addition to that, the meCUE will be compared to other well-established measurement tools of related constructs (convergent validity, DeVellis, 2003) and relevant external criteria (criterion validity, Kaplan & Saccuzzo, 2005).

So far, the questionnaire has been used to mainly evaluate physical technological devices such as portable mp3-players, mobile phones and medical technology. As it is likely that practitioners will apply the meCUE to evaluate digital interactive consumer goods, this study focuses on the question of whether the meCUE qualifies to examine the experience of mobile app users. Apps for route planning purposes with public transport were chosen for this study.

The empirical study of this thesis systematically analyzes the following research questions (RQ):

RQ 1: Is the meCUE able to detect expected differences on pre-determined dimensions? (Discriminative validity)

RQ 2: Do ratings on the meCUE correlate with other validated questionnaires? (Convergent validity)

RQ 3: Do ratings on the meCUE correlate with external criteria? (Criterion validity)

6. Study 1: The Expert Review

One aim of this study is to bring as much realism and ecological validity into the lab experiment as possible. Therefore, in the main study (see chapter 7), the questionnaire was applied to evaluate real products that are already available on the market. However, one disadvantage of this approach is the impossibility to manipulate the independent variables as the apps have a certain aesthetic appeal and usability that is out of the influence of the researcher. In order to solve this problem, an expert review was conducted prior to the lab experiment. This step was supposed to help identifying the apps with the most difference on both the dimensions aesthetic appeal and usability and to establish hypotheses about them.

6.1. Method

The expert review comprised two major steps. First, six public transport apps for navigation around Berlin were chosen from the iOS appstore. Based on her own expert opinion and customers' ratings in the iOS appstore, the researcher selected six apps that best suited the experimental purpose. The second step was the evaluation of the apps by four experts in order to prove the researcher's assumptions and to provide qualitative and quantitative data about the aesthetic appeal and the usability of the apps.

6.1.1. Participants

The experts were four German usability professionals with several years of experience in academia or industry. At the time of the study, two of them worked as researchers in the usability and user experience field at the *Institute of Technology* in Berlin (*Technische Universität Berlin*). One was a usability expert at a medical technology company and one a user experience researcher at an UX-consulting agency. It was assumed that the experts were familiar with the product characteristics as well as the evaluation criteria like best practice interaction design and usability heuristics. All four were smartphone users; two possessed iOS smartphones and two used Android smartphones.

6.1.2. Material

As described above, six iPhone apps from the iOS appstore were chosen for the expert review (see figure 3 and figure 4). All of them were for local route planning purposes

with public transport around Berlin. Some of them could also be used for long-distance travelling in Germany. Searching for apps on the iOS appstore revealed that there are mainly two different types of search result pages in public transport apps. On the one hand there are apps that use simplistic table-like information depiction listing connections (see *DB Navigator*, *iFahrinfo* and *Fahrplan* in figure 3). On the other hand there are apps that provide visual search results of possible connections, using size and color in a meaningful way to indicate duration and type of route (see *FahrInfo*, *moovel* and *Waymate* in figure 4). When choosing the apps for the expert review, it was made sure that both types of information depiction are equally often represented in the sample.

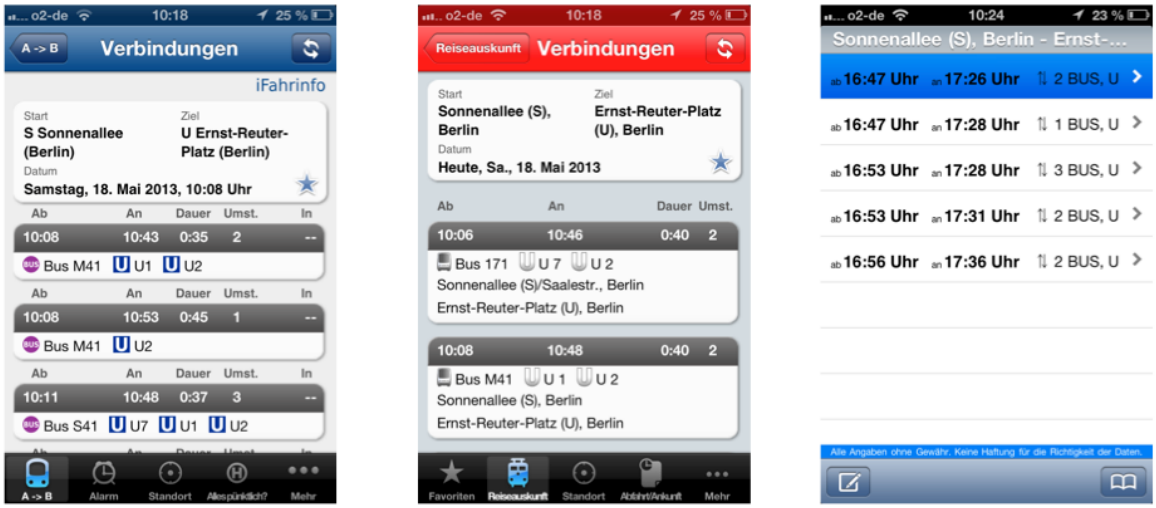

Figure 3: Screenshots of public transport apps with list search result page.
From left to right: iFahrinfo, DB Navigator, Fahrplan

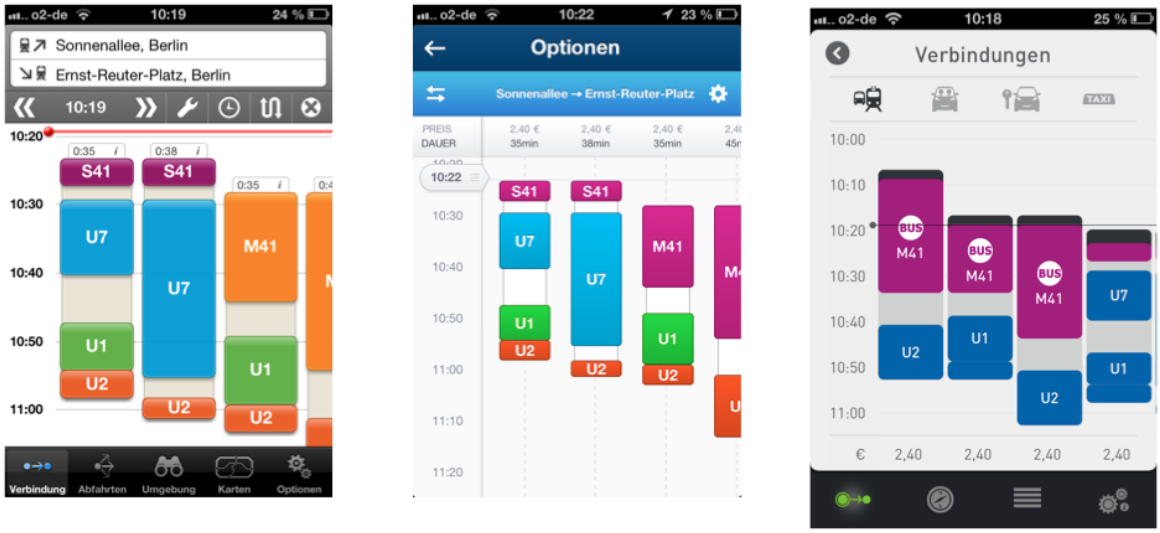

Figure 4: Screenshots of public transport apps with visual search result page.
From left to right: FahrInfo, Waymate and moovel.

All six apps provided functions to enter start and destination and set the time and date of either departure or arrival. They differed, however, in the quality of applied interaction patterns as well as the number of features that were provided. One of the apps, *Fahrplan*, had a bug that made it impossible to get details about specific connections, meaning that it showed a blank screen after clicking on one of the listed connections. As this functional imperfection was expected to give valuable insights in the perception of effectiveness and consequently (in case the app was chosen for the main study) the questionnaires potential to assess this perception, it was included in the sample.

Screenshots of the main output screen were provided for the aesthetic judgment on a photo paper print in 11.5 cm x 5.5 cm (size of iPhone 4). Moreover, the apps were installed on an iPhone 4S, which was given to the experts during the review session. Additional material included thumps-up and thumps-down icons on photo paper, a 20 cm x 11 cm sized cardboard template of an iPhone and paper snippets with the words „Usability“ and „Design“ on it (see figure 5).

6.1.3. Procedure

Each expert was asked to rate the apps in an individual one-hour evaluation session. In the beginning the expert was given the printed screenshots and was asked to give her judgment about the visual aesthetic of the app by means of a single item („the app has an attractive design“; 1= totally agree, 7 = totally disagree). There was no interaction with the app at this point but the expert could look at the photo and share her thoughts without any restrictions in time or scope. Each app was being rated directly after viewing it.

The next step contained two different use cases with tasks (see appendix B), in which the expert had to find her way from A to B with the apps and evaluate the apps concerning their usability. The goals of the tasks were to find out about the exact time of arrival in the first use case and the platform the train would arrive at in the second use case. Due to the above-mentioned bug, the successful completion of the second use case was not possible with *Fahrplan*. The apps were provided on an iPhone 4S and were shown in the same order as in the first step for each expert, but were counterbalanced over the four different experts. During the phase of interaction the experts were asked to think aloud and the researcher took notes on positive, negative

and neutral utterances of the expert. After carrying out the two use cases on an app, the app was rated by the expert on a single item („the app has a good usability“; 1= totally agree, 7 = totally disagree).

Once the app was evaluated concerning design and usability, the next step for the expert was to put the apps into an individual order from best overall judgment to worst overall judgment. In order to do so, the printed screenshots and thumps-up and thumps-down icons were used (see figure 5).

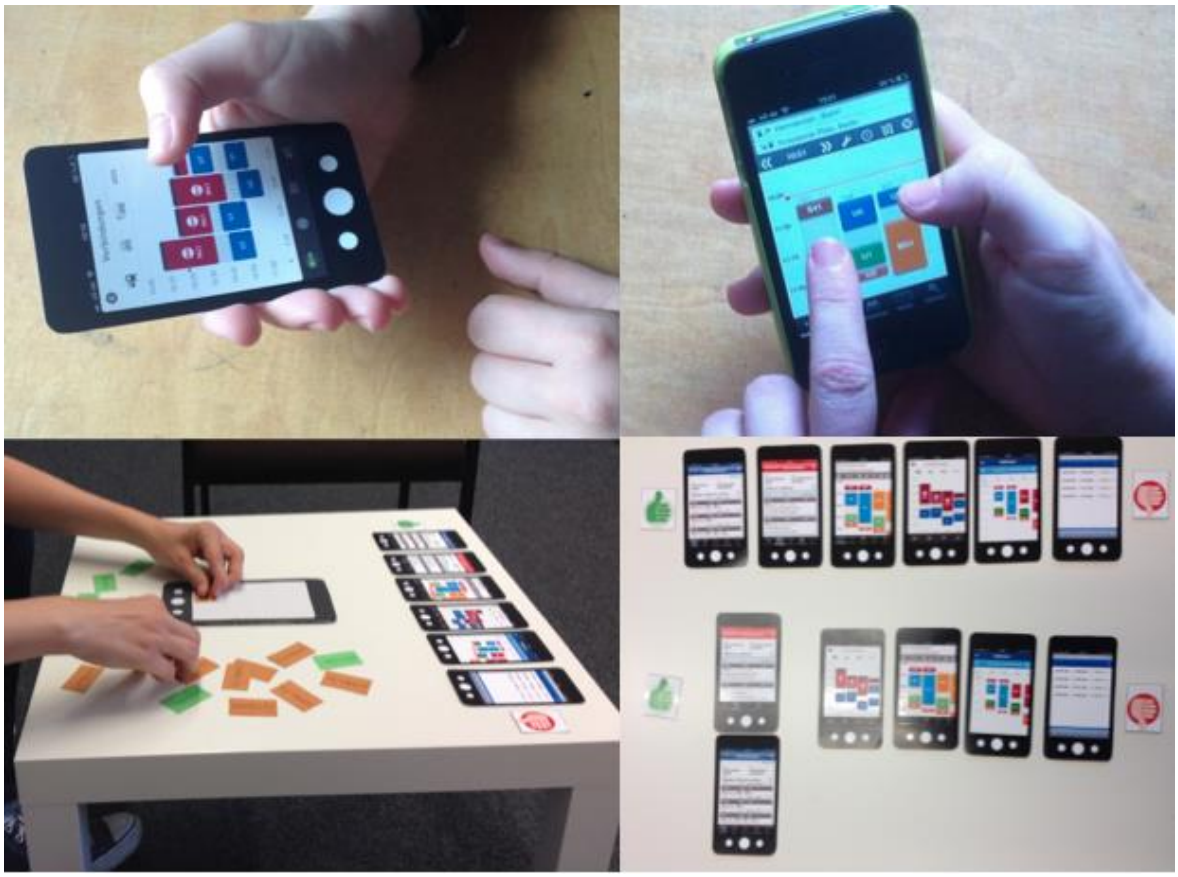

Figure 5: Impressions from the expert review:
Printed screenshot for the design evaluation (top left), interaction with the app (top right), weighting of the attributes *usability* and *design* (bottom left),
comparison of calculated rating and the rating by the experts (bottom right)

To get more insights about the expert's assigned relation of importance of design and usability to the apps, the next step contained a weighting of these attributes. Following the „coloring the black box“ approach, introduced by Pohlmeyer (2011), the expert filled an iPhone shaped cardboard template with colored paper snippets representing the attributes *usability* and *design*. The template could fit exactly ten paper snippets, leaving it up to the expert how many design- and usability-snippets she would use.

During this process, the researcher emphasized that overall appeal is shaped by more attributes than aesthetic appeal and usability alone, but that for this approach it was necessary to leave out other characteristics for a moment and focus on just the relation of these two attributes in the apps.

The multi-attributive rating was followed by a short break in which the expert could rest for a moment and the researcher calculated a weighted attribute-score for each app by multiplying the results from the single-item with the weight that was assigned to the attribute. The two scores, one for design and one for usability, were then summed up to an overall score for each app. From these scores the researcher calculated a ranking for the apps and confronted the expert with it. The calculated rating was placed next to the rating generated by the expert in step five (see appendix B) so that the expert was able to comment on possible deviations. The expert was also given the possibility to adjust her overall judgment from step three after having seen the weighted order.

6.2. Results

Qualitative data from the thinking aloud procedure as well as quantitative data from the single item questionnaires were analyzed. When interacting with the apps, the experts identified various usability issues. Qualitative analysis revealed that the experts hardly agreed on the severity of the discovered issues. Three of them stated for example that the icons used in *Waymate* are not clear and are consequently not intuitive. One of the experts stated that this was the main reason to rate this app as having a comparably low usability. Consequently, the qualitative data was used to get a better understanding of variation in the usability and design scores and helped the choice of the apps for the main study.

Figure 6 shows the mean scores of the single item design and the single item usability for all six apps. Based on the small sample size (n = 4) and the explorative purpose of the expert review, quantitative data was mainly used to detect maximum and minimum differences between the apps on the determined dimensions as well as a ranking of the apps. Exact mean scores and standard deviations were of lesser interest for the above-mentioned reasons.

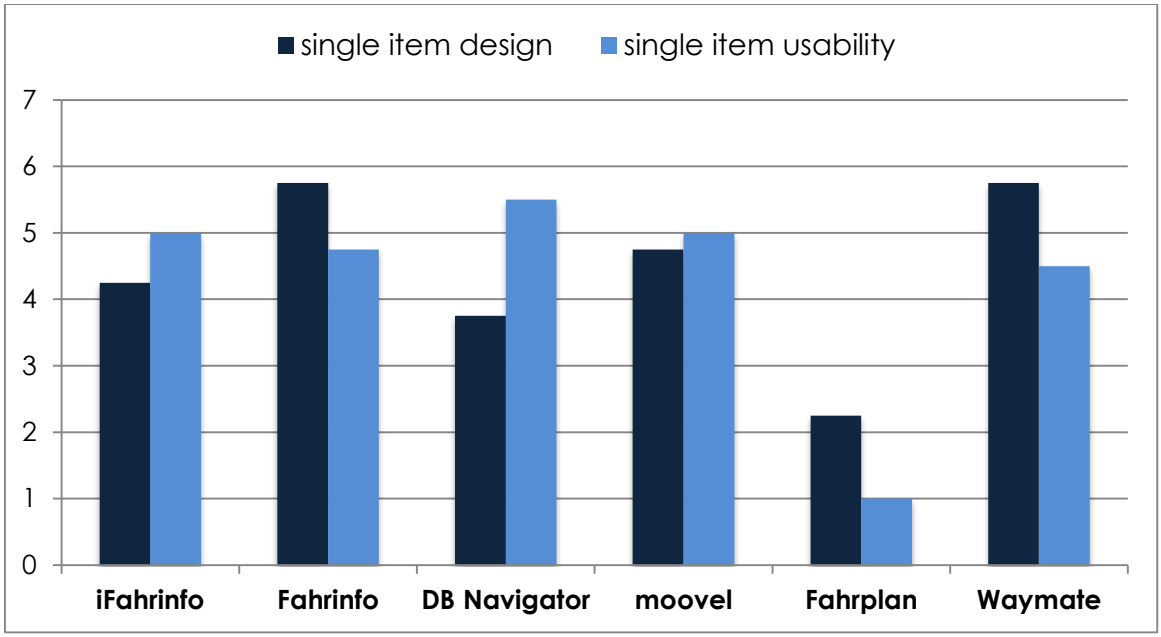

Figure 6: Results for single-item ratings on design and usability
(1= totally disagree; 7= totally agree, N=4).

The results show that the apps *Waymate* and *Fahrinfo* scored highest on the design dimension and the app *DB Navigator* got the highest scores on usability. *Fahrplan* scored lowest on both design and usability. The weighting of the attributes design and usability revealed that for all four experts usability (M=7) was more important than design (M=3).

Multiplying the single-items with the weight of the attributes for each expert and summing up the weighted design and usability scores calculated an overall ranking of the apps. For clarification, the formula used is:

Overall score = (s-i usability $expert_1$*weight usability $expert_1$) + (s-i design $expert_1$*weight design $expert_1$) + ... + (s-i usability $expert_n$*weight usability $expert_n$) + (s-i design $expert_n$ *weight design $expert_n$).

s-i = single-item

Out of the highest possible score of 280, the three apps *FahrInfo*, *DB Navigator* and *moovel* each got an overall score of 198, followed by *Waymate* with an overall score of 192 and *iFahrinfo* with an overall score of 188. *Fahrplan* scored comparatively low with an overall score of 55.

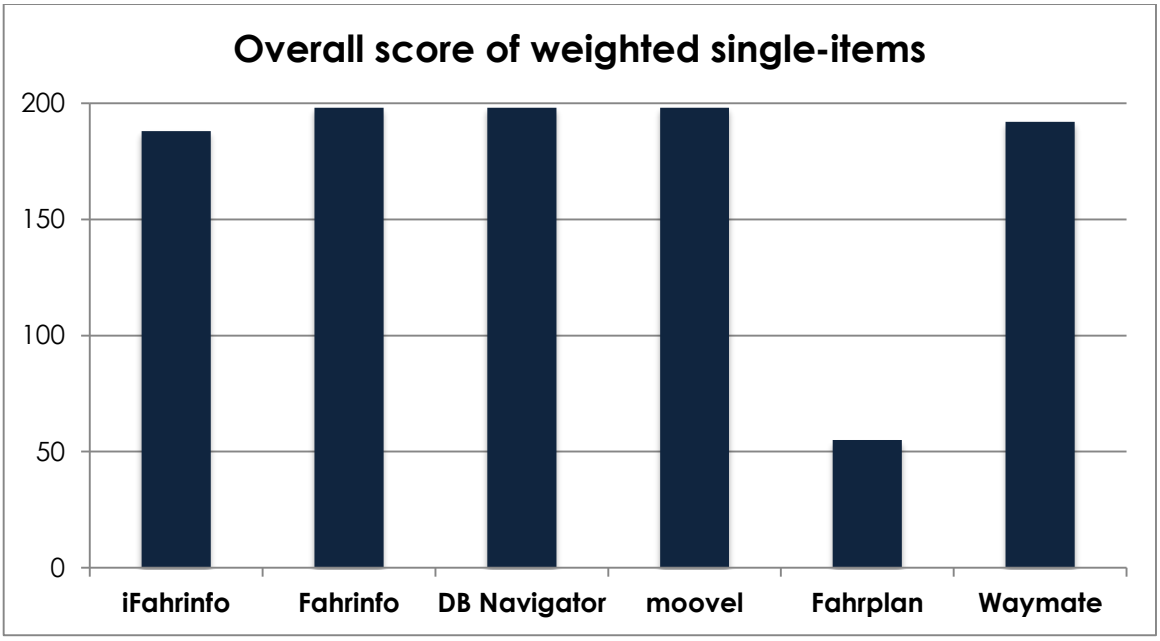

Figure 7: Overall scores of weighted single-items for the six apps.

Comparing the calculated ranking of each expert to the expert´s subjective ranking order revealed some minor deviations that the experts explained by putting more emphasis on a feature they liked or disliked.

Some design and usability issues such as inconsistent use of icons could be defined from the thinking aloud protocol and were also taken into consideration when deciding for the three apps for the lab study.

6.3. Summary and conclusion

Four usability experts were asked to give their opinion on the quality of design and usability of six public transport apps. The quantitative results show that the apps differ systematically in these dimensions. The qualitative results, however, reveal that subjective opinion to a great extent also contributes to the ratings of the apps. Usability issues that were perceived as severe by some experts did not seem to bother other experts. Similar disagreements could be discovered for the opinions of the apps' design. Here, the expert review failed to provide concrete criteria for the evaluation of design and judgments were rather made based on general taste of colors and fonts.

Therefore, several other criteria have also been taken into account in order to choose the apps for the main study. One requirement was maximum distances on the usability as well as the design scales, meaning that the apps with the lowest and the highest ranks were taken into consideration. Degree of consensus among the experts

concerning the evaluation was taken as another criterion since subjective opinion accounted for some variation in the ratings.

The results clearly showed that *DB Navigator* and *Fahrplan* qualify well for the main study, because they differed the most in their degree of usability. Moreover, *Fahrplan* was also rated as having the worst design. Due to a bug, *Fahrplan* did not show precise information on specific connections such as platforms or train directions. As this was expected to add to variation in the evaluation of the apps, *Fahrplan* was included in the choice. It was of special interest how the bug would affect usability related dimensions such as efficiency and effectiveness. The remaining question was, however, whether to choose *FahrInfo* or *Waymate* as the app with the best design. Based on the above-mentioned qualitative data and the higher amount of variation in the scores for *Waymate*, *FahrInfo* was selected for the main study.

7. Study 2: The Main Study

As described above, the goal of the expert review was twofold. One objective was to identify three apps that differ in usability and aesthetic appeal, so that they could be used as independent variables in the main study. In addition to that, the results of the expert review were the basis for establishing meaningful hypotheses about the independent variables (the apps), which could be tested in the subsequent main study in order to give answers to the research questions.

7.1. Hypotheses

Based on the empirical findings of the expert review, the following hypotheses were derived for each research question (RQ).

RQ 1	**Is the meCUE able to detect expected differences on pre-determined dimensions? (Discriminative validity)**	
Hypotheses	H1a	Of all three apps *DB Navigator* and *Fahrplan* will differ most on instrumental product perceptions/pragmatic qualities, with *DB Navigator* scoring higher than *Fahrplan*.
	H1b	Of all three apps *FahrInfo* and *Fahrplan* will differ most on the subscale visual aesthetics, with *FahrInfo* scoring higher than *Fahrplan*.
	H1c	*DB Navigator* will score higher on instrumental product perceptions/pragmatic qualities than *FahrInfo.*
	H1d	*FahrInfo* will score higher on instrumental product perceptions/pragmatic qualities than *Fahrplan.*
	H1e	*FahrInfo* will score higher on the subscale visual aesthetics than *DB Navigator.*
	H1f	*DB Navigator* will score higher on the subscale visual aesthetics than *Fahrplan.*

RQ 2	**Do ratings on the meCUE correlate with other validated questionnaires? (Convergent validity)**	
Hypotheses	H2a	The instrumental product perceptions of meCUE correlate positively with pragmatic qualities of AttrakDiff-mini.
	H2b	The instrumental product perceptions of meCUE correlate positively with the dimensions perspicuity, dependability and efficiency of UEQ.
	H2c	The subscale visual aesthetics of meCUE correlates positively with the VisAWI-S.

	H2d	The subscale positive emotions of meCUE correlates positively with the subscale positive emotions of PANAS.
	H2e	The subscale negative emotions of meCUE correlates positively with the subscale negative emotions of PANAS.
	H2f	The subscale status of meCUE correlates positively with hedonic qualities of AttrakDiff-mini.
	H2g	The subscale commitment of meCUE correlates positively with the hedonic qualities of AttrakDiff-mini.
	H2h	The module overall judgment of meCUE correlates positively with the subscale attractiveness of AttrakDiff-mini and the subscale attractiveness of UEQ.

RQ 3		**Do ratings on the meCUE correlate with external criteria? (Criterion validity)**
Hypotheses	H3a	The subscale efficiency of meCUE correlates negatively with task completion time

7.2. Method

The main study was conducted in a laboratory under controlled conditions. Quantitative data on the evaluation of the three selected apps were gathered by means of a within-subject design. The evaluation sessions included two use-cases that encouraged the participants to interact with the apps and form an opinion on the performance of the latter. The apps were being evaluated by filling out a series of questionnaires.

7.2.1. Participants

In total 26 individuals participated in the lab study. Two participants did not have experience with smartphone usage and were therefore excluded. Of the remaining 24 participants, eleven were female and 13 were male. All of the participants were German native speakers. The age of the sample ranged from 20 to 35 years (M = 26.1, SD = 3.5). The participants were fairly well educated, with 41.7% having a university degree and 50% higher education qualification (Abitur). All of the participants possessed a smartphone (Android phone: 9, iPhone: 9, Blackberry: 1, Windows phone: 1). The majority of the sample (66.7%) was familiar with public transport apps and used them at least once a week. The participants were recruited from an online research pool, administered by the graduate school *prometei* at the *Berlin Institute of*

Technology and received an incentive of 10 Euro for their participation. None of the participants was involved in the expert review.

7.2.2. Material

Three different public transport apps (*DB Navigator*, *FahrInfo* and *Fahrplan*) were used in this study. All of them were provided on an iPhone 4S. The two tasks that had to be accomplished with each of the apps were given as a printout. All of the questionnaires that had to be filled out by the participants were programmed in an online survey tool (LimeSurvey 2.00+) and provided on a 15.4 inch Laptop. An overview of the questionnaires is shown in table 1 (p. 29). Using the stopwatch function on an iPad mini, task-completion time was measured.

7.2.3. Independent Variables and Design

The three public transport apps served as the independent variables in this study and were the only variation used. The two use cases and tasks remained the same for each app. All participants evaluated the three apps, resulting in a within-subject design with the factor *app.* The order of the apps was counterbalanced over the participants.

7.2.4. Dependent Variables

The choice of the dependent variables that were included in this study was based on the comparison of the meCUE with other validated questionnaires in order to compare ratings that measure related constructs (see chapter 4). Table 1 p. 29 gives an overview of all dependent variables and their sources.

7.2.5. Additional Measures

The external criterion *task completion time* was assessed in the study to detect correlations with instrumental qualities and therefore to examine criterion-related validity. Since the time it takes to complete a task with an app also depends on the participants' familiarity with the iPhone, data from the text message task (see chapter 7.2.6.) were used in order to elaborate whether there is a correlation with task completion time. Questions about demographics, prior experience, and smartphone usage were asked to get a better understanding of the sample. Affinity for technology (TA) (Karrer, Glaser, Clemens, & Bruder, 2009) and centrality of visual product aesthetics (CVPA) (Bloch, Brunel, & Arnold, 2003) were used as quasi factors to examine their effect on the evaluation of the apps.

Table 1: List of the dependent variables: Subscales, sources and number of items

Subscales	Sources	Items
Effectiveness Efficiency Visual aesthetics Status Commitment Positive Emotions Negative Emotions Product loyalty Intention to use	meCUE (Minge & Riedel, 2013)	33
Overall judgment	meCUE (Minge et al., 2013)	1
Visual aesthetics	VisAWI-S (Thielsch & Moshagen, 2011)	4
Attractiveness Hedonic Quality Pragmatic Quality	AttrakDiff-mini (Diefenbach & Hassenzahl, 2010)	10
Positive Emotions Negative Emotions	PANAS (Tellegen et al., 1988)	20
Attractiveness Perspicuity Efficiency Dependability Stimulation Novelty	UEQ (Laugwitz et al., 2006)	26

7.2.6. Procedure

The study was conducted in one of the labs at the Center of Human-Machine Systems of the Berlin Institute of Technology with an average duration of 45 minutes. All participants evaluated each of the three apps. The presentation order of the apps was counterbalanced over the participants.

As can be seen in figure 8, each session contained ten steps. The sessions started with a general welcome and an oral introduction to the experiment by the researcher. The

introduction was also given on paper and was provided during the whole experiment in case the participant wanted to re-read the session rules. It was emphasized by the researcher that there are no right or wrong answers but that it is the participant's personal opinion that is of interest for this study. Participants were then introduced to the iPhone and were asked to type a given text message on the iOS messenger app (see appendix C). The text message was provided on a handout and was the same text for all participants. The researcher recorded the time needed and noted how many letters had been written after one minute. The purpose of this step was to familiarize the participants with the device, because more than half of them were no iPhone users. In addition to that, the data generated in this step (letters written in one minute) helped to control for individual differences in task completion time of the later accomplished tasks (see step 4, 6 and 8).

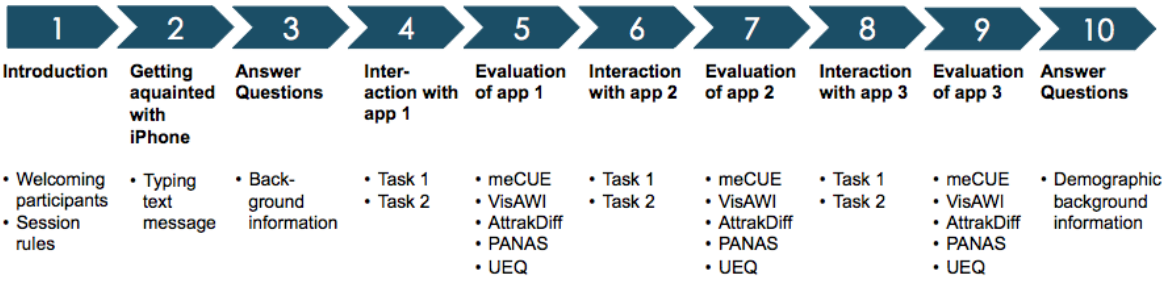

Figure 8: Procedure of the laboratory experiment in the main study

In step 3 participants were asked to give background information regarding their smartphone and their prior experience of public transport apps, by filling out an online questionnaire on the provided laptop.

The actual experiment began with the interaction with the first app (step 4). In order to do so the first use case was presented to the subjects, which was about finding a route for a visiting friend. The participants were given a maximum of two minutes to accomplish the task. The task was finished as soon as they were able to show the researcher the arrival time on the app. The second use case also involved finding a way from A to B, the start and destination point, however, were different from the first use case. Within a maximum of two minutes, participants had to figure out on which platform they would arrive at Berlin Main Station on a specific time and date. The task was finished once the participants were able to show her the relevant information on the app and here too, the researcher recorded the time. After having accomplished the two tasks, the participants were then asked to evaluate their experience with the app

by filling in the relevant questionnaires online. The order of the questionnaires was the same for all apps and participants. The items, however, were given in random order. Step 4 and 5 were then repeated for the two remaining apps. Since the use cases were exactly the same for all three apps it could be expected that participants quickly memorize the solution to the tasks and become faster in accomplishing them. To prevent this, they were explicitly asked to show where they find the relevant information on the app display instead of simply telling the researcher the solution.

In the final step the participants filled in the technology-affinity (TA) questionnaire and the scale for centrality of visual product aesthetics (CVPA) and provided some demographic background information by answering questions online. The sessions ended by thanking them for participating and handing out the incentive.

7.3. Results

The data from the online questionnaires as well as task-completion time and the amount of letters written in step 2, were analyzed and edited with SPSS 21 and Microsoft Excel 2011. For the comparison between apps on pre-determined dimensions a three-way multivariate analysis of variance (MANOVA) with repeated measures was carried out for all the dependent variables. From the TA-questionnaire a single score was computed and participants were categorized as scoring low (less than the mean score) or high (higher than mean score) on the scale. The same procedure was applied to the responses on the scale for CVPA. Consequently TA and CVPA served as quasi-factors and were entered as between-subjects factors into the MANOVA.

For the purpose of investigating whether the meCUE found differences between apps, evaluation scores of three apps were compared with each other. The Likert scores for each dimension were summed up and divided by the total number of included items in order to calculate mean scores for the corresponding subscales. These scores were used as dependent variables. Appendix D shows minimum and maximum ratings, mean scores and standard deviations for all subscales.

Post-hoc pairwise comparisons (Bonferroni) were conducted for the purpose of finding out which of the three apps are significantly different from each other. Correlations of the meCUE with the other questionnaires were analyzed by calculating Pearson bivariate correlation coefficients.

Response scores of negatively worded evaluation statements were reversed. As a consequence, for all statements, high responses indicate a high score on the corresponding subscale.

7.3.1. Discriminative Validity

Table 2: Mauchly's Test of Sphericity for the dependent variables

Questionnaire	Subscale	Mauchly's W	Approx. Chi-Square	df	Sig.
meCUE	Effectiveness	.32	21.88	2	<.001
	Efficiency	.37	18.66	2	<.001
	Visual aesthetics	.73	6.00	2	.050
	Status	.90	2.09	2	.351
	Commitment	.58	10.22	2	.006
	Positive emotions	.71	6.45	2	.040
	Negative emotions	.51	12.82	2	.002
	Intention to use	.67	7.60	2	.022
	Product loyalty	.69	7.03	2	.030
	Overall judgment	.37	18.79	2	<.001
VisAWI	Visual aesthetics	.57	10.69	2	.005
AttrakDiff-mini	Pragmatic qualities	.56	10.89	2	.004
	Hedonic qualities	.69	7.11	2	.029
	Attractiveness	.38	18.45	2	<.001
PANAS	Positive emotions	.99	.03	2	.985
	Negative emotions	.38	18.28	2	<.001
UEQ	Attractiveness	.25	26.61	2	<.001
	Perspicuity	.49	13.48	2	.001
	Novelty	.84	3.38	2	.184
	Stimulation	.48	14.06	2	.001
	Dependability	.31	22.21	2	<.001
	Efficiency	.67	7.56	2	.023

A three-way MANOVA with repeated measures was computed to check for statistical significant differences between the evaluations of the three apps. The factor *app* was entered as a within-subject factor. TA and CVPA served as between-subject factors. Using Pillai's trace, there was a significant effect of *app* on the independent variables (V=1.5, F(42,42)=2.93, p<.001) and there were no significant interaction effects between *app* and TA, between *app* and CVPA or between *app*, TA and CVPA.

Mauchly´s test (see table 2, p. 32) indicated that the assumption of sphericity had been violated for the main effect of app for meCUE´s EFFECTIVENESS, EFFICIENCY, COMMITMENT, POSITIVE EMOTIONS, NEGATIVE EMOTIONS, INTENTION TO USE, PRODUCT LOYALTY and OVERALL JUDGMENT. Further it had been violated for VisAWI, AttrakDiff-mini's PRAGMATIC QUALITIES, HEDONIC QUALITIES and ATTRACTIVENESS, PANAS' NEGATIVE EMOTIONS, UEQ´s ATTRACTIVENESS, PERSPICUITY, STIMULATION, DEPENDABILITY and EFFICIENCY. Therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity.

Instrumental Product Perceptions

For the instrumental product perceptions/pragmatic qualities, separate univariate ANOVAs revealed significant main effects of the within-subject factor *app* on meCUE´s EFFECTIVENESS, EFFICIENCY, on AttrakDiff-mini's PRAGMATIC QUALITIES and UEQ's PERSPICUITY, DEPENDABILITY and EFFICIENCY. Table 3 gives an overview of degrees of freedom, F-values and significance levels.

Table 3: Main effects of within-subject factor *app* on instrumental product perceptions/pragmatic qualities

Questionnaire	Subscale	Sphericity	df	F	Sig.
meCUE	Effectiveness	Greenhouse-Geisser	1.19, 23.75	27.72	<.001
	Efficiency	Greenhouse-Geisser	1.23, 24.61	13.87	.001
AttrakDiff-mini	Pragmatic qualities	Greenhouse-Geisser	1.39, 27.85	14.03	<.001
UEQ	Perspicuity	Greenhouse-Geisser	1.33, 26.52	11.48	.001
	Dependability	Greenhouse-Geisser	1.18, 23.68	19.98	<.001
	Efficiency	Greenhouse-Geisser	1.51, 30.11	4.17	.035

Figure 9 shows the means for the dependent variables and the individual apps. The results indicate a similar evaluation pattern to that of the expert review with *DB Navigator* scoring highest on all instrumental/pragmatic qualities. *Fahrplan* receives the lowest scores on all dimensions. A table with all means, standard deviations, minimum and maximum scores can be found in appendix D.

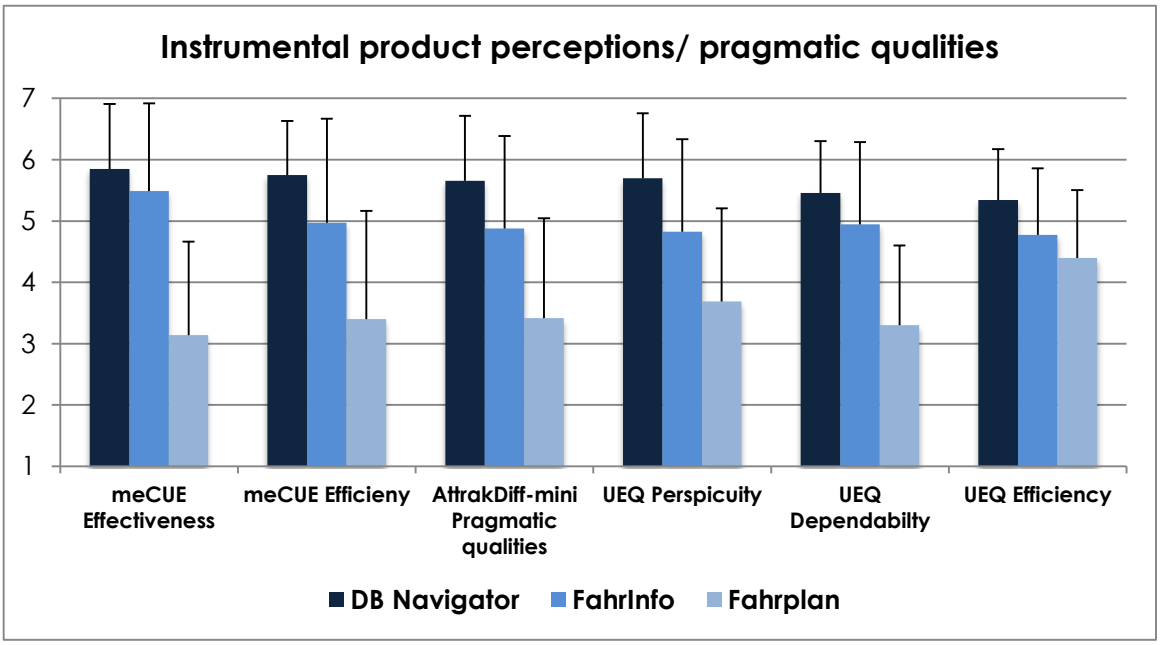

Figure 9: Mean scores and standard deviations
for instrumental product perceptions/ pragmatic qualities

Post-hoc tests for pairwise comparisons (Bonferroni) revealed significant differences between *DB Navigator* and *FahrInfo* on the AttrakDiff-mini's PRAGMATIC QUALITIES ($p<.05$) and UEQ´s PERSPICUITY ($p<.05$). For the comparison of *DB Navigator* with *Fahrplan* significant differences were found on meCUE´s EFFICIENCY ($p<.001$), EFFECTIVENESS ($p<.001$), on AttrakDiff-mini's PRAGMATIC QUALITIES ($p<.001$), PERSPICUITY ($p<.001$), DEPENDABILITY ($p<.001$) and EFFICIENCY ($p<.05$). Finally, significant differences for *FahrInfo* and *Fahrplan* were revealed by meCUE´s EFFECTIVENESS ($p<.05$), UEQ´s DEPENDABILITY ($p<.05$).

Non-Instrumental Product Perceptions

Table 4 gives an overview of the main effects of the within-subject factor *app* and shows that it has significant main effects on all the non-instrumental and pragmatic qualities.

Table 4: Main effects of within-subject factor *app*
on non-instrumental product perceptions/hedonic qualities

Questionnaire	Subscale	Sphericity	df	F	Sig.
meCUE	Visual aesthetics	Sphericity assumed	2,40	35.43	<.001
	Status	Sphericity assumed	2,40	5.32	.009
	Commitment	Greenhouse-Geisser	1.41,28.24	8.10	.004
VisAWI-S	Visual aesthetics	Greenhouse-Geisser	1.4, 27.97	33.61	<.001
AttrakDiff-mini	Hedonic qualities	Greenhouse-Geisser	1.52, 30.48	21.47	<.001
UEQ	Novelty	Sphericity assumed	2,40	35.90	<.001
	Stimulation	Greenhouse-Geisser	1.31, 26.27	25.40	<.001

In order to learn about the direction of these main effects, means for the dependent variables were computed and are shown in figure 10. Within the dimension of hedonic qualities, *FahrInfo* was evaluated as most visually appealing and *Fahrplan* as least visually appealing. This pattern is consistent with the VisAWI. Compared to hedonic qualities of AttrakDiff-mini and UEQ, the subscales STATUS and COMMITMENT of meCUE are comparatively low for all three apps. *FahrInfo* scores highest on these dimensions except for NOVELTY, where *DB Navigator* was rated best. *Fahrplan* received the lowest scores on all of the hedonic/non-instrumental subscales.

Post-Hoc tests for pairwise comparisons (Bonferroni) revealed no significant differences between *DB Navigator* and *FahrInfo* on the non-instrumental product perceptions/hedonic qualities.

Significant differences were found for the comparison of *DB Navigator* with *Fahrplan* on the meCUE´s VISUAL AESTHETICS ($p<.001$) and COMMITMENT ($p<.05$), on the VisAWI ($p<.001$), on the AttrakDiff-mini's HEDONIC QUALITIES ($p<.001$) and on the UEQ´s NOVELTY ($p<.001$) and STIMULATION ($p<.001$).

Comparing *FahrInfo* and *Fahrplan,* the results showed significant differences on the meCUE´s VISUAL AESTHETICS ($p<.001$), STATUS ($p<.05$), COMMITMENT ($p<.05$), by the VisAWI ($p<.001$), by the AttrakDiff-mini´s HEDONIC QUALITIES ($p<.001$) and by the UEQ´s NOVELTY ($p<.001$), STIMULATION ($p<.05$).

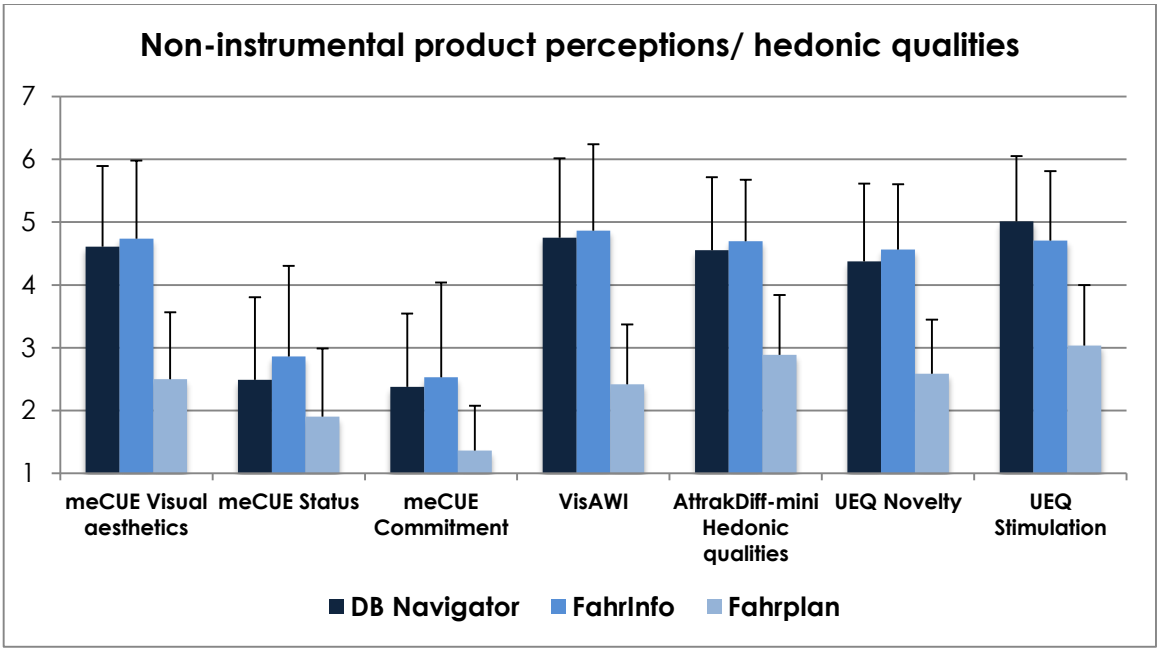

Figure 10: Mean scores and standard deviations
of non-instrumental product perceptions/ hedonic qualities

Emotions

Significant main effects of the factor *app* were found on all subscales of emotions (see table 5). According to the results of descriptive analysis (see figure 11), *DB Navigator* evoked the most positive emotions and *Fahrplan* evoked the most negative emotions. Both emotion scales – meCUE and PANAS – produce similar rankings of the apps, keeping in mind that the PANAS is 5-point Likert-scaled, whereas the meCUE is 7-point Likert scaled.

Table 5: Main effects of within-subject factor *app* on emotions

Questionnaire	Subscale	Sphericity	df	F	Sig.
meCUE	Positive emotions	Greenhouse-Geisser	1.55, 31.05	16.57	<.001
	Negative emotions	Greenhouse-Geisser	1.34, 26.83	18.59	<.001
PANAS	Positive emotions	Spherictiy assumed	2,40	16.77	<.001
	Negative emotions	Greenhouse-Geisser	1.24, 24.74	14.45	<.001

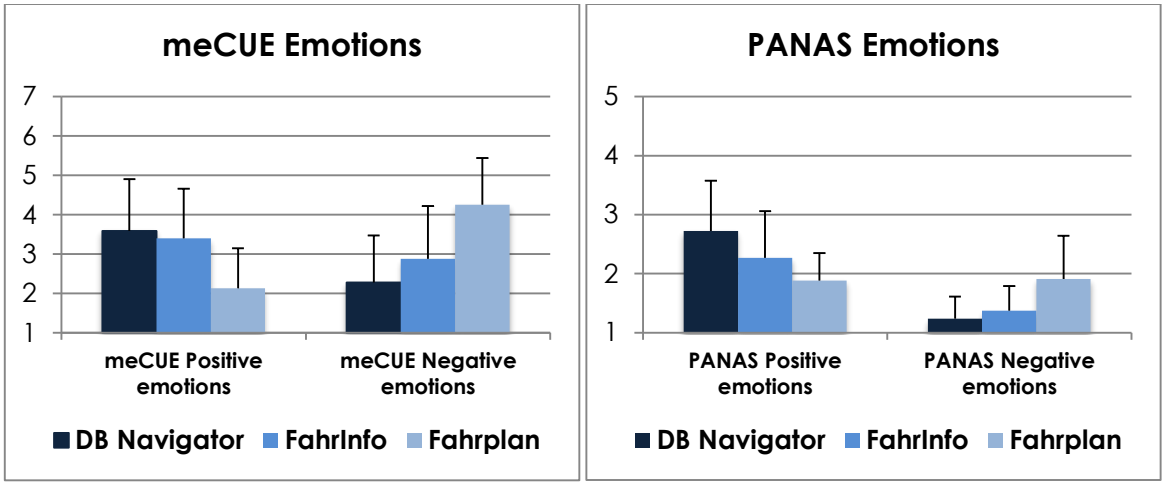

Figure 11*:* Mean scores and standard deviations of emotion-scales.

Bonferroni pairwise comparisons revealed significant differences between *DB Navigator* and *FahrInfo* on the subscales NEGATIVE EMOTIONS of the meCUE ($p<.05$) and the PANAS' POSITIVE EMOTIONS ($p<.05$).

For the comparison of *DB Navigator* with *Fahrplan* significant differences were found on the meCUE´s POSITIVE EMOTIONS ($p<.001$) and NEGATIVE EMOTIONS ($p<.001$). This pattern is consistent with the PANAS' POSITIVE EMOTIONS ($p<.001$) and NEGATIVE EMOTIONS ($p<.001$).

Finally, significant differences for *FahrInfo* and *Fahrplan* were revealed by the meCUE´s POSITIVE EMOTIONS ($p<.05$) and by the PANAS' NEGATIVE EMOTIONS ($p<.05$).

Consequences

As table 6 indicates, significant main effects of *app* were found for both subscales of the module consequences.

Table 6: Main effects of within-subject factor *app* on meCUE´s module *consequences*

Questionnaire	Subscale	Sphericity	df	F	Sig.
meCUE	Intention to use	Greenhouse-Geisser	1.5, 30.09	14.01	<.001
	Product loyalty	Greenhouse-Geisser	1.53, 30.55	21.25	<.001

The means of these subscales (see figure 12) show that *DB Navigator* produced the highest results for both INTENTION TO USE and PRODUCT LOYALTY. *Fahrplan* produced the lowest results.

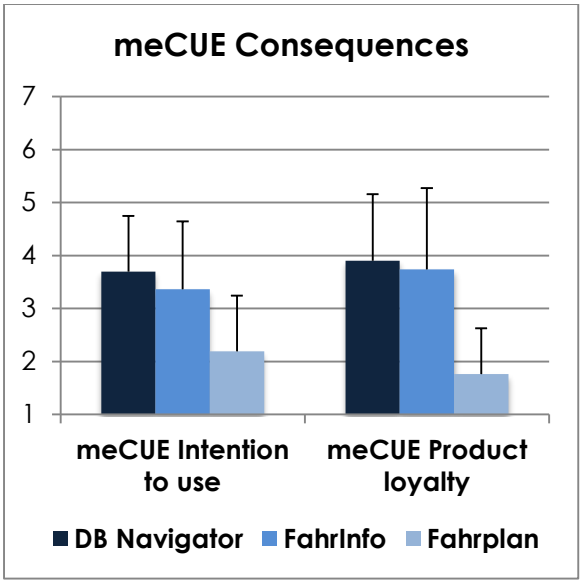

Figure 12*:* Mean scores and standard deviations of the module *consequences*

Significant differences were found on both subscales between *DB Navigator* and *Fahrplan* (INTENTION TO USE: p<.001; PRODUCT LOYALTY: p<.001) and between *FahrInfo* and *Fahrplan (*INTENTION TO USE: p<.05; PRODUCT LOYALTY: p<.05).

Overall Judgment

The within-subject factor *app* had main effects on all scales that assess a general evaluation (see table 7).

Table 7: Main effects of within-subject factor *app* on general evaluation

Questionnaire	Subscale	Sphericity	df	F	Sig.
meCUE	Overall judgment	Greenhouse-Geisser	1.23, 24.57	20.61	<.001
AttrakDiff-mini	Attractiveness	Greenhouse-Geisser	1.23, 24.67	21.61	<.001
UEQ	Attractiveness	Greenhouse-Geisser	1.14, 22.81	21.71	<.001

OVERALL JUDGMENT of the meCUE was assessed by a 21-point single item ranging from -5 to +5. *DB Navigator* scored best on this item; *FahrInfo* also got an average score in the positive range whereas *Fahrplan* produced a mean score in the negative range (see figure 13).

The attractiveness scales of both the AttrakDiff-mini and the UEQ produce comparable pattern of results. *DB Navigator* was rated best, *FahrInfo* received results close to *DB Navigator* and *Fahrplan* got comparatively low results.

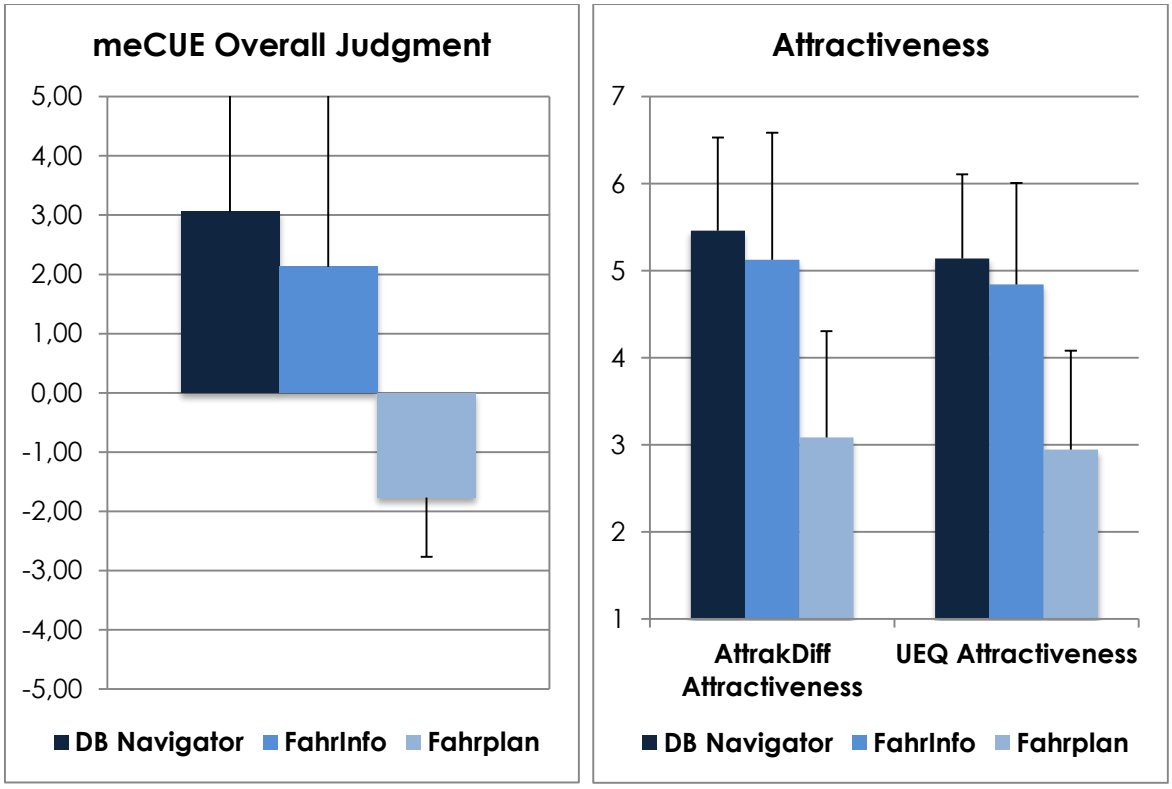

Figure 13*:* Mean scores and standard deviations of general evaluations*.*

There were no significant differences on these scales between *DB Navigator* and *FahrInfo*.

For the comparison of *DB Navigator* with *Fahrplan* significant differences were found on all three, the meCUE´s OVERALL JUDGMENT ($p<.001$), on the AttrakDiff-mini's ATTRACTIVENESS ($p<.001$) and on the UEQ´s ATTRACTIVENESS ($p<.001$).

The comparison of *FahrInfo* and *Fahrplan* also produced significant differences on the meCUE's OVERALL JUDGMENT ($p<.05$), the AttrakDiff-mini's ATTRACTIVENESS ($p<.05$) and the UEQ´s ATTRACTIVENESS ($p<.05$).

TA and CVPA

Significant main effects of the between-subjects factor TA can be reported for the AttrakDiff-mini´s HEDONIC QUALITIES ($F(1,20)=5.23$, $p<.05$) and for the UEQ´s subscale STIMULATION ($F(1,20)=4.44$, $p<.05$). There were no other significant main effects of TA on any of the dependent variables.

The second between-subjects factor CVPA had a main effect on the meCUE´s COMMITMENT ($F(1,20=5.72$, $p>.05$), but no other significant main effects. There were no significant interaction effects between TA and CVPA.

7.3.2. Convergent Validity

The quality of the convergent validity of a psychological measurement scale can be assessed by determining the correlation of its constructs with related constructs of already validated scales. As the assessed questionnaires are considered to be interval (Bortz, 2005), one-tailed bivariate correlations were calculated. As inclusion criterion for Pearson Product Moment correlations, each dimension was first analyzed for normal distribution. For the Pearson coefficient to be valid and to establish whether the correlation is significant, the variables should follow a normal distribution with skewness and kurtosis values close to zero (Field, 2009). The Kolmogorov-Smirnov Test as well as calculated skewness and kurtosis values show that the variables have a sufficiently normal distribution.

Whereas in the analysis of discriminative validity, dimension scores for each of the three apps were compared, in the analysis of convergent validity the focus was on the constructs themselves, which have been assessed by the questionnaire. Consequently, the subscales were averaged over the independent variables for this analysis. For each dimension of the meCUE (instrumental product perceptions, non-instrumental product perceptions, emotions and consequences) the correlation with related constructs was calculated. Appendix E gives an overview of the correlation coefficients of all constructs in question.

For all of the subscales that were being investigated in the scope of this study, there are significant relationships with the corresponding related subscales of other questionnaires.

Product Qualities

As table 8 shows, the non-instrumental product perceptions (EFFECTIVENESS and EFFICIENCY) of the meCUE are significantly correlated with PRAGMATIC QUALITY of the AttrakDiff-mini and the ergonomic quality aspects of the UEQ (PERSPICUITY, DEPENDABILITY and EFFICIENCY). EFFECTIVENESS of meCUE is significantly related to PRAGMATIC QUALITY of the AttrakDiff-mini ($r=.898$, $p<.01$) and to PERSPICUITY ($r=.859$, $p<.01$), DEPENDABILITY ($r=.920$, $p<.01$) and EFFICIENCY ($r=.700$, $p<.01$) of the UEQ. The data also show significant correlations of the meCUE´s EFFICIENCY with PRAGMATIC QUALITY of the AttrakDiff-mini ($r=.900$, $p<.01$) and the UEQ´s PERSPICUITY ($r=.903.$, $p<.01$), DEPENDABILITY ($r=.855$, $p<.01$) and EFFICIENCY ($r=.781$, $p<.01$).

Table 8*:* Correlations of instrumental product perceptions of meCUE with related subscales of other questionnaires; N=72, **p<0.01

		meCUE	
Questionnaire	Subscale	Effectiveness	Efficiency
AttrakDiff mini	Pragmatic Qualities	.898**	.900**
UEQ	Perspicuity	.859**	.903**
	Dependability	.920**	.855**
	Efficiency	.700**	.781**

Non-instrumental product perceptions contain the subscales VISUAL AESTHETICS, STATUS and COMMITMENT in the model of the meCUE. Each of these subscales was again related to similar subscales. The data (see table 9) reveal significant correlations for all pairs. Since the short version of the VisAWI was applied in this study, a single overall score for visual aesthetics was calculated by summing up the scores and dividing them by the number of items. This score for VISUAL AESTHETICS of the VisAWI-S correlates significantly with the corresponding subscale of the meCUE (r=.881, p<.01). STATUS and COMMITMENT of the meCUE were being related to HEDONIC QUALITY of the AttrakDiff-mini and show comparably lower, however significant correlations (r=.473, p<.01 respectively r=.537, p<.01).

Table 9: Correlations of non-instrumental product perceptions of meCUE with related subscales of other questionnaires; N=72, **p<0.01

		meCUE		
Questionnaire	Subscale	Visual Aesthetics	Status	Commitment
AttrakDiff mini	Hedonic qualities	.862**	.473**	.537**
UEQ	Stimulation	.793**	.506**	.557**
	Novelty	.816**	.511**	.492**
VisAWI-S	Visual Aesthetics	.881**	.404**	.489**

Emotions

Emotions are being assessed one-dimensionally by the meCUE, meaning that there are just two states of emotional experience that can be expressed in the questionnaire: positive and negative emotions. Since the PANAS shows a similarly simple underlying concept of the emotional expression through items of a questionnaire, a strong relationship was expected. Both pairs of scores - NEGATIVE EMOTIONS (meCUE and

PANAS) and POSITIVE EMOTIONS (meCUE and PANAS) - show significant correlations (see table 10).

Table 10*:* Correlations of the module *emotions* of meCUE with the PANAS; N=72, **p<0.01

		meCUE	
Questionnaire	Subscale	Positive Emotions	Negative Emotions
PANAS	Positive Emotions	.470**	-.503**
	Negative Emotions	-.456**	.717**

The relation of NEGATIVE EMOTIONS (r=.717, p<.01), however, is stronger than the relation of POSITIVE EMOTIONS (r=.470, p<.01). Furthermore it has to be mentioned that in the meCUE POSITIVE and NEGATIVE EMOTIONS correlate significantly with each other (r=-.372, p<.01), whereas they are uncorrelated in the PANAS (r=-.125, p=.296).

Overall judgment

The single-item on whether the participants generally liked or disliked the product was operationalized by an adjustable slider that ranged from -5 to 5. The slider could be moved in steps of 0.5, providing a 21-point scaled answer format. These scores were expected to show high correlations with other dimensions that assess positive or negative global evaluation. Correlation analysis confirms this expectation (see table 11) as the OVERALL JUDGMENT correlates significantly with the AttrakDiff-mini's dimension ATTRACTIVENESS (r=.919, p<.01) and the UEQ´s dimension ATTRACTIVENESS (r=.887, p<.01).

Table 11: Correlations of the module *overall judgment* of meCUE with the PANAS; N=72, **p<0.01

		meCUE
Questionnaire	Subscale	Overall Judgment
AttrakDiff-mini	Attractiveness	.919**
UEQ	Attractiveness	.887**

7.3.3. Criterion Validity

According to Kaplan & Saccuzzo (2005) "criterion validity evidence tells us just how well a test corresponds with a particular criterion. Such evidence is provided by high correlations between a test and a well defined criterion measure" (p. 137).

For the dimension *instrumental product qualities*, the criterion validity was evaluated by correlating the subscale EFFICIENCY with the time it took to accomplish the two tasks. Task completion time was summed up for both tasks and measured in seconds. The analysis revealed modest correlations with EFFICIENCY (r=-.602, p<.01).

Table 12: Correlations of the subscale *efficiency* of the meCUE, *pragmatic qualities* of the AttrakDiff-mini and the subscale *efficiency* of the UEQ with the external criterion task completion time; N=72, *p<0.05 ** p<0.01

	meCUE	AttrakDiff-mini	UEQ
External criterion	Efficiency	Pragmatic qualities	Efficiency
Task-completion time	-.507**	-.467**	-.257*

As table 12 shows, the relationship between PRAGMATIC QUALITY/EFFICIENCY and task completion time is less strong for the AttrakDiff-mini (r=-.467, p<.001) and the UEQ (r=-.257, p<.001). There was no significant relationship between task completion time and letters written in the text-messaging task.

7.4. Summary and Conclusion

The data indicate that the meCUE qualifies well for discriminating between different apps. The results of the quantitative study are consistent with the outcomes of the qualitative expert review conducted beforehand. Thus all hypotheses can be confirmed.

7.4.1. Research Question 1

DB Navigator and *Fahrplan* were expected to show the largest difference on subscales that directly or indirectly measure usability. Significant differences could be found on all of these subscales (H1a confirmed). The descriptive analysis showed that *DB Navigator* scores higher on the subscales related with usability than *FahrInfo* does (H1b confirmed). Whereas this comparatively smaller difference was significant on AttrakDiff-mini's subscale PRAGMATIC QUALITIES as well as on the UEQ's subscale PERSPICUITY, the meCUE did not show significant differences between these two apps on the instrumental product perceptions. At a first glance this could indicate that that the AttrakDiff(-mini) and the UEQ qualify better for detecting small usability differences. Concerning the subscales EFFICIENCY and DEPENDABILITY of the UEQ, however, no significant differences were found either.

As a reminder, in the expert review *FahrInfo* was rated as having the best design compared to the others apps. *DB Navigator* was rated as second best with scores close to *FahrInfo* whereas *Fahrplan* was rated as having a comparatively bad design.

The subscale VISUAL AESTHETICS of the meCUE is supposed to assess the perception of design and showed significant differences between *FahrInfo* and *Fahrplan* and between *DB Navigator* and *Fahrplan* (H1f confirmed). Although it is apparent from the analysis of means that *FahrInfo* scored higher than *DB Navigator* (H1e confirmed), the difference is rather small and not significant. This is not surprising as it was assumed from the results of the expert review that the two apps are close to each other on a design evaluation scale. The validity of these data is supported by the results of the VisAWI-S.

As visual appeal relates to non-instrumental respectively hedonic qualities of product perceptions (Mahlke, 2008) it was expected that *FahrInfo* will score comparatively high on these dimensions. The descriptive data show the expected tendency with the difference between *FahrInfo* and *DB Navigator* again being rather small. Except for the subscale STIMULATION of the UEQ, *FahrInfo* got the best ratings on the non-instrumental/hedonic product perceptions. However, none of the differences between *FahrInfo* and *Fahrplan* on non-instrumental/hedonic qualities became significant. Nevertheless it can be concluded that the meCUE was able to detect rather small differences as the quantitative data clearly reproduce the tendency of the expert review (H1c confirmed). In addition to that, the meCUE found significant differences between *FahrInfo* and *Fahrplan* on all non-instrumental product perceptions. This is supported by the data of the VisAWI-S, the AttrakDiff-mini´s HEDONIC QUALITIES and the UEQ´s NOVELTY and STIMULATION as these subscales revealed significant differences between these apps as well (H1d confirmed).

The module *emotions* of the meCUE shows the same pattern of results as the PANAS. The meCUE found significant differences for both types of emotions between *DB Navigator* and *Fahrplan* and *FahrInfo* and *Fahrplan* as well as for negative emotions between *DB Navigator* and *FahrInfo.* Yet, not all differences became significant on the PANAS as well. Reasons for this could be derived from the different rating formats of the both scales.

Affinity for technology (TA) and centrality of visual product aesthetics (CVPA) had few main effects on the dependent variables. None of them seem to be relevant for the research goals as it cannot be assumed that these user characteristics had major influences on the evaluation of instrumental product qualities or visual aesthetics.

Summarizing the results, it can be stated that the first research question ("Is the meCUE able to detect expected differences on pre-determined dimensions?") is confirmed and that discriminative validity of the meCUE is clearly supported by the data gathered in this study.

7.4.2. Research Question 2

The calculated correlations show the expected patterns and provide a strong indicator for the convergent validity of the meCUE. For the instrumental product perceptions of the meCUE (EFFECTIVENESS and EFFICIENCY) highly significant correlations with all related subscales could be established (H2a and H2b confirmed). For the non-instrumental product perceptions, significant correlations between the subscale VISUAL AESTHETICS of the meCUE and the VisAWI-S were revealed (H2c confirmed). In addition to that, significant correlations between STATUS and COMMITMENT and the corresponding HEDONIC QUALITIES of the AttrakDiff-mini support the validity of these subscales (H2f and H2g confirmed).

As expected, POSITIVE EMOTIONS of the meCUE correlate significantly with POSITIVE EMOTIONS of the PANAS (H2d confirmed) and NEGATIVE EMOTIONS of the meCUE correlate significantly with NEGATIVE EMOTIONS of the PANAS (H2e confirmed). Significant correlations between POSITIVE and NEGATIVE EMOTIONS were found in the meCUE but not in the PANAS. This does not seem to be problematic as the developers of the meCUE did not make any assumptions about the independency of negative and positive emotions.

The scale for OVERALL JUDGMENT is a rather new module and was not addressed within the construction of the meCUE. As the name implies it is supposed to measure a general evaluation of a product and gives a single score between -5 and +5. This study revealed highly significant correlations with both ATTRACTIVENESS-scales of the AttrakDiff-mini and the UEQ, which also assess the overall perception of a product with respect to its quality of interaction (H2h confirmed). As the OVERALL JUDGMENT-scale consists of only

one item, it has a major economic advantage over the corresponding ATTRACTIVENESS-scales.

All of the modules that were being examined in this study (*product perceptions, emotions* and *overall judgment*) correlate significantly with related subscales. These data provide a strong support for the convergent validity of the meCUE questionnaire. The module *consequences*, however, was left out in this study since it proved difficult to find related constructs or external criteria to assess its validity. Suggestions for further research on this module are discussed in the next chapter.

7.4.3. Research Question 3

The third research question addressed the criterion validity of the subscale EFFICIENCY. The analysis showed that task completion time correlates significantly with EFFICIENCY. The correlation is also higher for the meCUE than for the respective subscales of the AttrakDiff-mini and the UEQ. Hypothesis 3a can thus be confirmed.

As letters written per minute and task completion time did not correlate significantly it can also be concluded that task completion time does not depend on how well participants dealt with the iPhone interaction. Although *instrumental product perceptions* do not solely depend on how fast users can operate an interactive system, task completion time can be seen as an useful indicator for the ease of use and is thus associated with usability, effectiveness and efficiency.

8. Discussion

Detailed summaries and conclusions have followed the individual results of the expert review and the main study. In the following section general findings will be discussed concerning the research goals and the applied methodology. Finally, suggestions for future research areas will be presented.

8.1. Reflection on Research Questions

In Chapter 5 three research questions have been presented, which define the scope of this thesis and guided the research of this study. They addressed the discriminative, convergent and criterion validity of the newly developed meCUE questionnaire. The empirical data that were gathered and analyzed in this study successfully provided contributions to answering the research questions and they delivered strong indications for the presence of the different types of validity in the questionnaire. As illustrated in chapter 7, the data analysis showed the expected patterns and revealed highly significant results in many cases.

Concerning research question 1, it can be concluded that the meCUE showed the ability to detect expected differences on pre-determined dimensions. Most importantly, these results are consistent with the results of other related questionnaires, which supports the accuracy of the meCUE. As the descriptive analysis showed, even small differences between apps on usability-related and visual appeal-related subscales were detected.

The hypotheses for the main study were derived from the expert review and focused on instrumental and non-instrumental qualities. So far there is little known about the relation between the different modules as for example to what degree product perceptions influence emotions or consequences. Therefore, it proved difficult to establish any apriori hypotheses about these modules based on the expert review. Although the results revealed differences between apps on these subscales (*positive* and *negative emotions, product loyalty, intention to use* and *overall judgment*) there is more theory-based research needed on how the modules influence and affect each other.

Research question 2 addresses the convergent validity, “the evidence of similarity between measures of theoretically related constructs” (DeVellis, 2003, p. 56). The approach taken in this thesis was to identify similar measurement tools and assess

correlations between them and the meCUE. The data show significant correlations between all of the subscales in question, which are an indicator for the convergent validity of the meCUE.

However, a methodological imperfection of this approach becomes apparent in the choice of the related questionnaires. In order to assess convergent validity, it is necessary to match constructs of different measurement tools with each other. Even if the constructs are selected based on theory, it remains to the researcher to make an educated guess about how well constructs can be compared with each other. The data show, that it proved difficult for example to align the non-instrumental product qualities *status* and *commitment* of meCUE with the non-instrumental/hedonic qualities of other questionnaires, as correlations are comparably low for these subscales.

In order to answer research question 3 ("Do ratings on the meCUE correlate with external criteria?"), task completion time was chosen as an external criterion for instrumental product qualities. The correlations are significant and even higher than for the AttrakDiff-mini and the UEQ, which is a first reference for the criterion validity of the module *instrumental product perceptions*. Still, more research is needed on the relation of relevant external criteria. Although in this context, the time it takes to look up a route is a meaningful indicator for non-instrumental product qualities there is more to the usefulness and usability of an interactive product than the time it takes to accomplish a task with it. In addition to that no estimates can be made on the criterion validity of the other modules. In Chapter 8.3. some suggestions for future research on this topic will be discussed.

8.2. Methodological Reflections

The goal of the expert review was to identify the independent variables for the validation of the meCUE questionnaire. Four experts were asked to give their opinion following a structured research setting of a one-hour evaluation session. Although the hypotheses derived from the expert review could be approved in the main study, the approach also showed some methodological weaknesses. When analyzing the data from the expert review, it became apparent that findings are not as generalizable as expected. To better understand this, it should be stated that although they were not explicitly asked to identify usability issues, the four experts made comments on things

they liked and disliked in the apps. These comments revealed that the experts had very different opinions about the severities of usability issues in some cases, which led to a high variation in the quantitative results. Rather than relying on their expert knowledge about usability heuristics and best practice interaction patterns, their judgment mainly depended on subjective opinions. Having this in mind, the choice of the apps for the main study was based on a combination of quantitative and qualitative results rather than on quantitative data alone. For future studies that take a similar approach it is recommended to provide a more standardized expert evaluation structure. Suggestions include the method of heuristic evaluation (Nielsen & Molich, 1990; Nielsen, 1995) or other structured forms of expert reviews (e.g. Turner, 2011).

It should also be mentioned that there is an ongoing debate among usability professionals and researchers on the reliability of usability expert evaluations in general. Molich (2010), for example, started conducting comparative usability evaluation (CUE) studies in 1998 and found out that over 50% of the usability problems were uniquely reported when several professional teams evaluated the same website. An overview of all nine CUE-studies (1998 – 2011) can be found on the website of Molich (n.d.).

Further critique on the applied method concerns the distinction between usability and design. The expert review in this study should provide insights in the apps' aesthetic appeal and instrumental product qualities. In order to make these concepts more concrete and tangible for the experts, they were told to give their opinion on the apps' design and usability. When conducting the expert review it became apparent that it was nearly impossible to treat the two criteria independently as usability was often seen as a result of design characteristics. Consequently, apps that used color to indicate different metro lines or that used different sizes of fonts to give information on the length of a connection were perceived as more usable than other apps. As a result it is strongly advised that future studies should not rate design on par with visual aesthetics. More concrete criteria should be given in order to help experts to systematically judge the visual appeal of interactive systems. A theoretical basis for this can be found in the publication Hartmann, Sutcliffe, & Angeli (2008).

The purpose of the main study was to systematically investigate the validity of the questionnaire in controlled lab settings. Since the construction of the questionnaire as

well as the first validation studies were mainly conducted by evaluating hardware or work-related software, it was of special interest to 'test' the questionnaire's ability to evaluate interactive consumer goods. Although the empirical evidence suggests that the meCUE qualifies well for the application of assessing the user experience of public transport apps, it should be kept in mind that user experience is highly dependent on the context. Trying public transport apps out in a lab can be a different experience from using them in real life settings when for example being under time pressure or using the apps while walking around. It was attempted however to make the lab situation as realistic as possible by embedding the task into use cases and generating time pressure.

8.3. Outlook

This thesis provides a complex and thorough analysis of the meCUE questionnaire and its subscales. The tool was being investigated from different perspectives and a variety of research questions could be answered.

However, based on the nature of the methodological approach it was not possible to examine all of the modules of the meCUE in an equal manner. Some questions remain unanswered and provide substance for future research. Visual aesthetics and usability have been the focus of the examination of discriminative validity here. Research on other non-instrumental product qualities could shed some light on the discriminative power of these subscales. The apps used in this study are mainly used for instrumental goals (finding way from A to B and getting directions) and it was thus not expected that they would have much influence on motivational and symbolic qualities. For future research on this topic it would be interesting to evaluate rather *hedonic* consumer goods such as online games or apps that were developed for entertaining reasons.

Additionally, more external criteria should be identified to assess the questionnaire's criterion validity. Depending on the product's categories and tasks, suggestions include error rate for instrumental qualities, familiarity and frequency of usage for non-instrumental qualities *status* and *commitment* and actual usage over time for the module *consequences*. Here, a distinction should be made between subjective data and behavioral data, as it would be of special interest to get some insights if subjective opinion relates to or even predicts actual usage behavior. In order to pursue a theory-based approach on this research question, hypotheses can be derived based on the

theory of planned behavior (for a review see Ajzen, 1991). Methods such as contextual inquiry or diary studies could provide special benefits for the assessment of behavioral data in this context.

Moreover, the components of user experience should be examined over a longer period of time. Long-term studies or repeated measurements at specific points in time, could for instance give information about the changes in product loyalty or commitment for instance. The ContinUE-model (Pohlmeyer, 2011) discriminates between different phases of user experience and can serve as a framework for research in temporal aspects of the components of meCUE-model.

As mentioned before, there is little known yet about the interrelations between the different modules. In 2006 Mahlke studied the "aesthetic and symbolic qualities as antecedents of overall judgments of interactive products" and found out that instrumental and non-instrumental qualities interact with each other and that they contribute with different weights to an overall judgment of digital audio players. In a later study interactions between usability and aesthetics could not be replicated but it was interpreted from the results that usability had a greater influence on emotions than visual aesthetics (Mahlke & Thüring, 2007). Thielsch and Jaron (2012) on the other hand, found that aesthetics contribute to a larger extent to overall judgment of a website than usability.

To sum up, it remains unclear how the components of the (me)CUE-model are connected to each other and further research on these relations would provide a better understanding of the importance of system properties on user experience. Related research questions for future studies could be: "How do status and commitment influence product loyalty and intention to use?" and "What role do consequences play in the relation between product perceptions and overall judgment?" Regression analyses or structural equation modeling could lead to interesting insights into these topics.

Moreover, more attention should be devoted to user characteristics and context of use. Schmettow and colleagues (2013) criticize that individual differences remain a blind spot in user experience research. In their recent studies with the implicit method of *Stroop priming* they found indications for the presence of a 'geek personality'. According to the authors, these individuals "tend to think of computers as objects of

intellectual challenge and play, rather than tools and extensions of the self" (Schmettow et al., 2013, p. 1). The authors construct the hypothesis that geek personalities assert different claims to user experience than persons with other personalities do. Such hypotheses have to be studied empirically to gain validated knowledge about the influence of individual user characteristics on the user experience.

Finally, it is recommended to translate the meCUE to English. As the questionnaire is only available in German language so far, its field of application is rather narrow. Applying the meCUE in different cultural environments and contexts can further promote its validity.

In conclusion, a variety of open questions and interesting areas of research remain for the meCUE. Nonetheless this work has contributed to the quality of the questionnaire by showing that it offers a powerful tool for a comprehensive assessment of user experience. Furthermore it was demonstrated that the meCUE qualifies well for the evaluation of interactive consumer goods. Although this study focused on mobile applications, the major theoretical assumptions underlying this work as well as the recommendations for future research should be transferable to other domains as well.

References

Abbasi, M. Q., Lew, P., Rafique, I., & Li, Z. (2012). User Experience Evolution Lifecycle Framework. In *Proc. International conference on Information, Systems and Engineering (ICISE)* (pp. 15–17).

Ajzen, I. (1991). The theory of planned behavior. *Organizational behavior and human decision processes*, *50*(2), 179–211.

Ardito, C., Buono, P., Caivano, D., Costabile, M. F., Lanzilotti, R., Bruun, A., & Stage, J. (2011). Usability evaluation: a survey of software development organizations. In *SEKE* (pp. 282–287).

Ardito, C., Buono, P., Costabile, M. F., & Lanzilotti, R. (2012). Are software companies aware of UX? In *International Workshop on the Interplay between User Experience (UX) Evaluation and System Development* (Vol. 2012, p. 71). Presented at the NordiCHI, Copenhagen, Denmark.

Bargas-Avila, J., & Hornbæk, K. (2012). Foci and blind spots in user experience research. *Interactions*, *19*(6), 24–27.

Bloch, P. H., Brunel, F. F., & Arnold, T. J. (2003). Individual differences in the centrality of visual product aesthetics: concept and measurement. *Journal of consumer research*, *29*(4), 551–565.

Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler* (6th ed.). Wein: 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.

Burmester, M., Hassenzahl, M., & Koller, F. (2002). Usability ist nicht alles–Wege zu attraktiven Produkten (Beyond Usability–Appeal of interactive Products). *I-com*, *1*(1/2002), 32.

Capota, K., van Hout, M., & van der Geest, T. (2007). Measuring the emotional impact of websites: a study on combining a dimensional and discrete emotion approach in measuring visual appeal of university websites. In *Proceedings of the 2007 conference on Designing pleasurable products and interfaces* (pp. 135–147).

Clemmensen, T., Hertzum, M., Yang, J., & Chen, Y. (2013). Do Usability Professionals Think about User Experience in the Same Way as Users and Developers Do? In *Proceedings of the IFIP TC 13 International Conference on Human-Computer Interaction* (pp. 461–478). Presented at the INTERACT 2013, Springer.

Desmet, P. (2003). Measuring emotion: Development and application of an instrument to measure emotional responses to products. In M. A. Blythe, A. F. Monk, K. Overbeeke, & P. Wright (Eds.), *Funology: from Usability to Enjoyment* (pp. 111–123). Dordrecht: Kluwe Academic Publishers.

Desmet, P., Overbeeke, K., & Tax, S. (2001). Designing Products with Added Emotional Value: Development and Application of an Approach for Research Through Design. *The design journal*, *4*(1), 32–47.

DeVellis, R. F. (2003). *Scale development: theory and applications* (Vol. 26). Thousand Oaks, Calif.: Sage Publ.

Diefenbach, S., & Hassenzahl, M. (2010). *Handbuch zur Fun-ni Toolbox*. unpublished manuscript, Folkwang Universität der Künste.

Dooley, D. (2001). *Social research methods*. Upper Saddle River, N.J.: Prentice Hall.

Doria, L., Minge, M., & Riedel, L. (2013). User-centred evaluation of lower-limb orthoses: A new approach. Presented at the Dreiländertagung der Deutschen, Schweizerischen und Österreichischen Gesellschaft für Biomedizinische Technik (BMT), Graz.

Fawcett, A. J. L. (2008). *Principles of assessment for occupational therapists and physiotherapists: theory, skills and application*. Chichester: John Wiley & Sons.

Ferreira, J., Sharp, H., & Robinson, H. (2012). Agile Development and User Experience Design Integration as an Ongoing Achievement in Practice. In *Agile Conference (AGILE), 2012* (pp. 11–20).

Field, A. (2009). *Discovering Statistics Using SPSS* (3rd ed.). London: SAGE Publications Ltd.

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* (pp. 261–268). New York, NY, USA: ACM.

Fowler, M., & Highsmith, J. (2001). The agile manifesto. *Software Development*, *9*(8), 28–35.

Glanznig, M. (2012). User Experience Research: Modelling and Describing the Subjective. *Interdisciplinary Description of Complex Systems*, *10*(3), 235–247.

Gothelf, J., & Seiden, J. (2013). *Lean UX: Applying Lean Principles to Improve User Experience*. Cambridge: O'Reilly.

Gube, J. (2010, May 10). What is user experience design? Overviews, Tools And Resources. *smashing magazine*. [Blog-post]. Retrieved from http://uxdesign.smashingmagazine.com/2010/10/05/what-is-user-experience-design-overview-tools-and-resources/

Gunther, R. (2012). An open letter to the user experience community. *uxpa.org*. Retrieved from http://www.ovostudios.com/uxpa/uxpa.html

Halalka, R. (2013, August 6). Approaching Mobile UX - A Product Manager's Perspective. *onproductmanagement.net*. [Blog-post]. Retrieved from http://onproductmanagement.net/2013/08/06/approaching-mobile-ux-a-product-managers-perspective/

Hartmann, J., Sutcliffe, A., & Angeli, A. D. (2008). Towards a theory of user judgment of aesthetics and user interface quality. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *15*(4), 15.

Hassenzahl, M. (2007). The hedonic/pragmatic model of user experience. In *Towards a UX Manifesto* (pp. 10–13). Presented at the HCI 2007, Lancaster: COST.

Hassenzahl, M., Burmester, D. M., & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In G.

Szwillus & J. Ziegler (Eds.), *Mensch & Computer 2003* (pp. 187–196). Vieweg+Teubner Verlag.

Hassenzahl, M., Burmester, M., & Koller, F. (2008). Der User Experience (UX) auf der Spur: Zum Einsatz von attrakdiff. de. *Usability Professionals*, 78–82.

Hassenzahl, M., Diefenbach, S., & Göritz, A. (2010). Needs, affect, and interactive products–Facets of user experience. *Interacting with Computers*, *22*(5), 353–362.

Huisman, G., van Hout, M., van Dijk, E., van der Geest, T., & Heylen, D. (2013). LEMtool: measuring emotions in visual interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 351–360).

ISO 9241-11: Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs): Part 11: Guidance on Usability. (1998).

Jordan, P. W. (2002). *Designing Pleasurable Products: An Introduction to the New Human Factors*. London: CRC Press.

Kakar, A. K. (2012). Engendering Holistic User Experiences: A Hierarchical Model for Balancing the Utilitarian and Hedonic features of IS Products.

Kaplan, R. M., & Saccuzzo, D. P. (2005). *Psychological testing: principles, applications, and issues* (Vol. 6). Belmont, CA: Wadsworth Cengage Learning.

Karrer, K., Glaser, C., Clemens, C., & Bruder, C. (2009). Technikaffinität erfassen-der Fragebogen TA-EG. *Der Mensch als Mittelpunkt technischer Systeme*, *8*, 194–199.

Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., & Newell, C. (2012). Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, *22*(4-5), 441–504.

Lallemand, C. (2011). Toward a closer integration of usability in software development: a study of usability inputs in a model-driven engineering process. In *Proceedings of the 3rd ACM SIGCHI symposium on Engineering interactive computing systems* (pp. 299–302).

Laugwitz, B., Schrepp, M., & Held, T. (2006). Konstruktion eines Fragebogens zur Messung der User Experience von Softwareprodukten. In A. M. Heinecke & H. Paul (Eds.), *Mensch und Computer 2006: Mensch und Computer im Strukturwandel* (pp. 125–134). München: Oldenbourg Verlag.

Lavie, T., & Tractinsky, N. (2004). Assessing dimensions of perceived visual aesthetics of web sites. *International journal of human-computer studies*, *60*(3), 269–298.

Law, E. L.-C., Abrahão, S., Vermeeren, A. P., & Hvannberg, E. T. (2012). Interplay between user experience evaluation and system development: state of the art. In *International Workshop on the Interplay between User Experience (UX) Evaluation and System Development (I-UxSED 2012)* (p. 1).

Mahlke, S. (2006). Aesthetic and Symbolic Qualities as Antecedents of Overall Judgments of Interactive Products. Centre of Human-Machine Systems, Technische Universität Berlin.

Mahlke, S. (2008). *User experience of interaction with technical systems* (unpublished doctoral dissertation). Technische Universität Berlin, Berlin.

Mahlke, S., & Thüring, M. (2007). Studying Antecedents of Emotional Experiences in Interactive Contexts. In *Proceedings of CHI 2007 Emotion & Empathy*. Presented at the CHI 2007, San Jose, CA, USA.

McCarthy, J., Wright, P., Wallace, J., & Dearden, A. (2006). The experience of enchantment in human–computer interaction. *Personal and Ubiquitous Computing*, *10*(6), 369–378.

Minge, M. (2011). *Dynamische Aspekte des Nutzungserlebens der Interaktion mit technischen Systemen* (unpublished doctoral dissertation). Technische Universität Berlin.

Minge, M., Riedel, L., & Thüring, M. (2013). Modulare Evaluation von Technik. Entwicklung und Validierung des meCUE Fragebogens zur Messung der User Experience. Presented at the 10th Berliner Werkstatt Mensch-Maschine Systeme, October 10th - 12th, 2013, Berlin.

Minge, & Riedel, L. (2013). meCUE – Ein modularer Fragebogen zur Erfassung des Nutzungserlebens. Presented at the Mensch und Computer, September 8th - 11th 2013, Bremen.

Molich, R. (2010). A Critique of "How To Specify the Participant Group Size for Usability Studies: A Practitioner's Guide" by Macefield. *Journal of Usability Studies*, *5*(3), 124–128.

Molich, R. (n.d.). CUE-Comparative Usability Evaluation. Retrieved June 10, 2013, from http://www.dialogdesign.dk/CUE.html

Moshagen, M., & Thielsch, M. (2010). Facets of visual aesthetics. *International Journal of Human-Computer Studies*, *68*(10), 689–709.

Moynihan, B., Paul, S., & Markus, H. S. (2013). User Experience of a Centralized Hyperacute Stroke Service A Prospective Evaluation. *Stroke*, *44*(10), 2743–2747.

Nacke, L. E., & Drachen, A. (2011). Towards a framework of player experience research. In *Proceedings of the Second International Workshop on Evaluating Player Experience in Games at FDG*.

Nielsen, J. (1995). Ten usability heuristics. *nngroup.com*. [Blog-post]. Retrieved from http://www.nngroup.com/articles/ten-usability-heuristics/

Nielsen, J., & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems* (pp. 206–213).

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 249–256).

Norman, D. A. (2002). *The Design of Everyday Things* (Basic Pbk.). New York: Perseus Books.

O'Brien, H. L., & Toms, E. G. (2008). What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology*, *59*(6), 938–955.

Pohlmeyer, A. E. (2011). *Identifying Attribute Importance in Early Product Development* (unpublished doctoral dissertation). Technische Universität Berlin.

Riedel, L. (2013). *Entwicklung eines Fragebogens zur Messung des Nutzungserlebens interaktiver Produkte* (unpublished master thesis). Technische Universität Berlin, Berlin.

Roto, V., Obrist, M., & Väänänen-Vainio-Mattila, K. (2009). User experience evaluation methods in academic and industrial contexts. In *Interact 2009 conference, User Experience Evaluation Methods in Product Development (UXEM'09), Uppsala, Sweden*.

Russell, J. A. (1979). Affective space is bipolar. *Journal of personality and social psychology*, *37*(3), 345.

Scapin, D. L., Senach, B., Trousse, B., & Pallot, M. (2012). User Experience: Buzzword or New Paradigm? In *ACHI 2012, The Fifth International Conference on Advances in Computer-Human Interactions* (pp. 336–341).

Schmettow, M., Noordzij, M. L., & Mundt, M. (2013). An implicit test of UX: individuals differ in what they associate with computers. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems* (pp. 2039–2048). Presented at the CHI'13.

Sharp, H., Rogers, Y., & Preece, J. (2007). What is interaction design? In *Interaction design: beyond human-computer interaction* (Vol. 2). Chichester; Hoboken, NJ: Wiley.

Sy, D. (2007). Adapting usability investigations for agile user-centered design. *Journal of usability Studies*, *2*(3), 112–132.

Tellegen, A., Watson, D., & Clark, L. A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of personality and social psychology*, *54*(6), 1063–1070.

Thielsch, M., & Jaron, M. (2012). Das Zusammenspiel von Website-Inhalten, Usability und Ästhetik. In *Mensch & Computer* (pp. 123–132).

Thielsch, M., & Moshagen, M. (2011). Erfassung visueller Ästhetik mit dem VisAWI. *Usability Professionals*, 260–265.

Thüring, M., & Mahlke, S. (2007). Usability, aesthetics and emotions in human–technology interaction. *International Journal of Psychology*, *42*(4), 253–264.

Turner, N. (2011, November 2). A guide to carrying out usability reviews. *UX for the masses*. [Blog-post]. Retrieved from http://www.uxforthemasses.com/usability-reviews/

Väänänen-Vainio-Mattila, K., & Segerstahl, K. (2009). A tool for evaluating service user eXperience (ServUX): Development of a modular questionnaire. In *User Experience Evaluation Methods, UXEM'09 Workshop at Interact* (Vol. 9).

Weinschenk, S. (2010, May 19). The Psychologist's View of UX Design. *uxmag*. [Blog-post]. Retrieved from http://uxmag.com/articles/the-psychologists-view-of-ux-design

Wickens, C. D., & Hollands, J. G. (2000). *Engineering psychology and human performance*. Upper Saddle River, NJ: Prentice Hall.

Yu, Y., & Wu, X. (2010). Research on User Experience Design of B2C E-commerce Based on Personalized Needs. In *Proceedings of the 2nd International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)* (Vol. 2, pp. 100–

103). Presented at the 2010 2nd International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC).

Appendix

Appendix A: meCUE questionnaire (German)

Nachfolgend finden Sie einige Aussagen, mit deren Hilfe Sie das Produkt bewerten können. Kreuzen Sie bitte für jede Aussage an, wie sehr Sie persönlich finden, dass sie auf das Produkt zutrifft.

Es kann sein, dass einige Aussagen nicht so gut zum Produkt passen, kreuzen Sie bitte trotzdem immer eine Antwort an.

Denken Sie nicht zu lange über einzelne Aussagen nach, sondern geben Sie bitte die Einschätzung ab, die Ihnen spontan in den Sinn kommt.

Es gibt keine "richtigen" oder "falschen" Antworten - nur Ihre persönliche Meinung zählt!

Nützlichkeit

	lehne völlig ab	lehne ab	lehne eher ab	weder/ noch	stimme eher zu	stimme zu	stimme völlig zu
Insgesamt halte ich das Produkt für absolut nützlich.	○	○	○	○	○	○	○
Mithilfe des Produkts kann ich meine Ziele erreichen.	○	○	○	○	○	○	○
Die Funktionen des Produkts sind genau richtig für meine Ziele.	○	○	○	○	○	○	○

Benutzbarkeit

	lehne völlig ab	lehne ab	lehne eher ab	weder/ noch	stimme eher zu	stimme zu	stimme völlig zu
Die Bedienung des Produkts ist verständlich.	○	○	○	○	○	○	○
Das Produkt lässt sich einfach benutzen.	○	○	○	○	○	○	○
Es wird schnell klar, wie man das Produkt bedienen muss.	○	○	○	○	○	○	○

Visuelle Ästhetik

	lehne völlig ab	lehne ab	lehne eher ab	weder/ noch	stimme eher zu	stimme zu	stimme völlig zu
Das Design wirkt attraktiv.	○	○	○	○	○	○	○
Das Produkt ist stilvoll.	○	○	○	○	○	○	○
Das Produkt ist kreativ gestaltet.	○	○	○	○	○	○	○

Status/ Soziale Identität

	lehne völlig ab	lehne ab	lehne eher ab	weder/ noch	stimme eher zu	stimme zu	stimme völlig zu
Das Produkt verleiht mir ein höheres Ansehen.	○	○	○	○	○	○	○
Durch das Produkt werde ich anders wahrgenommen.	○	○	○	○	○	○	○
Meine Freunde dürfen ruhig neidisch auf das Produkt sein.	○	○	○	○	○	○	○

Bindung/ Individuelle Identität

	lehne völlig ab	lehne ab	lehne eher ab	weder/ noch	stimme eher zu	stimme zu	stimme völlig zu
Ohne das Produkt kann ich nicht leben.	○	○	○	○	○	○	○
Das Produkt ist wie ein Freund für mich.	○	○	○	○	○	○	○
Wenn ich das Produkt verlieren würde, würde für mich eine Welt zusammenbrechen.	○	○	○	○	○	○	○

Positive Emotionen

	lehne völlig ab	lehne ab	lehne eher ab	weder/ noch	stimme eher zu	stimme zu	stimme völlig zu
Durch das Produkt fühle ich mich ausgeglichen.	○	○	○	○	○	○	○
Das Produkt beruhigt mich.	○	○	○	○	○	○	○
Das Produkt entspannt mich.	○	○	○	○	○	○	○
Das Produkt stimmt mich euphorisch.	○	○	○	○	○	○	○
Das Produkt beschwingt mich.	○	○	○	○	○	○	○
Durch das Produkt fühle ich mich fröhlich.	○	○	○	○	○	○	○

Negative Emotionen

	lehne völlig ab	lehne ab	lehne eher ab	weder/ noch	stimme eher zu	stimme zu	stimme völlig zu
Das Produkt macht mich müde.	○	○	○	○	○	○	○
Durch das Produkt fühle ich mich erschöpft.	○	○	○	○	○	○	○
Durch das Produkt fühle ich mich passiv.	○	○	○	○	○	○	○
Das Produkt nervt mich.	○	○	○	○	○	○	○
Das Produkt verärgert mich.	○	○	○	○	○	○	○
Das Produkt frustriert mich.	○	○	○	○	○	○	○

Nutzungsintention

	lehne völlig ab	lehne ab	lehne eher ab	weder/ noch	stimme eher zu	stimme zu	stimme völlig zu
Wenn ich könnte, würde ich das Produkt täglich nutzen.	○	○	○	○	○	○	○
Ich kann es kaum erwarten, das Produkt erneut zu verwenden.	○	○	○	○	○	○	○
Wenn ich mit dem Produkt zu tun habe, vergesse ich schon mal die Zeit.	○	○	○	○	○	○	○

Produktloyalität

	lehne völlig ab	lehne ab	lehne eher ab	weder/ noch	stimme eher zu	stimme zu	stimme völlig zu
Ich würde mir genau dieses Produkt jederzeit (wieder) zulegen.	○	○	○	○	○	○	○
Ich würde das Produkt gegen kein anderes eintauschen.	○	○	○	○	○	○	○
Im Vergleich zu diesem Produkt wirken andere Produkte unvollkommen.	○	○	○	○	○	○	○

Globales Produkturteil

Geben Sie bitte abschließend an, wie Sie das Produkt insgesamt bewerten.

Appendix B: Guideline Expert Evaluation (German)

Vielen Dank, dass Sie sich bereit erklärt haben, an meiner Experten Evaluation teilzunehmen. Bei der nun folgenden ca. 1-stündigen Sitzung geht es darum, sechs verschiedene Apps zu bewerten. Die Bewertung soll mir dabei helfen, die richtigen Apps für meine Hauptstudie auszuwählen, in der es darum geht einen Fragebogen zu validieren, der User Experience Aspekte erfasst.

Zuerst werde ich Ihnen verschiedene Screenshots der Apps zeigen und sie bitten, das Design der Apps zu bewerten. Danach werde ich Ihnen ein Aufgabenszenario geben und Sie dürfen mit den Apps auf dem bereit liegenden Smartphone interagieren. Anschließend an jede Interaktion möchte ich Sie bitten die Usability der Apps zu bewerten. Sie können Ihre Bewertungen im Nachhinein noch anpassen, solange dieser Schritt noch nicht abgeschlossen ist. In den darauf folgenden Schritten wird es noch um Ihr Gesamturteil gehen.

Bei den Bewertungen möchte ich Sie bitten auf ihr Wissen, was Design und Usability betrifft, zurückzugreifen und die Apps aus Expertensicht zu beurteilen.

Schritt 1: Apps zeigen (Screenshots Fotopapier)

1.1. Fragen nach jeder App:

	stimme völlig zu	stimme zu	stimme eher zu	weder noch	lehne eher ab	lehne ab	lehne völlig ab
Die App hat eine gute Usability.	☐	☐	☐	☐	☐	☐	☐

Wie kommen Sie zu diesem Wert? Gibt es etwas, das Ihnen besonders gefällt/missfällt?

1.2. Anschließende Fragen:

Kennen Sie eine oder mehrere dieser Apps?

- Ja/nein
- Wenn ja, wie häufig benutzen Sie sie?
  - Mehrmals täglich
  - Mehrmals pro Woche
  - Mehrmals pro Monat

Schritt 2: Aufgabe

Eine Freundin besucht Sie gerade in Berlin und ist alleine unterwegs. Sie ruft Sie an, weil sie sich verlaufen hat und nicht recht weiß, wie sie zu Ihrem Treffpunkt kommt.

Sie befindet sich am **S+U Hermannstraße** und möchte zum **S+U Potsdamer Platz**. Finden Sie bitte mit Hilfe der App für sie heraus, mit welchen öffentlichen Verkehrsmitteln sie am schnellsten ist.

Bitte versuchen Sie dabei laut zu denken beim Bearbeiten dieser Aufgabe. D.h. teilen Sie mir bitte mit, was Sie gerade machen, ob die App Ihre Erwartungen erfüllt und was Ihr Eindruck von der App ist.

2.1. Fragen nach jeder App:

	stimme völlig zu	stimme zu	stimme eher zu	weder noch	lehne eher ab	lehne ab	lehne völlig ab
Die App hat eine gute Usability.	☐	☐	☐	☐	☐	☐	☐

Wie kommen Sie zu diesem Wert? Gibt es etwas, das Ihnen besonders gefällt/missfällt?

Schritt 3: Gesamturteil

Nachdem Sie die Apps nun hinsichtlich Ästhetik und Usability bewertet haben, bringen Sie sie bitte in eine Rangordnung. Ganz links sollte die App sein, die Ihnen insgesamt am wenigsten gefällt und ganz rechts die App, die Ihnen am besten gefällt.

Schritt 4: Gewichtung

Sie haben eben angegeben, welche Apps Ihnen gut oder weniger gut gefallen. Angenommen, Ihr Gesamturteil setzt sich nur aus Ästhetik und Usability zusammen, stellen Sie sich vor: Ästhetik + Usability = 10.

Vor Ihnen liegen 20 Papierstreifen. Wie groß sind die Anteile, die Ästhetik und Usability in Ihrem Gesamturteil jeweils ausmachen? Legen Sie bitte so viele Papierstreifen der entsprechenden Farbe auf die Smartphone-Fläche, wie der Anteil jeweils ausmacht. Wie viele Steine bekommt Ästhetik und wie viele Steine bekommt Usability?

****3 min Pause*** Berechnung***

Schritt 5: Berechnete Gewichtung

Durch Ihre Bewertungen der Apps hinsichtlich Ästhetik und Usability und der Gewichtung ergibt sich folgende Rangordnung.

Möchten Sie Ihre Rangordnung noch einmal ändern? Optional: Wie erklären Sie sich diese Abweichung?

Appendix C: Guideline Main Study (German)

Schritt 1: Willkommen heißen und Ablauf erklären

Herzlich willkommen zu meiner Studie und vielen Dank, dass Sie teilnehmen.

In meiner Studie geht es darum, einen neuen Fragebogen zu untersuchen, der verschiedene Aspekte bei dem Umgang mit technischen Produkten erfasst. Das heißt, der Fragebogen „misst“, wie gut oder schlecht ein technisches Produkt bewertet wird.

Die Produkte, um die es heute geht, sind drei verschiedene Apps zur Routenplanung mit öffentlichen Verkehrsmitteln.

Der Ablauf des Experiments sieht wie folgt aus: Zunächst bitte ich Sie, am Computer einige Fragen zu Ihren Vorerfahrungen mit Smartphones und ÖPNV Apps zu beantworten. Danach werden Sie Aufgaben mit insgesamt drei verschiedenen Apps bearbeiten, die wir Ihnen auf dem bereit liegenden Smartphone zur Verfügung stellen. Jede App bewerten Sie direkt im Anschluss anhand mehrerer Fragebögen am Computer.

Dabei werden Sie merken, dass einige Fragen sehr ähnlich klingen werden, bzw. sich sogar wiederholen. Dies dient der Untersuchung unseres Fragebogens. Bitte beziehen Sie sich beim Ausfüllen der Fragebögen darauf, wie gut oder schlecht Sie kurz zuvor mit der App zurecht kamen. Antworten Sie einfach spontan. Es gibt keine richtigen oder falschen Antworten, nur Ihre Meinung zählt!

Das Ganze wird in etwa 60 min dauern. Wenn Sie eine Pause brauchen, können Sie jederzeit Bescheid sagen.

Haben Sie noch Fragen?

Schritt 2: Interaktion mit dem iPhone

Bevor wir loslegen, möchte ich Sie noch bitten einen kurzen SMS-Text mit dem Smartphone abzutippen. Sie haben eine Minute Zeit und ich werde mir notieren, wie viel Zeichen Sie in einer Minute geschrieben haben. Es geht dabei nicht darum, Sie zu testen, sondern einen Referenzwert zu produzieren, der eine Aussage darüber gibt, ob Sie die Interaktion mit dem iPhone gewohnt sind.

SMS-Text:

Hallo Anja,

ich schaffe es leider nicht bis 18:00 Uhr bei Dir zu sein Ich muss heute länger arbeiten und bin wahrscheinlich erst gegen 19:00 Uhr im Büro fertig. Danach würde ich dann noch eben im Supermarkt vorbeifahren und Wein kaufen. Soll ich noch etwas anderes mitbringen? Ich bin dann wahrscheinlich gegen 19:30 Uhr bei Dir. Ihr könnt aber ruhig schonmal ohne mich mit dem Essen anfangen.

Bis nachher.

Schritt 3: Ausfüllen Fragebogen Vorerfahrung (am PC)

Schritt 4: Aufgabenszenario 1 mit App 1

Sie bekommen demnächst Besuch von einer Freundin aus Leipzig und Ihre Freundin hat Sie gebeten, Ihr eine gute Verbindung für den Weg vom Ostbahnhof zu Ihrem Treffpunkt rauszusuchen. Ihre Freundin kommt am 14. August um 16:20 Uhr am Ostbahnhof an und Sie möchten sich danach am Kottbusser Tor treffen. Bitte finden Sie mit Hilfe der App eine gute Verbindung für Ihre Freundin heraus. Ich werde für diese Aufgabe die Zeit stoppen, sie haben maximal zwei Minuten Zeit. Die Aufgabe ist erfolgreich geschafft, sobald Sie mir zeigen, wo Sie die Information ablesen, **wann Ihre Freundin am U Kottbusser Tor ankommt**.

Aushändigen Handout mit Information:

Start:	Ostbahnhof Berlin
Ziel:	U Kottbusser Tor Berlin
Ankunft:	14. August 2013, 16:20 Uhr

Aufgabeszenario 2 mit App 1

Sie möchten eine Reise planen. Am 2. September fahren Sie mit dem Zug in den Urlaub und um den reibungslosen Ablauf am Tag der Abreise gut planen zu können, möchten Sie vorab schon einmal wissen, wie Sie am besten zum Hauptbahnhof kommen und auf welchem Gleis Sie sich mit Ihrer Reisebegleitung treffen sollen. Nehmen Sie an, Sie wohnen in der Nähe des Kottbusser Tors und möchten von dort aus zum Hauptbahnhof. Sie möchten die Route so planen, dass Sie vor 12:00 Uhr mittags am Hauptbahnhof ankommen.

Ich werde für diese Aufgabe die Zeit stoppen, sie haben maximal zwei Minuten Zeit. Die Aufgabe ist erfolgreich geschafft, sobald Sie mir zeigen, wo Sie die Information ablesen, **auf welchem Gleis Sie gegen 12:00 Uhr am Hauptbahnhof ankommen werden.**

Aushändigen Handout mit Information:

Start:	U Kottbusser Tor Berlin
Ziel:	Berlin Hbf
Ankunft:	2. September 2013, letzte Verbindung vor 12:00 Uhr

Schritt 5: Ausfüllen der Fragebögen (am PC)

MeCUE
VisAWI-S
AttrakDiff-mini
PANAS
UEQ

Kennen Sie die App? Wenn ja, wie häufig nutzen Sie die App?

Schritt 6: Aufgabenszenarien mit App 2

Schritt 7: Ausfüllen der Fragebögen (am PC)

Kennen Sie die App? Wenn ja, wie häufig nutzen Sie die App?

Schritt 8: Aufgabenszenarien mit App 2

Schritt 9: Ausfüllen der Fragebögen (am PC)

Kennen Sie die App? Wenn ja, wie häufig nutzen Sie die App?

Schritt 10: abschließende Fragen

Entscheidung für eine App

Bitten planen Sie nun Ihren Weg von hier nach Hause mit einer der Apps. Sie dürfen sich aussuchen, mit welcher App Sie das machen möchten.

Ausfüllen CVPA und TA-EG am PC

Ausfüllen Demographischer Fragebogen (am PC)

Schritt 11: Bedanken und auf Wiedersehen

Appendix D: Descriptive Statistics

See Excel file *Appendix_D_and_E* on CD-Rom for a complete list of descriptive statistics (N, Minimum, Maximum, Mean, Standard Deviation) of all independent variables

Appendix E: Correlation Table

See Excel file *Appendix_D_and_E* on CD-Rom for a complete table of correlations between all independent variables.

Eidesstattliche Versicherung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und eigenhändig sowie ausschließlich unter Verwendung der aufgeführten Quellen und Hilfsmittel angefertigt habe.

Berlin, 28. Oktober 2013

_____

(Carina Kuhr)