


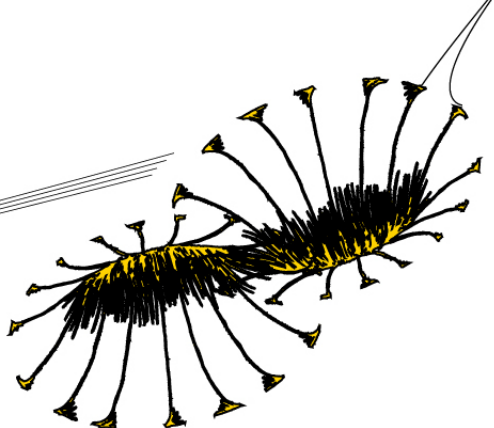
EXPLORING THE APPLICABILITY OF IMPLICIT MEASURES TO ASSESS UX IN HCI RESEARCH



Victoria Sinram
s1200100

Master thesis

December 2016



University of Twente
Faculty of Behavioral Science
Master of Human Factors and Engineering Psychology

Supervisors

Dr. Martin Schmettow, University of Twente, Netherlands

Dr. Matthijs Noordzij, University of Twente, Enschede, Netherlands

Dr. Matthias Peissner, Kathrin Pollmann & Nora Fronemann, Fraunhofer Institute for Industrial Engineering IAO,
Stuttgart, Germany

Table of Content

ABSTRACT	1
SAMENVATTING	1
1 INTRODUCTION	2
1.1 A DEFINITION OF USER EXPERIENCE	2
1.2 CURRENT RESEARCH METHODS ON USER EXPERIENCE AND LIMITATIONS	4
1.3 NEW METHODS FOR ASSESSING UX	5
1.4 MAIN AIM OF THE STUDY	7
1.5 THE MEASURES	8
1.6 THE DEVELOPMENT OF STIMULUS MATERIAL	10
1.6.1 THE NUX PROTOTYPE	11
1.6.2 THE UX PROTOTYPE	11
1.7 CROSS-CORRELATIONAL ASSUMPTIONS	13
1.8 PRE-REQUISITES FOR THE CROSS-CORRELATIONAL ANALYSIS	13
1.8.1 PRE-STUDY	13
Manipulation Assumption – Prototypes	13
Method	13
Results	14
Discussion	16
2 METHOD	17
2.1 EXPERIMENTAL DESIGN	17
2.2 PROCEDURE	17
2.3 PARTICIPANTS	17
2.4 MATERIAL	18
2.4.1 TASK MATERIAL	18
2.4.2 STIMULUS MATERIAL	18
2.5 TASKS - INTERACTION WITH THE PROTOTYPES & UEQ-MECUE	18
2.6 MEASURES	18
2.6.1 AAT	19
2.6.2 AMP	19
2.6.3 SAM	19
3 RESULTS	20
3.1 DATA PREPARATION	20
3.2 CROSS-CORRELATION – A COMPARISON OF IMPLICIT AND EXPLICIT MEASURES	20
3.3 PRE-REQUISITES FOR THE CROSS-CORRELATIONAL ANALYSIS	21
3.3.1 MANIPULATION CHECKS	21
UEQ-meCUE	21
SAM	22
AAT	23
AMP	23
4 DISCUSSION	24
4.1 THE RESEARCH QUESTION	24
4.2 EXPLANATIONS	24
4.2.1 THE MANIPULATION OF THE PROTOTYPES	24
4.2.2 THE ROLE OF EMOTIONS	25
4.3 LIMITATIONS	26
4.4 FUTURE PROSPECTS	28
4.5 CONCLUSION	29
5 REFERENCES	29
6 APPENDIX	34

Abstract

Aim of this study was to examine whether implicit measures can be used as a contribution to explicit measures for assessing user experience (UX) in applied research. For this, a two-fold study design was established: In a pre-study, two newly developed software prototypes were examined concerning their UX and usability in order to obtain stimulus material for the subsequent tests of the main-study. Herein, the attitudes of 43 participants concerning the prototypes were investigated by two implicit measures, the Affect Misattribution Procedure (AMP) and Approach-Avoidance Task (AAT) as well as two explicit measures, the UEQ-meCUE-questionnaire and Self-Assessment-Manikin (SAM) valence scale. Concerning the research question, no meaningful correlations between the two implicit tests and the SAM were found. Thus, there is no relationship between the ratings on these tests. Likewise, a correlation between the implicit tests also showed no meaningful results, making a validation of the two measures impossible. To reach this conclusion, each test was considered separately: The analyses revealed that both explicit measures show higher UX mean-ratings for the UX prototype, accordingly to the expectations. The implicit tests did not replicate these findings, indicating that in this context and with this stimulus material, the implicit tests were not able to detect the difference in UX. In sum, the explicit tests show the expected difference in the manipulation of UX separately from the products' usability. However, the implicit tests were not sensitive enough to detect this difference and therefore cannot add a contribution to the explicit measures of this study.

Samenvatting

Doel van dit onderzoek was om te kijken of impliciete tests als bijdrage aan expliciete tests kunnen gebruikt worden om user experience (UX) in de toegepaste wetenschap te meten. Hiervoor werd voor een tweezijdig studie design gekozen: In een voorstudie werden twee versies van dezelfde software onderzocht met betrekking op hun UX en usability om stimulus materiaal voor de navolgende tests van de hoofdstudie te verkrijgen. Hierin werden de attitudes van 43 proefpersonen met betrekking op de prototypes onderzocht door middel van twee impliciete tests, de Affect Misattribution Procedure (AMP) en Approach-Avoidance taak (AAT) zowel als de twee expliciete tests, de UEQ-meCUE questionnaire en Self-Assessment-Manikin (SAM) valence scale. Met betrekking op de onderzoeksvraag van deze studie werden er geen betekenisvolle correlatie tussen de impliciete tests gevonden. Dit impliceert dat er geen relatie bestaat tussen de scores van de tests. Tegelijk liet de correlatie tussen de impliciete tests ook geen betekenisvolle resultaten zien, waardoor het niet mogelijk was om de twee tests te valideren. Om tot deze conclusies te komen, werd elke test apart

bekeken: De analyses lieten zien dat de expliciete tests hogere gemiddelde UX-schattingen voor de UX-prototype vertoonden, wat overeenkomt met de veronderstelde verwachtingen. Echter konden de impliciete tests deze tendentie niet weergeven, wat betekent dat de impliciete tests in deze context en met dit stimulus materiaal niet in staat waren om de UX-verschillen te meten. In som laten de expliciete tests de verwachte verschillen in de manipulatie van UX zien, separaat van de usability van het product. Echter waren de impliciete tests niet gevoelig genoeg om de verschillen te detecteren en kunnen daarom geen bijdrage aan de expliciete tests in deze studie leveren.

1 Introduction

The early years of research in the field of human-computer interaction (HCI) were spent around terms, such as ‘efficiency’ and ‘usability of products’ (Bargas-Avila & Hornbæk, 2011). Within two decades, a new research focus has emerged and been recognized as a novel movement within the field of research in HCI: this focus emphasizes a holistic perspective on technologies, devices and products, as it includes traditional factors, e.g. usability, functionality or task efficiency, as well as additional qualities that put emphasis on the emotional components of an experience with technology (Bargas-Avila & Hornbæk, 2011). However, despite the shift from traditional usability terms to UX, the methods utilized until now have not yet been revised fundamentally: the explicit measurement techniques remain influenced by factors such as attribute substitution or missing validity and reliability estimations. Though, in different fields of psychology, the pitfalls of these techniques have been discussed and gradually addressed by the development of implicit measures, the techniques in UX research remained restricted to explicit methods, such as interviews or self-report measures. Therefore, this study will deal with the question of how to apply a new measurement paradigm as well as what such measures can contribute to explicit methods.

1.1 A Definition of User Experience

Despite the amount of research that has been conducted on UX, a concise definition of this term is still lacking. The ISO 9241-210 on human-centered design defines UX as a concept that considers “all the users’ emotions, beliefs, preferences, perceptions, physical and psychological responses, behaviours and accomplishments that occur before, during and after use.” (ISO 9241-210 2.15). However, this definition lacks to define terms precisely, e.g. ‘all the emotions’, which hinders the process of making the concept measurable (Spröll, Peissner & Sturm, 2010). Other definitions are more precise on terms concerning UX: for example, Jordan (2002) proposes that UX is composed of functionality, usability and four different kinds of pleasures, namely ‘physiological’, ‘sociological’, ‘psychological’ and ‘ideological’

pleasure. Besides, Battarbee and Forlizzi (2004) introduce UX as compounded by three types of experiences, 'Experience', 'an experience' and 'co-experience', while Norman (2004) present UX as composed of three types of designs: 'visceral design', 'behavioral design' and 'reflective design'. The most commonly used definition in the field of HCI, however, comes from Hassenzahl (2003): He focuses more than the previous definitions on particular aspects of emotions and their incorporation into products and devices to induce a particularly good experience. Additionally, he proposes that UX is more a matter of emotionality for human beings rather than only task-oriented, as he defines UX as a "momentary, primarily evaluative feeling (good-bad) while interacting with a product or service" (Hassenzahl, 2003, p.12). He adds that a good UX is constituted by the fulfillment of as many underlying psychological needs and values as possible (Hassenzahl, 2008; Sproll, Peissner, Sturm & Burmester, 2010). Therefore, he proposes a dichotomous model of UX which "assumes that people perceive interactive products along two different dimensions" (Hassenzahl, 2007, p.10): hedonic and pragmatic.

Hedonic qualities put emphasis on the emotional and affective components of a product, such as the enhancement of a person's psychological wellbeing (Hassenzahl, 2003; Väänänen-Vainio-Mattila, Roto & Hassenzahl, 2008; Bargas-Avila & Hornbæk, 2011). Hassenzahl (2003) as well as Hassenzahl and Tractinsky (2006) propose that hedonic qualities are composed of three aspects: stimulation, identification and evocation. The former refers to the notion that "individuals strive for personal development" (Hassenzahl, 2003, p. 5). In order to create a positive UX, products must fulfill people's needs by providing "new impressions, opportunities, and insights" to stimulate the evolvement of new experiences (Hassenzahl, 2003, p.5; Sproll, Peissner, Sturm & Burmester, 2010). The second aspect deals with social components of an interaction (Hassenzahl & Tractinsky, 2006): Hassenzahl (2003) as well as Hassenzahl and Tractinsky (2006) describe that people's possessions, such as a smart phone, a watch or a car, are used to "express their self through [these] physical objects" (Hassenzahl, 2003, p. 5). The third aspect deals with the memories a person makes or has made in the past (Hassenzahl, 2003). Products are presumed to be means to store or provoke meaningful memories made in the past (e. g. a video of a child's first smile) (Hassenzahl, 2003). Hassenzahl and Tractinsky (2006) add that such memories contribute to a person's actualization of the self. Furthermore, Hassenzahl (2008) describes 'hedonic' as 'be-goals' or needs of a person, such as "being related to others" or "being special" (p.12).

Pragmatic qualities, on the other hand, center on the idea of the product's functionality as well as usability and task-efficiency, meaning that the product shall serve as means to an end for the human operator to manipulate his/her environment (Hassenzahl, 2003; Väänänen-Vainio-Mattila, Roto & Hassenzahl, 2008). Hassenzahl (2003; 2007) as well as Hassenzahl and Tractinsky (2006) mention that it is important for a product to fit to the behavioral goals

of a user in order to support him/her meaningfully and successfully. A product should thus incorporate “functionality (i.e., utility) and ways to access this functionality (i.e., usability)” (Hassenzahl, 2003, p. 4). Additionally, the pragmatic component of UX is associated with peoples’ achievement of their particular ‘do-goals’ (Hassenzahl & Tractinsky, 2006; Hassenzahl, 2007; Hassenzahl, 2008). Such do-goals could include desired activities, e.g. “making a phone call” (p. 12) or ‘searching the web for particular information’ (Hassenzahl, 2008). Additionally, Hassenzahl (2008) states that the pragmatic qualities form a prerequisite for hedonic goals to evolve at all: being able to accomplish a desired activity (e. g. solving a difficult calculation) eases the fulfillment of the corresponding hedonic goal ‘being competent’ (Hassenzahl, 2003). Furthermore, he enhances that do- and be-goals always have to be considered together (Hassenzahl, 2008).

1.2 Current Research Methods on User Experience and Limitations

Even though, the shift towards UX began 20 years ago, new methods to measure UX have not yet been considered, intensively. Bargas-Avila and Hornbæk (2011) state that researchers still utilize mainly traditional methods, which rely on the users having explicit access to their recently gained experiences with a product: 53% of all research is accomplished with questionnaires, such as Likert Scales and similar, 20% of all studies use semi-structured interviews to deduce the experiences of the user and 15% make use of focus groups, in which a group of users discusses strengths and weaknesses of products they are given (Bargas-Avila & Hornbæk, 2011).

However, the practices associated with the usage of these methods have been criticized recently. In their literature review, Bargas-Avila and Hornbæk (2011) found that 51% of the studies analyzed utilize self-developed questionnaires whose items are not made publicly available in the articles. In doing so, these authors prevent other researchers from reusing the established measures, which is necessary for the validation of measures. According to Cohen and Swerdlik (2010), the validation of a test incorporates the repeated gathering of information on the validity of that tests over time, by means of assessing the content, criterion-related and construct validity. However, by means of not making items or questionnaires available, the assessment of validity falls short. An interesting example of such practices is mentioned by Bargas-Avila and Hornbæk (2011): Lankes et al. (2008) fall short of providing the reader with any information on the “standardized questionnaire” (p.255) they used, such as reliability and validity estimates or the name of the questionnaire. Consequently, it decreases the credibility of the study enormously.

Additionally to such practices, Nosek, Hawkins and Frazier (2011) criticize the traditional methods as such: They indicate that the measures rely on experiences and access to these that people might actually not have. Further, they explain that people lack motivation,

opportunities, abilities or even awareness with respect to reporting their experiences in questionnaires (Nosek, Hawkins and Frazier, 2011). Likewise, Kahneman (2002) also investigated this problem and proposed the phenomenon of ‘attribute substitution’. This implies that whenever people are not able to instantly answer a question posed, because they might not have an opinion on that particular topic, or they think the question is too difficult for them to answer, they will answer another question that is perceived more easy (Kahneman, 2002). Thus, people will always give an answer, but sometimes, it is not the answer that fits to the question asked. Kahneman (2002) states further that, in doing so, people will bias and distort results of those questionnaires (e.g. Likert scales) and thereby affecting the validity and reliability of measure.

Strasser, Weiss and Tscheligi (2012) highlight that self-report measures are at risk of being answered wrongly, either intentionally (faking) or unintentionally. Also, they searched for an explanation of this phenomenon: They found that people are often not able to express what they recently experienced. However, by trying to verbalize these, “they distort their answers” (p. 243) and thus bias the test results. Besides this, Devezas and Giesteira (2014) add that self-report measures are prone to social-desirability, which is, according to Adams et al. (2005), the “tendency of individuals to portray themselves in keeping with perceived cultural norms” (p. 389), and self-presentation motivations describing “the process in which people control how they are perceived and evaluated by others (Leary, Tchividjian & Kraxberger, 1994, p.461). A measure being prone to this, might lead to false conclusions due to unauthentic answering of the self-reports (Devezas & Giesteira, 2014). Van de Mortel (2008) states that social desirability poses a severe threat to the construct validity of a test, as it influences the way in which people answer, which, in turn, influences how well the test can measure the construct it is intended to measure. In sum, these above-mentioned limitations claim for a new type of measurement in HCI and applied UX research.

1.3 New Methods for Assessing UX

In other fields of psychology, the demands to overcome these aforementioned deficits in measurements have led to the emergence of implicit measures: First approaches to assess implicit vs. explicit memory as well as the construct of implicit cognition have been made in cognitive psychology. Later, the social psychology branch referred to these constructs and practices and tried to establish own methods to access and assess implicit attitudes in individuals. This resulted in the development of the Implicit Association Task (IAT) by Greenwald and Banaji (1995). These authors state that a clear advantage of implicit measures is that they assess the required information as such that the participant neither must know about what he is being assessed on, nor by relying on any self-report techniques or requirements at all. By avoiding these techniques, Nosek, Hawkins and Frazier (2011) state

that researchers are able to address the previously mentioned pitfalls such as attribute substitution or non-accessibility of recently made experiences, because implicit measures are not affected by these concepts at all.

Likewise, Bar-Anan and Nosek (2014) mention that implicit measures infer cognitions or experiences without the need for introspection: the participant does not have to verbalize, comprehend, interpret and recall the experiences. Thereby, these methods address the limitation given by Strasser, Weiss and Tscheligi (2012). Besides, Nosek, Hawkins and Frazier (2011) explain that in implicit measures, the construct is “inferred through a within-subject experimental design: [by] comparing behavioral performance between conditions (e.g. different primes)” (p. 154), which means that the same participant has to accomplish multiple different tests after another, whose results are compared: These conditions in implicit measures merely consist of two contrary extremes, such as positive vs. negative, soft vs. alcoholic drinks or calmative vs. anxious stimuli. Behavioral performance is then determined by the comparison of response latencies or categorization across the conditions (Tractinsky, Cokhavi, Kirschenbaum & Sharfi, 2006; Nosek, Hawkins & Frazier, 2011). From this, the attitude is then inferred.

To my knowledge, there are three studies in the context of HCI that already use experimental implicit measures to assess UX. The first study by Strasser, Weiss and Tscheligi (2012) was conducted to measure participants’ affections towards robots by making use of the Affect Misattribution Procedure (AMP). In this, participants are presented, one by one, with emotionally-loaden prime pictures shortly before pictures of Chinese characters appear. These latter shall in turn be rated by the participant (Payne, Cheng, Govorun & Steward, 2005). The authors divided 30 participants across two groups and showed videos of robots moving or being static to them. Afterwards, the participants received a questionnaire as explicit and the AMP as implicit measure. Aim of this study was to assess the applicability of the AMP in Human Robot Interaction (HRI). The study found that the implicit measures revealed a negative tendency towards a certain type of robots, while the utilized explicit measures did not show this tendency.

The second study by Schmettow, Noordzij and Mundt (2013) was conducted to investigate participants’ implicit associations and attitudes towards technical devices, such as computers or tablets. 41 participants were asked to perform the Stroop priming task as implicit measure, in which they had to react to colored words of three categories (hedonic, utilitarian and geekism), after seeing a picture of a technical devices. Afterwards, their need-for-cognition level was compared with the latencies retrieved from the task. The study’s main aim was to examine the suitability of the Stroop priming task to extend current methods. The results supported this notion in such that the task could even help to assess associations more directly than traditional methods: “implicit [...] methods [...] may serve to better understand

the nature of rating scales in HCI, and give more direct access to users' spontaneous associations and affects" (Schmettow, Noordzij & Mundt, p. 2046).

In the third study, Devezas and Giesteira (2014) compared implicit and explicit measures in order to obtain an estimation of how well implicit measures can be applied to HCI. They assessed the aesthetic judgment of eight participants concerning the 'valence' and 'self-identification' towards pictures of different interfaces, by using the Picture implicit association test: the P-IAT. Two bipolar scales were used as explicit measures for valence and self-identification. Devezas and Giesteira (2014) reported a "medium" correlation ($r = .42$, $p > 0.05$) between implicit and explicit measures (p.15). The authors argue that this is reason to believe that implicit measures can even work as "complementary or substitutive method for self-report measures" (p.15).

1.4 Main Aim of the Study

On the basis of the aforementioned arguments, a number of requirements for an additional measurement paradigm of UX have been established to account for the described pitfalls: First, it should not require participants' conscious introspection. Second, it should account for social-desirability and self-presentation motives and be able to deal with wrong answers in an acceptable manner, and third, it should incorporate considerable validity and reliability to avoid the usage of ad hoc constructed questionnaires.

The discussed literature shows that implicit measures incorporate multiple advantages in comparison to explicit measures, as they bypass many of the pitfalls reported with explicit measures, such as social desirability, missing validity and reliability estimations, or not being able to express recent experiences (Devezas & Giesteira, 2014). Though a control for these pitfalls would call for a replacement of explicit by the implicit measures, the nature of UX requires an addition of implicit to explicit measures rather than a replacement.

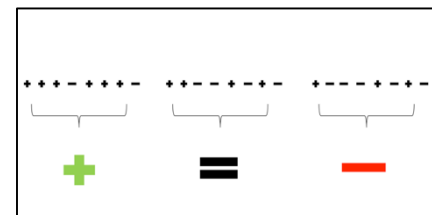


Figure 1: Implicit Measures' Insights to the Formation of Users' Opinions Concerning a

Hassenzahl, Diefenbach and Göritz (2010) describe experience as "a stream of feelings, thoughts and action; a continuous commentary on our current state of affairs." (p. 353), which suggests that experience is constituted especially by the subjectivity of the person. Besides, Hassenzahl (2003) as well as Hassenzahl and Tractinsky (2006) state that experience is also influenced by social factors, such as identification with a product: by means of possessing certain products, people present themselves to others thereby creating emotions of being admired or liked for that. This, in turn, leads to the conclusion that social desirability or self-presentation motivates are part of experience. Therefore, it is thought that implicit measures can be utilized to assess the construct under consideration *additively* to the results retrieved by explicit measures: Implicit measures can, for example, contribute to the assessment of a

product's UX as such, as they can reveal those moments during the interaction that are meaningful for the formation of a user's overall rating of the product at hand retrieved from explicit measures. This particular information could be used for iterative design processes, in which several parts of the product could be assessed by means of such tests. Therefore, the current study will exploratorily investigate the relationship between implicit and explicit measures. By means of a correlational analysis of both, it will be examined to what extent implicit measures can provide additional information to explicit measurement results. Besides, this study will utilize *two implicit* instruments for the sake of validation. The assessment of both tests' criterion-validity will be carried out by correlational analysis of the mean ratings per stimulus on both tests (Cohen & Swerdlik, 2010).

An additional unique selling proposition of this study is that the participants are able to actively interact with and explore the prototypes designed for this study. Prior studies, such as Strasser, Weiss and Tscheligi (2012) or Devezas and Giesteira (2014), have examined the applicability of implicit measures by utilizing video material or pictures to familiarize participants with the device or software at hand. However, the active interaction shall produce deep experiences that are rich in emotions and connected to particular moments which shall in turn enhance the applicability of implicit measures.

1.5 The Measures

To assess UX implicitly, it is necessary to determine two tests from the field of implicit tests. The determination of the two tests was directed by the requirements established before. Additionally, as most but not all implicit tests work with reaction times (RTs) (Payne & Lundberg, 2014), it was a prerequisite choose one test working with RTs and one without.

The first chosen implicit test is the AMP task by Payne et al. (2005). In this task, the participants saw four different stimuli displayed consecutively: a picture (for 75ms), a mask (125ms), a Chinese character (100ms) and a pattern mask. Afterwards, they were instructed to rate the Chinese character according to their visual pleasantness by clicking on either of the buttons for 'pleasant' or 'unpleasant' (Payne et al., 2005). Payne et al. (2005) found that the valence of the priming picture influences the ratings of the Chinese character significantly. Due to these promising results, the same presentation times were chosen for this study. The test was chosen on the basis of its good validity tests: "the AMP predicted behavior with an average effect of $r = .35$ " (Payne & Lundberg, 2014, p. 674). Additionally, Payne and Lundberg (2014) report promising reliability estimates: a Cronbach's alpha ranging from .49 to .95 (average: .81) in the consulted studies and a split-half reliability ranging from .37 to .92 (average: .58). Besides, the AMP was chosen as it is one of few implicit measures that does not rely on reaction times, but on the categorization of the stimulus material (Bar-Anan & Nosek, 2014). Furthermore, Bar-Anan and Nosek (2014) proposed that the AMP is more

suitable than other implicit measures for assessing other than social constructs, because it does not mention the assignment categories, explicitly.

The second chosen test is the Approach-Avoidance Task (AAT) by Rinck and Becker (2007). The task requires the participants to respond by pulling or pushing a joystick to a certain picture format, either landscape or portrait. According to Heuer, Rinck and Becker (2007) as well as Wiers, Rinck, Dictus and van den Wildenberg (2009), every stimulus of this test contains a valence ranging from positive to negative. Wiers et al. (2009) add that pulling (arm flexion) is linked to a positive interpretation of the stimulus, while pushing (arm extension) is related to a negative interpretation of a stimulus. The AAT bases its functionality on the assumption that reaction times are shorter for compatible (pull positive/push negative) trials, and longer than incompatible (pull negative/push positive) trials. The AAT has been chosen, because it incorporates active movement representing positive or negative reactions. The inclusion of this is thought to enhance the emotional connectedness to the stimuli. Additionally, Rinck and Becker (2007) have shown the test's good Spearman-Brown reliability estimates: $r = .71$. Besides, the AAT has been used extensively in therapeutic studies and yielded promising results with respect to social anxiety, phobias or alcohol disorders (Heuer et al., 2007; Rinck & Becker, 2007).

The first explicit test, the Self-Assessment Manikin (SAM) by Lang, Greenwald & Bradley (1988), contains three dimensions of to measure human emotions: valence, arousal and dominance. Each scale shows five figures as well as four in-between spaces resulting in a 9-point Likert scale ranging from '1' to '9' (Bradley & Lang, 1994). It has been used as a non-verbal method to assess "emotional responses in a variety of situations, including reactions to pictures, images, [...]" (Bradley & Lang, 1994, p. 51). Here, only the valence scale will be utilized, as this scale measures the hedonic aspects of UX best. The emotions assessed range from 'happy' or 'pleased' ('1') to 'unhappy' or 'annoyed' ('9') (Bradley & Lang, 1994). The SAM was chosen in this study on the basis of its ability to measure subjective emotions and assess these in different kinds of populations, as well as its usage and promising results in combination with the IAPS pictures (Bradley & Lang, 1994; 2007). Additionally, according to Bargas-Avila and Hornbæk (2011), the SAM is the most commonly utilized measurement technique in UX research to assess emotions.

The second explicit test is the mCUE questionnaire by Minge and Riedel (2013), which consists of 3 modules that measure different components of UX by utilizing a 7-point Likert scale. A selection of the items from the positive and negative emotions modules was made as these fit best to an interaction with prototypes instead of an end-product (items: AP.1-3, and AN.1-3, see Minge & Riedel, 2013). The mCUE was chosen on the basis of good validity and reliability estimates reported by Minge and Riedel (2013). The two chosen modules from mCUE correlated considerably with the valence scale of the SAM: a positive

correlation ($r = .65$) between ‘valence’ and ‘positive emotions’ and a negative correlation ($r = -.66$) between ‘valence’ and ‘negative emotions’. This is of particular importance, as the SAM will represent the explicit measures during the correlation with the implicit measures.

The last test is the ‘User Experience Questionnaire’ (UEQ) by Laugwitz, Held and Schrepp (2006). It assesses UX through 26 items on a 7-point Likert scale, which are grouped on six factors: “attractiveness”, “perspicuity”, “efficiency”, “dependability”, “stimulation” and “novelty” (Laugwitz et al., 2006, p. 63). According to Hinderks, Schrepp, Rauschenberger, Olschner, and Thomaschewski (2012) and Hinderks et al. (2012), this questionnaire is used to assess pragmatic and hedonic components of UX as well as the positive or negative attitude towards end-products. Laugwitz et al. (2008) report good internal consistency ($\alpha > .73$) with the UEQ. Additionally, these authors conducted first studies on the validity of the UEQ and found promising results: in a series of usability testings, these authors correlated the individual scales of the UEQ with the time the participants needed to accomplish a given task and found significant correlations ($r > -.54$) that were in accordance with the hypotheses established.

1.6 The Development of Stimulus Material

As both implicit tests presuppose the usage of dichotomous stimulus material, two versions of the same software tool were required to produce such stimulus material. In this study, it is required to measure UX separately from usability. As proposed by Hassenzahl (2003), hedonic qualities presuppose the existence of pragmatic qualities to evolve, it is important that the tools contain the same estimation of pragmatic qualities, but differ in their estimation of hedonic qualities. To establish such material, multiple apps were considered, but neither of them fulfilled the aforementioned requirements. For most applications, either the basis for comparison was missing or one of the applications contained bad usability. For these reasons, a new software with two versions (‘the two prototypes’) were designed.

The two prototypes were based on the ideation tool by Sonnleitner, Pawlowski, Kässer and Peissner (2013): it was developed for a previous study to show that the inclusion and stimulation of particular user needs through design features lead to positive experiences with the product. As this study’s focus lies within the investigation of the overall emotional UX rather than individual user needs, no specific user needs were included in the design of the two prototypes. Rather, the version that was assumed to create a positive UX, was enhanced by including design aspects and features that were developed on the basis of the experience categories by Zeiner, Laib, Schippert and Burmester (2016). In their study, Zeiner et al. (2016) have examined the emergence of positive experience in work contexts with and without technical devices. The context of positive emotions at work is especially valuable, here, as the interaction with the two prototypes was planned to take place in a ‘working

environment' scenario (description, see Appendix 6.2). The implementation of these functionalities was undertaken by means of an iterative design process in which multiple concept versions of the prototypes were generated, tested and adjusted, first on paper and later by means of the prototyping tool Axure RP 8. The basic functionality of the prototypes can be described as the following: the two prototypes are note applications which enable and support people in the generation of ideas about certain topics. The tools provide them with a predetermined set of topics, one after another, and give them opportunities to write down their ideas in text boxes. The two prototypes differ in the amount of functions and support included during the idea generation process, as described below.

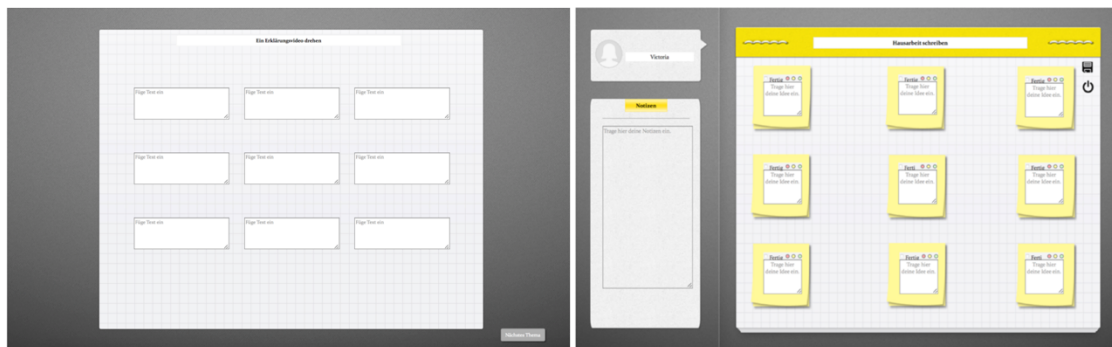


Figure 2: nUX (left) and UX Prototype (right) - Interaction Page

1.6.1 The nUX Prototype

The neutral-UX (nUX) prototype was created with a clean and clutter-free outer appearance. The utilized colors of this prototype were greyscale. The initial interaction page was designed in an unobtrusive graph paper and did not incorporate any exceptional functions except for being a note pad. The participants could, therefore, only type in associated keywords or sentences corresponding to the topic given by the header above the note pad (see Figure 2, left). If being finished, they could go on with the next topic by clicking the button stating 'next topic'. Likewise, this prototype included only one of the experience categories given by Zeiner et al. (2016), which is 'Finishing a task'. This rather sterile approach was chosen in order to develop a pure, but neutrally valenced prototype that functions only as a means to an end for the users, thereby creating good usability estimates, but no other (positive) emotions and experiences.

1.6.2 The UX Prototype

The basic functionality concept remained the same for the UX prototype. This prototype, though, contained additional functions and possibilities developed on the basis of the experience categories by Zeiner et al. (2016). First, the prototype included possibilities to personalize its outer appearance, such as by choosing a color, choosing an avatar and entering

a name. The latter, especially, served the purpose of greeting each participant with its own name at the beginning of the interaction to increase the emotional connectedness between the prototype and the participant. Additionally, participants were able to prioritize their generated ideas by means of buttons that would color the post-its correspondingly to the clicked button's color: 'green', 'yellow' and 'red' for the priorities ranging from 'not yet important' to 'urgent' (see Figure 2, right). Besides, the participants were appraised for the ideas they already generated: during the interaction, appraisal pop-ups appeared stating, e.g. how good the generated ideas were. Lastly, the participants were able to save the current map of ideas and send it to someone if they wished to do so. The inclusion of these categories into the design is assumed to lead to an increase of the positivity of the UX. Thereby, it is expected that the pictures are more positively valenced to the user. For an overview of the included categories and their corresponding (visual) implementation in the prototype, see Table 1.

Table 1. Experience categories included in the UX prototype

Experience Category	Implementation in the UX prototype
Receiving feedback	<i>Pop-ups</i> : pop-ups displayed at random during the interaction give feedback on the user's ideas
Appreciation	<i>Favorite color</i> : participants can choose their favorite color (personalization) <i>Greeting</i> : participants are greeted with their whole name to create a connection between the product and the user <i>Profile picture</i> : participants can customize their profile picture in order to create a connection between the product and the user
Keeping track of things	<i>Note pad on the left</i> : participants can add notes that do not fit on post its, or that they might come up with and have not yet grouped
Prioritizing	<i>Priority post-its</i> : participants can prioritize the ideas they generated by coloring the post-its in red, yellow and green for major, medium & lower importance
Exchanging ideas	<i>Send maps via e-mail</i> : participants are able to distribute their idea maps via e-mail which enables them to work together with others
Stimulating experience	<i>Send maps via e-mail</i> : by means of sending own ideas to others who can work on the same map, it is possible to socialize and to improve ideas through feedback of others
Aesthetics	<i>Outer appearance</i> : the prototype looks appealing to the participants; free from clutter
Finishing a task	<i>Text fields</i> : these give participants space for their ideas <i>'Ready' button</i> : participants are able to tick ideas they are done with and thereby disable the text field

1.7 Cross-Correlational Assumptions

For the exploratory investigation of the possible contribution of the implicit to the explicit tests, a correlation across the three measures per stimulus will be conducted testing the following assumption: If the implicit measures add a contribution to current methods of measuring UX, a correlation between the AAT and AMP shall reveal a closer relationship between these than between the AMP and the SAM and the AAT and the SAM. In order to estimate the construct-validity, the AAT and the AMP will be correlated. As both tests measure attitudes implicitly and utilize the same dichotomous rating scale (pull/push or pleasant/unpleasant), the following assumption is presumed: The mean RTs per stimulus of the AAT correlate strongly (≥ 0.85) with the mean proportions per stimulus on the AMP.

1.8 Pre-Requisites for the Cross-Correlational Analysis

To reach the aforementioned analyses steps, an exploratory design containing multiple sub-assumptions has been established (for the detailed assumptions, see Appendix 6.1). First, the manipulation of the prototypes was assessed in a pre-study by means of the UEQ and meCUE questionnaires. If successful, the individual tests will be examined concerning their functionality during the main study: the AAT, AMP and SAM. The manipulation checks per test function as guidelines for the further analyses.

1.8.1 Pre-Study

Manipulation Assumption – Prototypes

As the two prototypes were based on the same underlying tool, the following assumptions are expected (derivation of values, see Appendix 6.1): Both prototypes yield a mean rating of ‘4’ for their usability estimation. For their UX, it is assumed that the UX prototype receives a UX estimation of ‘5.5’, while the nUX prototype yields an estimation of ‘4’.

Method

In order to test whether the manipulation of the two prototypes is successful, a two-phased pre-study was conducted: a user study and a UX expert review. Throughout these, the software versions were iteratively changed and improved resulting into the prototype versions that were be used in the main study.

Participants. For the user review, five potential end-users (3 female; $M_{Age} = 23.8$, $SD_{Age} = 1.64$) as well as four UX experts (1 female; $M_{Age} = 32.8$, $SD_{Age} = 2.99$) were recruited at the Fraunhofer Institute for Industrial Engineering IAO in Stuttgart, Germany.

Material. The user review was conducted on a 21.5 inch iMac and the application was launched in Google Chrome. The expert review took place at the work space of each expert. Their review was sent to the experimenter after finishing the review.

Tasks. The users had to process through the two interactions with the prototypes. For a task description, see ‘User Scenario’ described in Appendix 2. After each interaction, they received two questionnaires to be filled in. The walkthrough of the prototypes for the experts was similar to the user review, but they had to set the focus on different objectives: instead of actually generating ideas about the given topics, they were asked to elaborate on remarkably positive or negative aspects during the interactions. After each interaction, the experts had to fill in both questionnaires and mark those experience categories by Zeiner et al. (2016), they found in the recently experienced prototype. For a full task description, see Appendix 2.

Procedure. For both groups, users and experts, the procedure followed this approach: first, they all received an instruction explaining the tasks of this study. Second, the first interaction took place. Then, the users and the experts filled in the two questionnaires as well as the experts marked the experience categories. Afterwards, the same procedure followed for the second prototype.

Results

To be able to test the prototype assumptions, the results of the meCUE and UEQ questionnaires were examined. In order to analyze the data, several items of both questionnaires had to be rescaled (full description, see Appendix 3.3).

Quantitative Results. For the quantitative analysis, a Wilcoxon signed-rank test was performed to compare the differences in means per factor of each questionnaire (meCUE: 6; UEQ: 5; full description, see Appendix 3.3.1). However, these results have to be considered carefully, as this pre-study’s sample size is only very small: $n = 5$ (users) and $n = 4$ (experts). For the users, an overall tendency for the nUX prototype to receive higher mean ratings than the UX prototype was found, which is contrary to the established prototype assumptions. Few factors assumption-conform (see Table 2): ‘aesthetics’, ‘use intention’ and ‘overall rating’ (meCUE) as well as ‘attractiveness’, ‘novelty’ and ‘stimulation’ (UEQ). These showed slightly higher mean ratings for the UX prototype, though, none of these were significant at a significance level of $\alpha = .05$.

Table 2. Descriptives of the meCUE and UEQ

Test/Factor	User Review				UX Experts			
	Mean		SD		Mean		SD	
	UX	nUX	UX	nUX	UX	nUX	UX	nUX
<u>meCUE</u>								
Usability	5.80	6.60	.84	.43	5.38	5.75	.48	.50
Aesthetics	<u>3.40</u>	3.07	.60	.64	3.42	4.66	.50	.72
Usefulness	3.27	3.73	1.88	1.50	3.67	4.25	1.09	.32
Use Intention	<u>2.50</u>	2.30	1.37	1.35	<u>2.88</u>	2.50	1.38	.58
Pos. emotions	2.87	2.90	1.36	1.07	<u>3.38</u>	3.33	1.11	.95
Neg. emotions*	5.20	4.33	1.02	1.50	4.84	4.65	1.35	1.25
Overall	<u>3.84</u>	3.82	.69	.34	3.93	4.19	.45	.43
<u>UEQ</u>								
Attractiveness	<u>3.90</u>	3.70	.81	1.05	4.29	4.96	.53	.76
Perspicuity	5.60	6.05	.82	.94	3.88	5.50	1.09	.54
Novelty	<u>3.80</u>	3.30	1.04	.82	<u>4.13</u>	3.81	.48	.43
Stimulation	<u>3.95</u>	3.85	.82	1.07	4.31	4.69	.24	.94
Dependability	5.00	5.70	.92	.74	4.00	5.13	1.20	.78
Efficiency	4.95	5.10	.69	1.47	3.71	6.00	1.23	.35
Overall	4.54	4.62	.43	.69	4.05	5.01	.47	.40

*On this factor, lower mean ratings are desired; underlined values represent an accordance with the expected assumption

The analysis of the expert data reflects the same tendency, but even stronger: only three of the 15 factors analyzed argue in favor of the UX prototype and show higher mean ratings for this: ‘use intention’ and ‘positive emotions’ (meCUE) and ‘novelty’ (UEQ). However, similarly to the user data, all factors showed rather slightly than remarkably higher ratings for the UX prototype and neither of these three yields significance.

Qualitative Results. The qualitative analysis was accomplished as proposed in the overview of qualitative content analyses by Mayring (1988) (results, see Table C in Appendix 3.3.2). For the user reviews, there were two important findings taken from the UX prototype. First, the nature of the appraisal pop-ups formed a severe interruption for the interaction. Three participants were massively interrupted in their ideation process, as the pop-up repeatedly appeared when clicking into the text field. Second, the duration per prototype exceeded the expected duration of 15-20 minutes by more than 30 minutes. Therefore, the total number of topics reduced to two in the final study. For the expert review, there was one important remark about the UX prototype: the text in combination with the buttons of the ‘quit’-pop-up were misleading: The text of the pop-up was understood as such that a click on ‘yes’ implied a return from the pop-up to the current map. However, the opposite happened: the expert was directed to the feedback page and could not return to the map. This led to the creation of

negative emotions. Therefore, the text of the pop-up was changed carefully after the reviews by means of clarifying it.

Discussion

The main aim of this pre-study was to examine whether the prototypes evoked more positive emotions than the nUX prototype. Additionally, both were expected to yield similar estimations in both for their usability. For this, five users and four UX experts interacted with the two and assessed their UX and usability by means of the UEQ and meCUE. The experts also judged the included experience categories.

The quantitative analyses of the data revealed contradictory results with respect to the prototype assumptions: Most of the factors of the two questionnaires for both groups implied a higher and thus more positive rating for the nUX prototype rather than for the UX prototype. This implies that the manipulation of the prototypes has not been successful. Likewise, the qualitative analysis shows a similar tendency. The UX prototype revealed most problems during the interaction, while there are only few reported concerning the nUX prototype. This fact, in turn, leads the users and experts to experience more positive emotions with this latter prototype, because fewer negative emotions evolved during the interaction. This notion is also supported by Hassenzahl (2008): he states that pragmatic qualities “refer [...] to the product’s perceived ability to support the achievement of ‘do-goals’, such as ‘making a telephone call [...]’” (p. 12). Further, he claims that pragmatic qualities refer to the usability and utility of a product which are a pre-requisite for the fulfillment of be-goals, the hedonic qualities of UX: “lack of usability might impose a barrier to the fulfillment of [...] be-goals” (Hassenzahl, 2008, p. 12). Based on this definition, one could argue that the UX prototype received a more negative evaluation, because it showed a greater number of usability problems which left no room for the hedonic qualities to evolve and be fulfilled, thereby preventing the generation of a positive UX. Therefore, it is expected that the adjustment of the recommended design changes by the UX experts as well as the changes retrieved from the user reviews shall lead to a more positive UX in the UX prototype. What can be concluded from this pre-study is that the manipulation of the two prototypes has not worked as expected: the UX prototype receives generally lower estimations on its UX than the nUX prototype. Especially the qualitative analysis revealed that the UX prototype still incorporates potential for improvements, as most of all problems occur with this prototype. Therefore, it is expected that an adjustment of the recommended changes will lead to the confirmation of the investigated assumptions (overview of the most important adjustments, see Table A, Appendix 4). Accordingly, these assumptions must be tested again in the main study.

2 Method

2.1 Experimental Design

The experiment was conducted as a within-subject design with one independent variable, ‘prototype’, which had two levels: ‘UX’ and ‘nUX’. Additionally, the experiment included four dependent variables. The first consisted of the subjective perception of UX and usability, measured by the merged UEQ-meCUE. The second and third represented the implicit evaluation of the two prototypes based on the RTs assessed by the AAT and the implicitly given proportion of positive answers per prototype measured by the AMP. The last dependent variable was the valence rating each prototype retrieved by the SAM. Additionally, the order the prototypes were presented to the participants was counterbalanced (UX-first or nUX-first). Likewise, the keys (AMP), the responses to the picture format (AAT), the order of the AAT and AMP as well as the runs of the SAM (‘Run1’-first or ‘Run2’-first) were counterbalanced to control for potential order effects.

2.2 Procedure

Participants had to walk through four phases of the experiment: the interaction, the two implicit tests and the SAM (see Figure 3). In

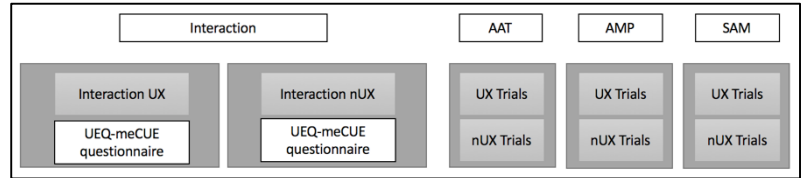


Figure 3: Overview of the Study's Procedure

advance to the study, the

participants were briefed about the setup of the electrocardiogram (ECG) and electroencephalogram (EEG) and the tasks. After signing the informed consent, the participants received the instructions for the 15-minutes interactions with the prototypes (Appendix 2.1). After each interaction, the participants filled in a paper version of the merged UEQ-meCUE questionnaire. Subsequently, the three tests were accomplished. The data collection time per participant for the whole experiment took approximately 90 minutes. At the end of the study, the participants were debriefed about the actual construct assessed, their implicit attitudes on the prototypes.

2.3 Participants

43 participants (24 females) were recruited via the internal participant database of the Fraunhofer Institute for Industrial Engineering IAO in Stuttgart, Germany, as well as via Facebook and flyers. Their mean age was $M_{Age} = 25.33$ years, $SD_{Age} = 4.47$ years. 39 participants were right handed and 39 participants gained a university's degree. Three participants passed their final exams and one received the ‘general certificate of secondary education’. The participants received a compensation of 15€ for their participation.

2.4 Material

2.4.1 Task Material

The interactions and the experiments were set up on two computers: the interactions were started on a Windows 7 laptop connected to an external screen (27 Inch), a qwertz-keyboard and a computer mouse. The AAT and the AMP were launched on a Windows 8 computer utilizing Inquisit 5 by Millisecond, while the SAM was launched in Matlab, Version R2014b, by MathWorks with two screens (27 Inch), one for the experimenter and one for the participant. Besides, another keyboard and a joystick were used. The ECG and EEG measured with three and 34 electrodes, respectively. However, the analysis of the ECG and EEG data is beyond the scope of this thesis.

2.4.2 Stimulus Material

For the choice of the stimulus material, the improved prototypes derived from the pre-study were taken as basis. For the nUX prototype, pictures from every step during the interaction were taken, as number of presented pages in this prototype was very small, resulting in nine pictures. For the UX prototype, a selection of 16 stimuli was made to receive approximately the same amounts of pictures. It was important to include stimuli with similar content to those of the nUX prototype, but to also include stimuli that would represent the design features based on by Zeiner et al. (2016). For an overview over the stimuli, see Appendix 6.2.

2.5 Tasks - Interaction with the Prototypes & UEQ-meCUE

For the both interactions, the participants received a scenario which they should imagine. Each of the two prototypes gave the participants two topics on which they should generate a minimum of six up to a maximum of nine ideas. Additionally, they should, if possible, cluster the ideas, save the document and send it to a colleague mentioned in the scenario. Afterwards, they could start the next topic. After each interaction, the participants received a pen-and-paper version of the UEQ-meCUE.

2.6 Measures

All tests contained IAPS pictures (Lang, Öhman & Vaitl, 1988) in the practice runs to explain the functionality of the test (AAT & AMP: 10 IAPS pictures, SAM: 6 IAPS pictures). The two experimental runs of all tests consisted of 128 pictures, 64 pictures each: 1x16 landscape and 1x16 portrait pictures of the UX prototype and 2x8 landscape and 2x8 portrait pictures of the nUX prototype were shown. The color of the UX prototype pictures was matched to the color chosen during the interaction.

2.6.1 AAT

In the AAT, the participants were presented with a screenshot from the interaction with both prototypes (see Figure 4). The UX-first group had to pull portrait and push landscape pictures, while the nUX-first group had to do so vice versa. A pull movement increased the screenshot's size and a push movement did so vice versa.

Each screenshot remained until the joystick was moved into one of the maximum positions. In the practice runs, an error message in red letters was shown to indicate a wrong movement. The participants could independently decide when to start the second run, by moving the joystick to the left.

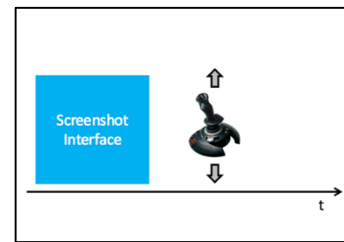


Figure 4: Schematic Representation of the AAT

2.6.2 AMP

Each trial of the AMP task consisted of four elements: a screenshot from the interaction (displayed for 75ms), a grey mask (125ms) to cover the screenshot, a Chinese character (100ms) and a pattern mask to cover the character (see Figure 5). Participants received a prepared keyboard on which the keys 'd' and 'k' were labelled with a happy and an unhappy smiley. The label was counterbalanced across the two groups: for group 1, 'd' represented the word 'pleasant' while it represented 'unpleasant' for group 2. By pressing one key, a new trial was started. The participant could independently start the second run by pressing the space bar.

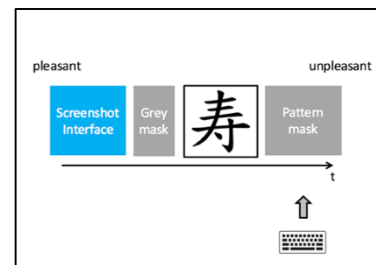


Figure 5: Schematic Representation of the AMP

2.6.3 SAM

For the SAM, the participants were instructed to use the keys '1' to '9' on the keyboard and to place their preferred hand in front of the keyboard and to always move back and forth from this position, see Figure 6. Before each trial, a white cross was displayed on a black screen to mark the beginning of the trial. Then a screenshot from either the UX or nUX prototype was presented for 3 seconds. After the screenshot, the two scales from the SAM were displayed one by one for a duration of 3 seconds each: firstly, the valence and then, the arousal scale. However, the arousal scale will not be taken into account within this thesis. After the arousal scale, the next trial was presented. In total, the SAM consisted of three runs: One practice run

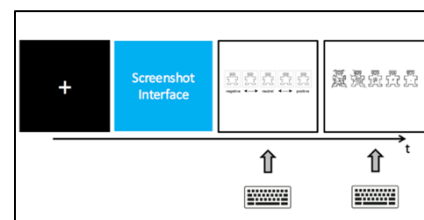


Figure 6: Schematic representation of the SAM

containing six IAPS pictures to explain participants the functionality of the test and two experimental runs, which were started in Matlab showing the same number and formats of pictures as the AMP and the AAT.

3 Results

3.1 Data Preparation

The following participants had to be excluded from the analysis due to changes in the setup of the experiment, too short exposure to the prototypes, errors or interruptions during the interaction and problems in understanding the task: 1, 4, 17 and 19. The analyses of the retrieved results were conducted in IBM SPSS 21 for Windows 8 and Mac OS X as well as in R version 3.2.2 for Mac OS X, making use of the packages “stats” and “graphics”.

3.2 Cross-Correlation – A Comparison of Implicit and Explicit Measures

In order to find out about the consensus between the results of the two implicit measures, as well as to examine the contribution of implicit to explicit measures, a cross-correlational analysis was conducted. Per test, the mean ratings per stimulus were calculated. These, in turn, were then correlated with the ratings of the other tests. With respect to Table 3, it becomes apparent that neither of the correlations between the three tests yielded remarkable and significant scores, even though both tests utilized the same stimulus material. The correlation between the AMP and the AAT was even slightly negatively correlated: $r = -.119$, $[-.4913; 0.2895]$. Likewise, the correlations between the AMP and the SAM ratings, as well as the AAT and the SAM ratings were insignificant. Figure 7 displays the cross-correlational plot, strengthening the results from the correlational analysis: there are no clear tendencies or patterns visible in the individual plots per test correlation.

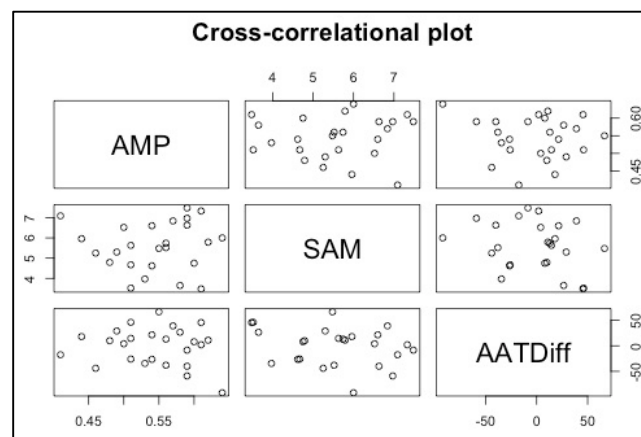


Figure 7: Cross-Correlational Scatterplot

Table 3. Cross-correlational Table

	AAT	AMP	SAM
AAT			
AMP	$\sigma^2 = -.262$ $[-0.4913; 0.2895]$ $r = -.119$		
SAM	$\sigma^2 = -11.007$ $[-0.5864; 0.1621]$ $r = -.249$	$\sigma^2 = .003$ $[-0.3533; 0.4354]$ $r = .049$	

3.3 Pre-Requisites for the Cross-Correlational Analysis

To reach the aforementioned conclusions, some prior steps were necessary: First, it was investigated by means of the UEQ-meCUE whether the manipulation of the prototypes was successfully adjusted from the pre- to the main study. For this, an exploratory factor analysis (EFA) was conducted to assess whether the factors retrieved from the questionnaires' handbooks were also found in this study (see Appendix 7.1). Hereafter, these factors were examined concerning the prototype assumption. Second, each of the implicit and explicit tests was considered individually and assessed whether it showed the expected tendencies.

3.3.1 Manipulation Checks

For the analysis of the individual tests, the following structure was defined: first, an exploratory data analysis (EDA) was conducted to visually inspect the data (see Appendix 7.2). Second, the statistical analysis followed to examine the differences in means of both prototypes. This study makes use of a Linear Mixed Effects (LME) model for this analysis, as required by the data: an Analysis of Variance (ANOVA) claims for independent and identically distributed variables. However, this assumption is violated by the repeated measures design utilized in this study: the subjects' attitudes are assessed at different times throughout the study. An LME model accounts for this violation by adding random factors. The model is thus constructed with two fixed effects, 'prototype' and 'order' and the random effect 'participant', as follows: Dependent Variable ~ Prototype + Order + (1|Participant).

UEQ-meCUE

The EFA of the UEQ-meCUE revealed two factors, 'hedonic qualities' and 'pragmatic qualities', which visually showed assumption-conform tendencies during the EDA. The statistical analysis confirmed these tendencies (see Table 4): on 'hedonic qualities', the UX prototype yields significantly higher mean ratings, as supported by the confidence interval

[.8229; 1.5227]. For the factor ‘pragmatic qualities’, however, the analysis revealed a slight difference between the prototypes. The nUX prototype receives a mean score of 5.72, while that of the UX prototype is .40 ratings *lower*. The confidence intervals for this difference even advocate for a significant difference: [-.6604; -.1262]. Additionally, the variance of the residuals implies that there is variance in the data that is unexplained by the factor ‘participant’. In sum, the manipulation of the prototypes was mainly successful: The UX prototype shows, indeed, higher means on the factor ‘hedonic qualities’, but lower than expected estimations of its ‘pragmatic qualities’.

Table 4. Coefficient table for the LME of the factor ‘hedonic and pragmatic qualities’

Groups	Hedonic qualities			Pragmatic qualities		
	Name	Variance	Std. Dev.	Name	Variance	Std. Dev.
Random effects						
Participant	(Intercept)	.2891	.5376	(Intercept)	.09096	.3016
Residual		.8746	.9352		.54497	.7382
Fixed effects						
	Estimate	Std. Error	t value	Estimate	Std. Error	t value
(Intercept)*	3.83526	.20366	18.832	5.7154	.1488	38.42
Prototype	1.17282	.21178	5.538	-.4026	.1672	-2.41
Order**	-.08967	.28051	-.320	.1198	.1984	.60

* β_0 is representative for the nUX prototype; **order in which the prototypes were shown

SAM

The EDA of the SAM reveals that the UX prototype receives noticeably higher mean ratings, which is confirmed by the statistical analysis below (see Table 5): the LME model presents a mean rating of 4.31 for the nUX prototype and a 1.60 higher rating for the UX prototype. This difference is even significant as stated by the confidence interval: [1.0497;2.1487]. According to the analysis of the random effects, the participants’ ratings are very homogeneous: $\sigma^2 <$

.000, which, however, is not supported this notion by the

EDA (see Figure 8). There is a general tendency for the UX

prototype to receive higher ratings, but there are also exceptions. Some participants show a reversed tendency of rating the nUX prototype more positively, suggesting that this effect is not very much representative. It can be observed that the lines of the participants are not parallel, implying no participant intercept random effect. In sum, the SAM showed clear tendencies that the UX of the UX prototype is more positive than that of the nUX prototype, though the expected values are only approached.

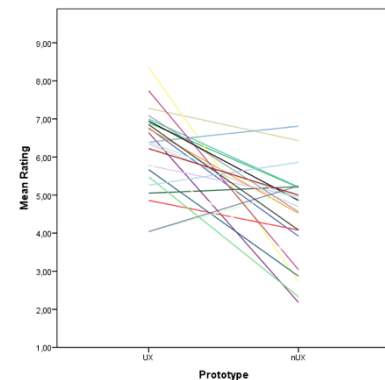


Figure 8: Spaghettiplot on the distribution of ratings per participant and prototype

Table 5. Coefficient table for the LME of the SAM Score

Groups	Name	Variance	Std. Dev.
<u>Random effects</u>			
Participant	(Intercept)	.000	.000
Residual		1.444	1.202
<u>Fixed effects</u>			
	<u>Estimate</u>	<u>Std. Error</u>	<u>t value</u>
(Intercept)	4.3146	.2984	14.458
Prototype	1.5992	.3399	4.705
Order	.4181	.3402	1.229

AAT

For the analysis of the AAT, the difference scores per participant and prototype were calculated by subtracting the RTs of “Pull” by those of “Push”. The EDA revealed slight differences between the two prototypes, which are disconfirmed by the LME model (see Table 6). The nUX prototype yields a difference score of -3.683ms, while that of the UX prototype is 3.736ms *higher*. However, this difference is insignificant: [-18.8332; 27.8813]. Likewise, the order effect of ‘prototype’ detected during the EDA is also insignificant: [-30.9273; 21.4020]. Additionally, there is much idiosyncratic difference between the RTs, suggesting that the participants’ reactions to the stimuli are very heterogeneous. In short, the AAT revealed no confirming evidences for the established prototype assumption.

Table 6. Coefficient table for the LME of the AAT

Groups	Name	Variance	Std. Dev.
<u>Random effects</u>			
Participant	(Intercept)	688.2	26.23
Residual		3835.7	61.93
<u>Fixed effects</u>			
	<u>Estimate</u>	<u>Std. Error</u>	<u>t value</u>
(Intercept)	-3.683	12.906	-.285
Prototype	3.736	14.598	.256
Order	-2.069	17.713	-.117

AMP

For the AMP, the proportion of positive ratings per participant ($n = 14$) and per prototype were calculated. The EDA showed similar medians but differing variances for the proportion of positive answers per prototype. The statistical analysis revealed that the proportions differ only marginally from each other: the UX prototype yields 0.2% more positive answers than the nUX prototype (see Table 7). This difference is also insignificant as suggested by the confidence interval, [-.0467; .0475]. The variances retrieved are likewise very small and

indicate a homogeneous tendency for the participants to rate. In sum, the manipulation check of the AMP revealed none of the assumed differences in positive ratings between the prototypes.

Table 7. Coefficient table for the LME for the AMP

Groups	Name	Variance	Std. Dev.
Random effects			
Participant	(Intercept)	.012768	.11300
Residual		.004776	.06911
Fixed effects			
	<u>Estimate</u>	<u>Std. Error</u>	<u>t value</u>
(Intercept)	.529301	.039517	13.394
Prototype	.002308	.027107	.085
Order	.074545	.094637	.0788

4 Discussion

4.1 The Research Question

The overarching research question of this study was to find out whether implicit measures are suitable additions to explicit measurement techniques to assess UX in applied research. To examine this question, a three-fold research design was constructed: First, the designed software tools were considered regarding the expected difference in rating the two prototypes. Second, a manipulation check per test was conducted to find out whether each individual test functioned properly and showed the expected results. Lastly, in order to investigate the contribution of implicit measures, a cross-correlational analysis between the implicit and explicit tests was conducted. However, this analysis mainly revealed contradictory results rather than those expected: the analysis showed that the expected relationships between the SAM and the two implicit tests were only minor and insignificant. This suggests that there are no relationships between the ratings per stimuli of the three tests. Therefore, it must be argued that the implicit tests cannot contribute to the results of the explicit tests in this study. Additionally, to investigate the implicit tests' construct validity, a correlation between the two implicit tests was conducted. However, this correlation was also insignificant, making a validation of the tests impossible.

4.2 Explanations

4.2.1 The Manipulation of the Prototypes

Explanations for the null-results of the correlations may have its roots in the manipulations of the prototypes and the derived stimulus material for the tests. However, this possibility seems rather unlikely, based on the efforts undertaken to develop the desired stimulus material: The

prototypes were based on the tool by Sonnleitner, Pawlowski, Kässer and Peissner (2013), which was successfully established to investigate the effects of user needs fulfillment on the experiences of the users with the devices. The further fine adjustments were based on the inclusion of design features retrieved from the experience categories by Zeiner et al. (2016). The prototypes were designed, tested and adjusted multiple times. Two complete prototype iteration cycles were passed through: In the first cycle, five potential end-users, and in the second cycle, four user experience experts evaluated the usability and experience of the whole product as well as the single functions based on the experience categories.

Support for the success of this technique comes from the results of the SAM and the UEQ-meCUE questionnaire. For the SAM, it was expected that the statistical analysis reveals higher mean ratings for the UX than for the nUX prototype, which was found to be true, thereby showing strong evidence for a successful manipulation. Likewise, also the results of the UEQ-meCUE questionnaire advocate for this. Higher mean ratings for the factor 'hedonic qualities' were expected to be found, which was confirmed by the statistical analysis: the mean ratings were smaller than expected, but the difference between the two prototypes showed a clear tendency into the expected direction. Both measures' results provide evidence for the successful manipulation of the UX separately from the prototypes' usability, which make it unlikely that the manipulation of the prototypes is responsible for the null-results retrieved in this study.

An investigation of the results shows that explanations come from the two implicit tests themselves, as these are the only measures showing the null-results. For example, for the AAT, it was expected that the UX prototype yields a negative and the nUX prototype a positive mean difference score, representing the tendencies of 'liking' and 'disliking'. However, the analyses showed that the scores of both prototypes revolved around 0ms, supposing that the RTs per condition 'Pull' or 'Push' per prototype relativize each other. Likewise, for the AMP, it was assumed that the UX prototype yields 75% positive answers, while the nUX prototype yields only 50% positive answers. Similarly to the results of the AAT, the visual and statistical analysis also showed no significant difference between the ratings of the prototypes: the proportion of positive answers was approximately equal across both prototypes. Thus, the null-results must have its roots in the implementation of the implicit measures.

4.2.2 The Role of Emotions

Support for this notion comes from an analysis of the contexts in which the implicit tests are mostly used: The AAT, for example, is utilized in therapeutic circumstances, such as treating addictive behaviors (Wiers, Rinck, Kordts, Houben & Strack, 2010; Wiers, Eberl, Rinck, Becker & Lindenmeyer, 2011) as well as anxiety disorders, in particular phobias (Rinck &

Becker, 2007) and social anxiety (Heuer, Rinck & Becker, 2007). Likewise, the AMP has been used successfully to assess political attitudes, self-esteem and racism stereotypes (Bar-Anan & Nosek, 2014) as well as people's attitudes towards affective pictures (IAPS Pictures, Lang, Öhman & Vaitl, 1988; Payne, Cheng, Govorun & Stewart, 2005). All these studies have led to promising results towards assessing attitudes by means of the AMP and the AAT. Linking this to the present results, it becomes apparent that the evaluation of the prototypes by the two tests lacks one aspect amongst others which is of major importance for their functionality: emotional connectedness of the individual to the topic. It appears that the participants are not as deeply connected as thought and that their emotions concerning the topic are too superficial. Indeed, the participants are thought to build up an emotional connectedness to the prototypes by means of the design aspects developed on the basis of the experience categories of Zeiner et al. (2016). However, these certainly do not incorporate as much emotional connectedness as pictures of anxiety-evoking vs. non-anxious stimuli (e.g. spiders vs. kittens) would do for someone suffering from spider phobia. In order to yield the expected results, the emotional connectedness must be enhanced, e.g. by factors such as longer interactions or products that are more meaningful to the person, e.g. social media or medication applications. Consequently, it can be concluded that the emotional experience produced by the prototypes and induced by the stimulus material is too low to be measured by the AAT and the AMP.

4.3 Limitations

Several issues to explain the null-results retrieved during the analyses of this study have been discussed above. In addition, the following limitations had an influence on the results.

The first limitation of this study concerned the EFA conducted on the UEQ-meCUE questionnaire: The analysis investigated whether the eight factors retrieved from the handbooks of the questionnaires, were also included within this new, merged version, which, however, was not the case: on first glance, the 32 items loaded on six different factors. Though, on second glance, it became apparent that all items loaded high on only two factors. This leads to questioning the discriminant validity of the factors mentioned by Minge and Riedel (2013), as well as Laugwitz, Held and Schrepp (2006). The factors appear to incorporate high correlations between each other, which implies that they reflect similar constructs. This study could only differentiate between the factors 'hedonic qualities' and 'pragmatic qualities'. Nevertheless, the questionnaire was used, as it represented the exact two factors that were the subject under examination during this study.

The second limitation of this study concerns the results retrieved on the factor 'pragmatic qualities'. It was expected that both prototypes yield similar estimations on this, but the nUX prototype yields slightly but even significantly better estimations than the UX

prototype. This lower fulfillment of the pragmatic qualities weakens the effect of the design aspects based on Zeiner et al. (2016) and thereby decreases the positivity of the UX induced by that prototype. Support for this phenomenon comes from Hassenzahl (2003): He argues that ‘hedonic qualities’ presuppose the fulfillment of ‘pragmatic qualities’ to evolve at all, which implies that the lower the usability of a product, the lower the positivity of the UX. This interference of the perceived usability, as already seen in the pre-study, could not be fully removed from this study and remains a problem, influencing the results obtained on the other tests as well. Drawing on this, it is important to modify and retest the prototypes to adjust the usability of the UX prototype before any further utilization of the material.

Third, the study incorporated a lot of missing values, which were due to technical problems during the main study. Through this, a varying number of stimulus ratings per participant on the SAM as well as the full data of 29 participants of the AMP was lost. Additionally, this limitation gave rise to another problem: During the debriefing, multiple participants mentioned they had rated the Chinese characters according to their aesthetical



Figure 9: 'Pleasant' (r) and 'Unpleasant' (l) Characters

construction. Those containing few and “gently curved” lines’ (Quote: Participant 28, see Figure 9) were rated as ‘pleasant’ and those containing many lines as were rated as ‘unpleasant’. Taking this into account, it must be argued that the AMP shows little construct validity, here. In fact, the test did not measure the construct it was intended to measure, namely the implicit attitude towards the stimuli derived from the prototypes.

An explanation for this comes from Bar-Anan and Nosek (2014): They argue that the AMP is particularly sensitive compared to other implicit tests to the time participants get to evaluate the stimuli presented. Linking this to the test design utilized, it becomes apparent that all times were restricted, just like in Payne et al. (2005), except for the duration of the pattern mask. This stayed until one of the designated buttons was pressed. Thus, the participants actually had an indefinite amount of time to think about the Chinese characters and their aesthetical construction. On the basis of these comments, it is logical to question the construction of the AMP in this study, as it could not replicate as promising results as in Payne et al. (2005) and Payne and Lundberg (2014). Additionally, although, the characters were randomized across participants, trials and prototypes, the small number of observations on this test gave rise to the aesthetical nature of the characters having more impact than expected.

Additionally, one issue concerning the usage of implicit tests in practical UX environments has emerged throughout the conduction of this study: the tests require the existence of clearly dichotomous stimulus material, which complicates the usage of the tests in applied research. Most prototypes do not exist in either of the two extremes (positive-

negative or positive-neutral). Mostly, only one version of the prototype exists that shall be improved by means of assessing its UX.

The fourth limitation of this study is that only the definition of UX by Hassenzahl (2003; 2008) has been utilized. To retrieve a more complete picture of UX, it is also necessary to broaden the view on it and, hence, to consult other definitions for additional perspectives on design elements that could be included in a prototype. For example, Jordan (2002) provides a differing view on the emotional component, as he defines it by four distinct pleasures: ‘physiological’, ‘sociological’, ‘psychological’ and ‘ideological’ pleasure. One particular aspect defined by Jordan (2002) as well as Forlizzi and Battarbee (2004) that is more deeply and distinctively considered than in the definition by Hassenzahl (2003; 2008) is the social aspect of UX, the ‘sociological pleasure’: herein, Jordan (2002) describes technology as a mediator of the relationships between people, e.g. an application like ‘Whatsapp’ or ‘Facebook’. The enhancement of the communication between people makes it easier to stay in contact with meaningful people, which, in turn, leads to the evocation of positive emotions when using the application (Jordan, 2002).

Another limitation concerns the AAT: although the trials, formats and prototypes were randomized across participants, this tests showed an insignificant but visible order effect. Additionally, there are two remarks with respect to the induction of emotions by the UX prototype: the first point concerns that the topics might differ in the emotional valence they elicit: ‘organizing a weekend trip’ might contain a more positive valence than topics such as ‘preparing to move to a different place’. Second, the UX prototype contains two motivators of which the latter was not as positive as the first which influences the judgment of the UX negatively due to a recency effect (Gleitman, Gross & Reisberg, 2011). An additional limitation of this study lies within the fact that there were no calibration tables available for the UEQ and meCUE. The estimations for the prototype assumptions were derived by literature about people’s behavior concerning Likert scales, e.g. ‘avoiding extremes’ or ‘central tendency bias’ (Bertram, 2007). The last point of concern deals with the (environmental) circumstances during the interaction: the wiring to the EEG and ECG and high temperatures in the lab ($\geq 25^{\circ}\text{C}$) might have led to induction of other emotions than planned (e.g. shame or aversion).

4.4 Future Prospects

On the basis of the aforementioned findings, it must be argued that the usage of the two implicit measures did not yield the expected results concerning the prototype assumptions. However, several suggestions for adjustments result from the conduction of this study: an adjustment of the UX prototypes’ usability, thereby improving its experienced UX, an enhancement of the emotional connectedness of the participants to the prototypes, for

example by means of letting participants interact longer in order to establish a personal history with the software at hand. Additionally, future research should choose for a software that incorporates more emotional value to the participants itself, such as social media or medication applications.

4.5 Conclusion

What can be concluded from this research is that the manipulation of the UX separately from the usability of the prototypes has worked, which is shown by the results of the explicit tests. Both show the expected higher mean ratings for the UX prototype. For the implicit tests, the crucial emotional connectedness of the individual to the topic was too low to influence their judgments noticeably. Additionally, the conduction of this study has shown that the dichotomous nature of the stimulus material that is required by the implicit tests make the application of implicit tests in practice rather difficult. As both utilized tests did not reveal any potential contributions to the results of the explicit measures used with this stimulus material and context, this study argues against the usage of implicit measures in practice.

5 References

- Adams, S. A., Matthews, C. E., Ebbeling, C. B., Moore, C. G., Cunningham, J. E., Fulton, J. & Herbert, J. R. (2005). The effect of social desirability and social approval on self-reports of physical activity. *American Journal of epidemiology*, 161(4), 389-398. doi: 10.1093/aje/kwi054
- Bar-Anan, Y. & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. *Behav Res*, 46, 668-688. doi: 10.3578/s13428-013-0410-6
- Bargas-Avila, J. A. & Hornbæk, K. (2011). Old wine in new bottles or novel challenges? A critical analysis of empirical studies of user experience. *Proceedings of the 2011 annual conference on Human factors in computing systems – CHI '11*. ACM Press, (2011), 2689-2698. doi: 10.1145/1978942.1979336
- Bradley, M. M. & Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal Behavior Therapy & Experimental Psychiatry*, 25 (1), 49-59. doi: 10.1016/0005-7916(94)90063-9
- Bradley, M. M. & Lang, P. J. (2007). The international affective picture system (IAPS) in the

- study of emotion and attention. In J. A. Coan & J. B. Allen (Eds.), *Handbook of Emotion Elicitation and Assessment* (pp. 29-46). Oxford, CA: Oxford University Press.
- Chan, K.W.L. & Chan, A.H.S. (2009). Spatial stimulus-response (S-R) compatibility effect for hand controls with visual signals on horizontal plane. *Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol II IMECS 2009, March 18-20, Hong Kong*.
- Cohen, R. J. & Swerdlik, M. E. (2010). *Psychological assessment and testing*. Singapore: McGraw-Hill Education (Asia).
- Devezas, T. & Giesteira, B. (2014). Using the implicit association test for interface-based evaluations. *ACHI 2014: The Seventh International Conference on Advances in Computer-Human Interactions*, 9-16.
- Forlizzi, J. & Battarbee, K. (2004). Understanding experience in interactive systems. *Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques*. ACM Press, (2004), pp. 261-268.
- Gleitman, H., Gross, J. & Reisberg, D. (2011). *Psychology, eighth edition*. New York, United States: North & Company.
- Greenwald, A. G. & Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological Review*, 102(1), 4-27. doi: 10.1037//0033-295X.102.1.4
- Hassenzahl, M. (2003). The thing and I: understanding the relationship between user and product. In M. A. Blythe, A. F. Monk, K. Overbeeke & P. C. Wright (Eds.), *Funology: from usability to employment* (pp. 1-12). Dordrecht: Kluwer Academic Publishers.
- Hassenzahl, M. (2007). The hedonic/pragmatic model of user experience. In E. Law, A. Vermeeren, M. Hassenzahl, & M. Blythe (Eds.), *Towards a UX Manifest – Proceedings of a cost294-affiliated work-shop on HCI 2008* (pp. 10-14).

- Hassenzahl, M. (2008). User experience (UX): towards an experiential perspective on product quality. *Proceedings of the 20th International Conference of the Association Francophone d'Interaction Homme-Machine on - IHM '08*. ACM Press (2008), 11-15. doi: 10.1145/1512714.1512717
- Hassenzahl, M., Diefenbach, S. & Göritz, A. (2010). Needs, affects, and interactive products – facets of user experience. *Interacting with Computers*, 22, 353-362. doi: 10.1016/j.intcom.2010.04.002
- Hassenzahl, M. & Tractinsky, N. (2006). User experience – a research agenda. *Behavior & Information Technology*, 25(2), 91-97. doi: 10.1080/01449290500330331
- Heuer, K., Rinck, M. & Becker, E. S. (2007). Avoidance of emotional face expressions in social anxiety: the approach-avoidance task. *Behaviour Research and Therapy*, 45 (2007), 2990-3001. doi: 10.1016/j.brat.2007.08.010
- Hinderks, A., Schrepp, M., Rauschenberg, M., Olschner, S. & Thomaschewski, J. (2012). Konstruktion eines Fragebogens für jugendliche Personen zur Messung der User Experience. In Brau, H., Lehmann, A., Petrovic, K., Schroeder, M. (Eds.), *Usability Professionals 2012* (pp. 78 – 83).
- ISO. 2009. *Ergonomics of human-system interaction – Part 210: Human-centered design for interactive systems 2.15*.
- Kahneman, D. (2002). Maps of bounded rationality: a perspective on intuitive judgment and choice. In T. Frangmyr (Ed.), *Les Prix Nobel: The Nobel Prizes 2002* (pp.449-489). Stockholm: Nobel Found.
- Lang, P.J., Öhman, A. & Vaitl, D. (1988). *The International Affective Picture System*. Gainesville, Florida: University of Florida.
- Laugwitz, B., Schrepp, T. & Held, M. (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.

- Leary, M. R., Tchividjian, L. R. & Kraxberger, B. E. (1994). Self-presentation can be hazardous to your health: impression management and health risk. *Health Psychology, 13* (6), 461-470. doi: 10.1037/0278-6133.13.6.461
- Mayring, P. (1983). *Qualitative Inhaltsanalyse, Grundlagen und Techniken*. Weinheim, Basel: Beltz Verlag.
- Minge, M. & Riedel, L. (2013). meCUE – Ein modularer Fragebogen zur Erfassung des Nutzungserlebens. In S. Boll, S. Maaß & R. Malaka (Eds.), *Mensch und Computer 2013: Interaktive Vielfalt*, (pp. 89-98). München: Oldenbourg Verlag.
- Minge, M., Riedel, L. & Thüring, M. (2013). Modulare Evaluation von Technik. Entwicklung und Validierung des meCUE Fragebogens zur Messung der User Experience. In E. Brandenburg, L. Doria, A. Gross, T. Güntzler & H. Smieszek (Eds.), *Grundlagen und Anwendungen der Mensch-Technik-Interaktion. 10. Berliner Werkstatt Mensch- Maschine-Systeme* (pp. 28-36). Berlin: Universitätsverlag der TU Berlin.
- Nosek, B. A., Hawkins, C. B. & Frazier, R. S. (2011). Implicit social cognition: from measures to mechanisms. *Trends in Cognitive Science, 14* (4), 152-159. doi: 10.1016/j.tics.2011.01.005
- Payne, B. K., Cheng, C. M., Govorun, O. & Steward, B. D. (2005). An inkblot for attitudes: affect misattribution as implicit measurement. *Journal of Personality and Social Psychology, 89*(3), 277-293. doi: 10.1037/0022-3514.89.3.277
- Rinck, M. & Becker, E. S. (2007). Approach and avoidance in fear of spiders. *Journal of behavior therapy and experimental psychology, 38*, 105-120. doi: 10.1016/j.jbtep.2006.10.001
- Schmettow, M., Noordzij, M. L. & Mundt, M. (2013). An implicit test of UX: individuals differ in what they associate with computers. *Proceeding of the thirty-first annual CHI conference on Human factors in computing systems - CHI '13*. ACM Press (2013), 2039-2048. doi: 10.1145/2468356.2468722
- Sonnleitner, A., Pawlowski, P., Kässer, T. & Peissner, M. (2013). Experimentally manipulating positive user experience based on the fulfilment of user needs. In P.

- Kotzé, G. Mardsen, G. Lindgaard, J. Wesson & M. Winckler (Eds.), *Human-Computer Interaction – Interact 2013* (pp. 555-562). Cham: Springer International Publishing AG.
- Sproll, S., Peissner, M. & Sturm, C. (2010). From product concept to user experience: exploring UX potentials at early product stages. *NordiCHI 2010, October 16-20, 2010*. Reykjavik, Iceland.
- Sproll, S., Peissner, M., Sturm, C. & Burmester, M. (2010). UX Concept Testing: Integration von User Experience in frühen Phasen der Produktentwicklung. In H. Brau, S. Diefenbach, K. Göring, M. Peissner & K. Petrovic (Eds.), *Usability Professionals 2010* (pp. 195-200). Stuttgart: Fraunhofer Verlag.
- Strasser, E., Weiss, A. & Tscheligi, M. (2012). Affect misattribution procedure: an implicit technique to measure user experience in HRI. *Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction – HRI '12*. ACM Press (2012), 243-244. doi: 10.1145/2157689.2157776
- Tractinsky, N., Cokhavi, A., Kirschenbaum, M. & Sharfi, T. (2006). Evaluating the consistency of immediate aesthetic perceptions of web pages. *International Journal of Human-Computer Studies*, 64, 1071-1083. doi: 10.1016/j.ijhcs.2006.06.009
- Väänänen-Vainio-Mattila, K., Roto, V. & Hassenzahl, M. (2008). Towards practical user experience evaluation methods. *Proceedings of the 5th COST294-MAUSE Open Workshop on Meaningful Measures: Valid Useful User Experience Measurement* (VUUM 2008). Reykjavik, Iceland.
- Van de Mortel, T. F. (2008). Faking it: social desirability response bias in self-report research. *Australian Journal of Advanced Nursing*, 25 (4), 40-48.
- Wiers, R. W., Rinck, M., Dictus, M. & van den Wildenberg, E. (2009). Relatively strong automatic appetitive action-tendencies in male carriers of the OPRM1 G-allele. *Genes, Brain and Behavior*, 8 (1), 101-106. doi: 10.1111/j.1601-183X.2008.00454.x

6 Appendix

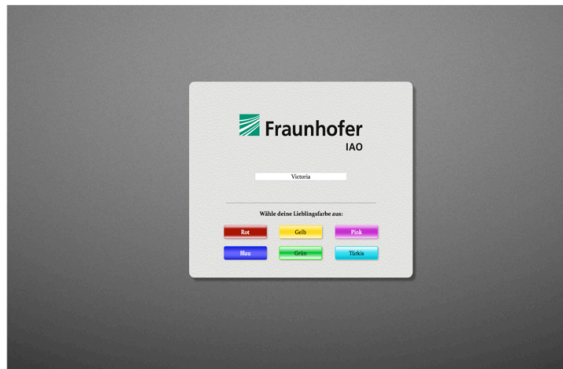
Table of Content

1 PROTOTYPES PRE-STUDY	3
1.1 UX PROTOTYPE	3
1.1.1 EXPERIENCE CATEGORIES AND THEIR CORRESPONDING FUNCTIONS – UX PROTOTYPE	4
1.2 NUX PROTOTYPE	6
1.2.1 EXPERIENCE CATEGORIES AND THEIR CORRESPONDING FUNCTIONS – NUX PROTOTYPE	6
2 PRE-STUDY – SCENARIOS & TASKS	7
2.1 USER TASK DESCRIPTION	7
2.2 EXPERT TASK DESCRIPTION	7
3 PRE-STUDY	8
3.1 MANIPULATION ASSUMPTION - PROTOTYPES	8
3.2 METHOD	8
3.2.1 PARTICIPANTS.	8
3.2.2 MATERIAL.	8
3.2.3 TASKS.	8
3.2.4 PROCEDURE.	9
3.3 RESULTS	9
3.3.1 QUANTITATIVE ANALYSIS	9
3.3.2 QUALITATIVE ANALYSIS	11
3.4 DISCUSSION	14
CONCLUSION.	15
4 ADJUSTMENTS OF UX PROTOTYPE	16
5 PROTOTYPE MAIN STUDY	17
5.1 UX PROTOTYPE	17
5.2 NUX PROTOTYPE	18
5.3 STIMULUS MATERIAL PER PROTOTYPE	19
6 MAIN STUDY	19
6.1 ASSUMPTIONS	19
6.1.1 PROTOTYPE ASSUMPTION	19
6.1.2 SAM-RATINGS	20
6.1.3 AAT	20
6.1.4 AMP	20
6.2 SCENARIO	21
6.3 R – CODE	22
6.3.1 ANALYSES	22
6.3.2 PLOTS	23
7 RESULTS OF THE MAIN STUDY	26
7.1 VALIDATION OF MERGED UEQ-MECUE QUESTIONNAIRE	26
7.2 EXPLORATORY DATA ANALYSIS	28
7.2.1 UEQ-MECUE	28
7.2.2 SAM	30
7.2.3 AAT	31
7.2.4 AMP	33
8 REFERENCES – PICTURES	34
8.1 OWLS	34
8.2 FIGURE 5&6 – KEYBOARD	34
8.3 FIGURE 4 – JOYSTICK	34
8.4 FIGURE 6 – SAM (VALENCE/AROUSAL)	34
8.5 UX PROTOTYPE – PRE-STUDY (APPENDIX 1)	35
8.5.1 QUIT + SAVE (INTERACTION PAGE)	35
8.5.2 APPRAISALS (APPRAISAL PAGE)	35

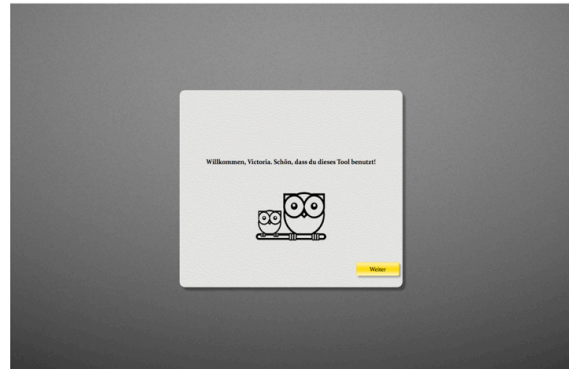
1 Prototypes Pre-Study

1.1 UX Prototype

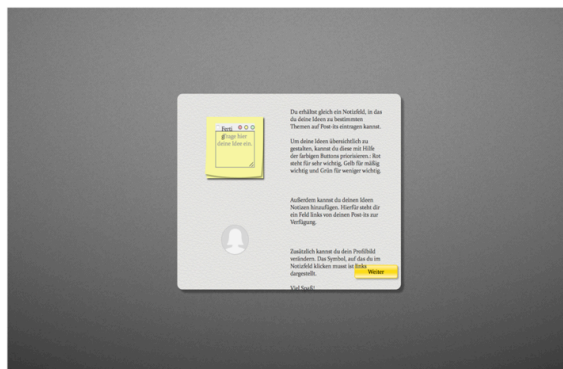
Beginning



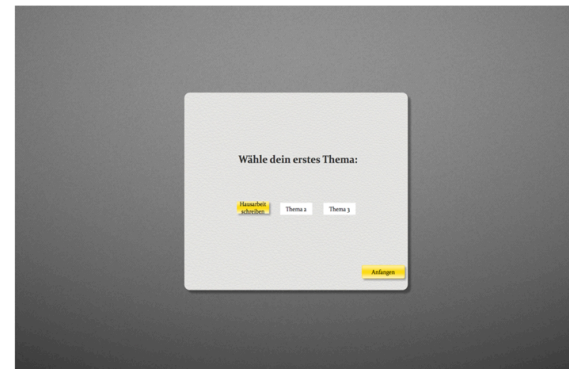
Greeting Page



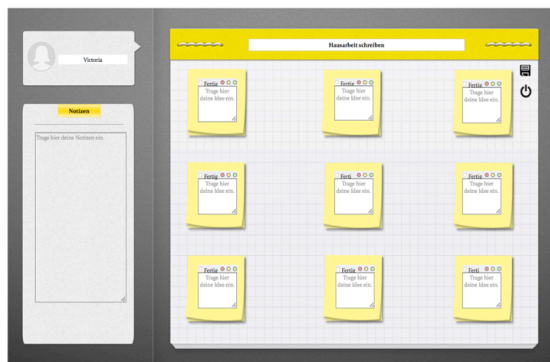
Tutorial Page



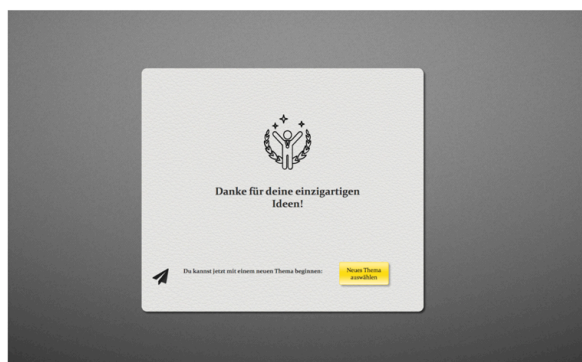
Topic Page



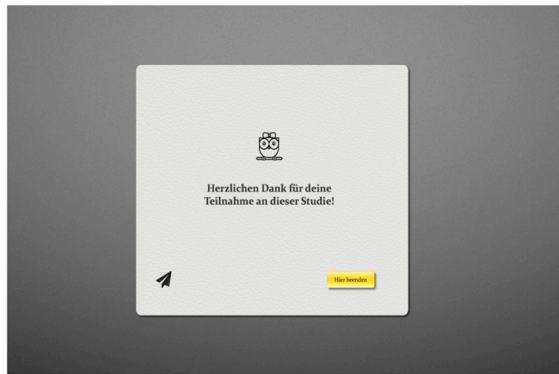
Interaction Page



Appraisal Page

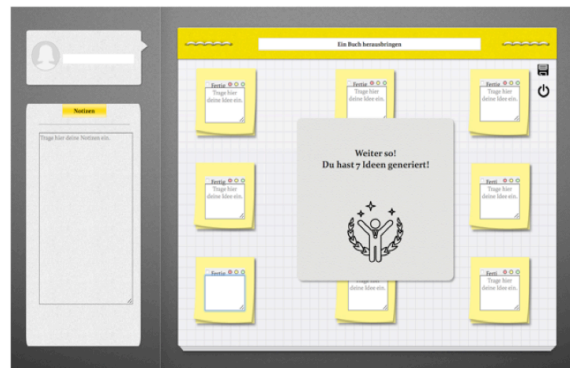
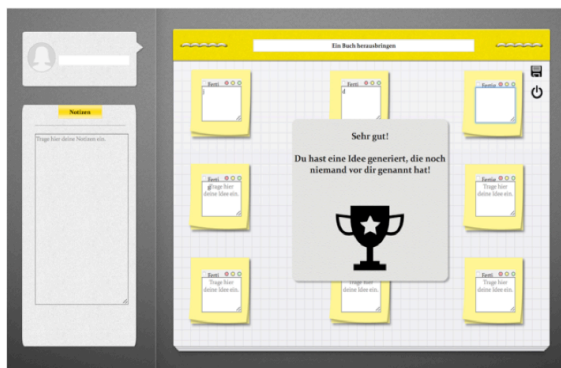


End Page



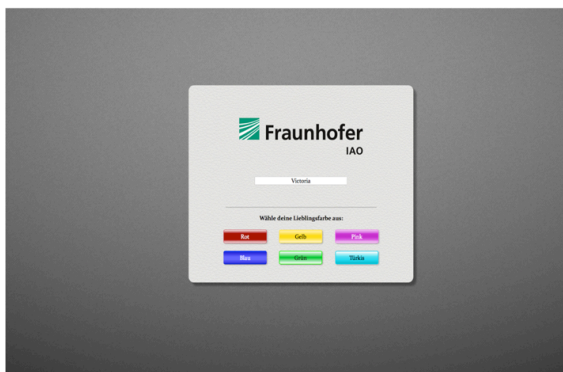
1.1.1 Experience Categories and Their Corresponding Functions – UX Prototype

Receiving feedback through pop-ups

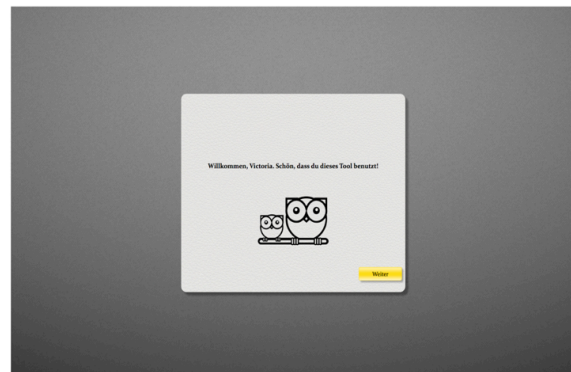


Appreciation

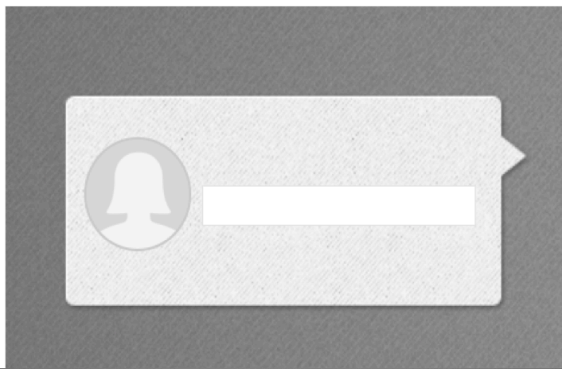
a) Changing Colors



b) Greeting



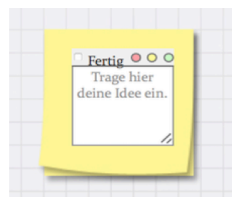
c) Profile Picture



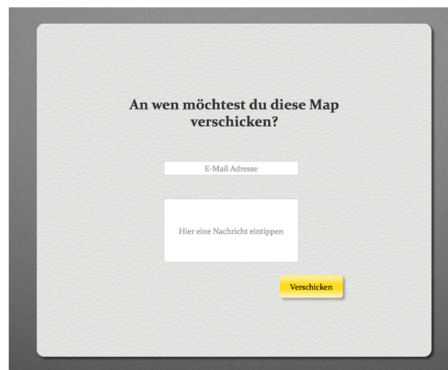
Keeping Track of Things



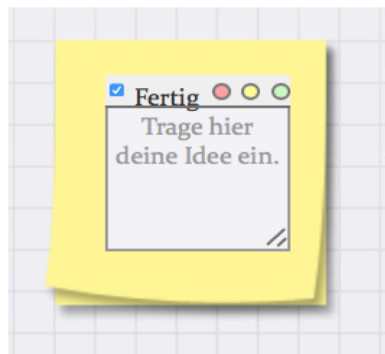
Prioritizing



Exchanging Ideas & Stimulating Experiences



Finishing a Task



- Text fields
- 'Fertig' Button

Aesthetics

For this experience category, there is no picture available as it involves the appearance of the whole prototype.

1.2 nUX Prototype

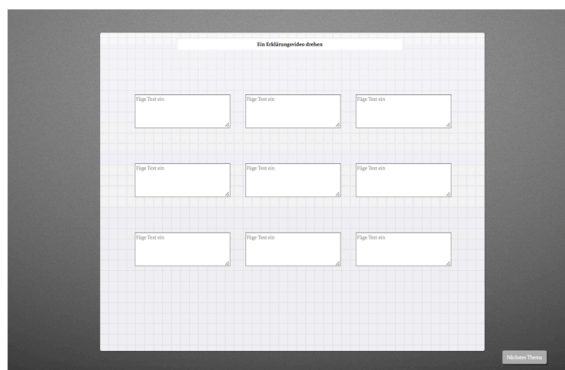
Beginning Page



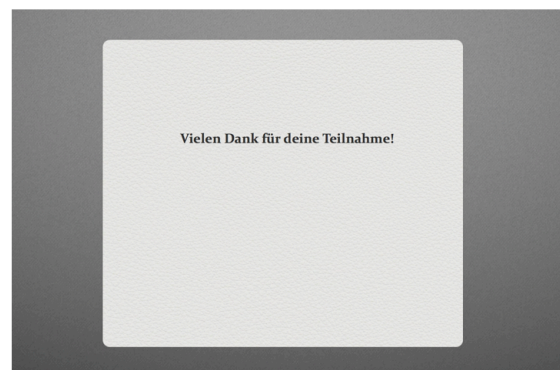
Topic Page



Interaction Page

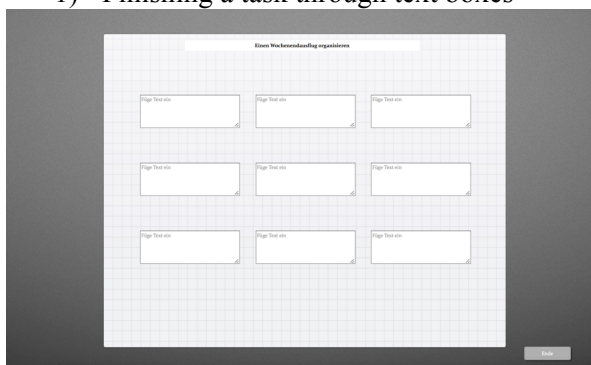


End Page



1.2.1 Experience Categories and Their Corresponding Functions – nUX Prototype

1) Finishing a task through text boxes



2 Pre-Study – Scenarios & Tasks

2.1 User Task Description

(Equal to Main Study – Scenario)

2.2 Expert Task Description

Das Review beginnt mit einer Interaktion des Prototypen B.

Die Probanden der eigentlichen Studie erhalten für diesen Prototypen die folgenden Aufgaben:

- 1) Den Hinweisen des Prototypen folgen (Namen eingeben, etc.)
- 2) Thema auswählen
- 3) Mindestens 6 bis maximal 9 Ideen pro Thema generieren
- 4) Die Ideen speichern
- 5) Die Ideen verschicken an: @iao.de
- 6) Neues Thema auswählen

Hinweise für das Experten Review mit Prototyp B:

- Bitte der Interaktion des Programms folgen
- Hierbei ist es weniger wichtig, Ideen zu den jeweiligen Themen zu finden, sondern auf Dinge während der Interaktion zu achten, die positiv oder negativ auffallend sind.
- Diese Punkte bitte während der Interaktion aufschreiben und beschreiben, was positiv oder negativ daran wahrgenommen wurde.
- Nach der Interaktion bitte die Erlebniskategorien zur Hand nehmen, --> bitte die Kategorien aus diesem Dokument löschen, die nicht im Prototypen B vorkommen
- MeCUE Fragebogen + UEQ Fragebogen ausfüllen (auf Papier)

Nächster Prototyp: Prototyp G

Hinweise für die Probanden: Den Hinweisen des Prototypen folgen

Hinweise für das Experten Review mit Prototyp G:

- Bitte der Interaktion des Programms folgen
- Ebenso, wie bei B, auf positive und negative Elemente während der Interaktion achten, diese aufschreiben und beschreiben
- Nach der Interaktion bitte die Erlebniskategorien zur Hand nehmen, --> bitte die Kategorien aus diesem Dokument löschen, die nicht im Prototypen G vorkommen
- MeCUE Fragebogen und den UEQ Fragebogen ausfüllen (auf Papier)

3 Pre-Study

3.1 Manipulation Assumption - Prototypes

Both prototypes were based on the same underlying tool, a note application in which participants are able to generate ideas on a set of predetermined topics. With respect to the usability of the two prototypes, it is assumed that they yield similar estimations of usability: both prototypes shall function equally well and be able to fulfill the do-goals, the users of the prototypes have during the task (Hassenzahl, 2008). Therefore, it is expected that the ratings of both prototypes on the UEQ and meCUE questionnaire center around '4', as this marks the neutral position of the 7-point Likert scale. With respect to their estimation of UX, the design of the two prototypes differed by the number of experience categories by Zeiner et al. (2016) included: the UX included more. Therefore, it is assumed that the UX of this latter prototype is also more positive than that of the nUX prototype. Thus, it is expected that the UX prototype yields a mean rating of '5.5', as this marks the central position between the neutral '4' and the positive extreme '7'.

In order to test whether the manipulation of the two prototypes was successful, a two-phased pre-study is conducted. Five user reviews are performed to evaluate the usability, determine the duration of the interaction and identify individual preferences for either of the two prototypes. The prototypes are iteratively changed and improved throughout the user reviews. Afterwards, four UX experts assess the final versions of both prototypes from the user reviews. Afterwards, the last modifications according to the comments of the experts are accomplished and the resulting versions are used during the experiment of the main study.

3.2 Method

3.2.1 Participants.

For the user review, five participants (3 female, $M_{Age} = 23.8$, $SD_{Age} = 1.64$) as well as four UX experts (1 female, $M_{Age} = 32.8$, $SD_{Age} = 2.99$) from the Fraunhofer Institute for Industrial Engineering IAO participated in this pre-study.

3.2.2 Material.

The user review was conducted on a 21.5 inch iMac and the application was launched in Google Chrome. The expert review took place at the work space of each expert. Their review was sent to the experimenter after finishing the review.

3.2.3 Tasks.

The users had to walk through the two prototype interactions and follow the instructions given on paper, which were handed to them before the study (see 'User Scenario', Appendix 8.2.1). For a visual representation of such an interaction, see Appendix 8.1. After each interaction, the participants

received two questionnaires to be filled in. Then, they could start with the next interaction. For the UX experts, the walkthrough the prototypes was similar to the users, but they had to set the focus on different objectives: instead of actually generating ideas about the given topics, they were asked to elaborate on remarkably positive or negative aspects during the interactions. After each interaction, the experts had to fill in both questionnaires and mark those experience categories by Zeiner et al. (2016), they found in the recently experienced prototype. For a full task description, see Appendix 8.2.2.

3.2.4 Procedure.

For both groups, users and experts, the procedure followed this approach: first, they all received an instruction explaining the tasks of this study. Second, the first interaction took place. Then, the users and the experts filled in the two questionnaires as well as the experts marked the experience categories. Afterwards, the same procedure followed for the second prototype.

3.3 Results

In order to be able to test the hypothesis of whether the reviewers evaluated the UX prototype more positively than the nUX prototype, the results of the meCUE and UEQ questionnaire were examined. The analysis of the quantitative data from the user as well as the expert review began with rescaling the negatively formulated items of both questionnaires; the factor ‘negative emotions’ of the meCUE as well as the following items on the UEQ: items 2, 5 and 6 on the factor ‘aesthetics’, items 2 and 4 on ‘perspicuity’, items 1 and 2 on ‘novelty’, items 1 and 4 on ‘stimulation’, as well as items 3 and 4 on ‘dependability’ and items 1 and 4 on ‘efficiency’. Hereafter, the descriptives for both prototypes on the two tests were examined, separately for the user and the expert review.

An inspection of the residuals implied a non-normal distribution and advocated for a non-parametric test. Therefore, a Wilcoxon signed-rank test was performed to compare the differences in means of the UX and nUX prototypes per factor of each test (meCUE: 6, UEQ: 5). For all subsequent tests of the user and expert review, a significance level of $\alpha = .05$ was used in order to determine the statistical significance of the obtained test results. However, with respect to the small sample sizes worked with in this pre-study, the results as such as well as any significances in differences have to be considered carefully and are rather considered as tendencies in ratings. The qualitative data retrieved from the reviews was analyzed separately. The procedure to do so is described in the corresponding sections for user reviews and UX expert reviews.

3.3.1 Quantitative Analysis

Descriptives and Wilcoxon Signed-Rank Test. Table X below shows the descriptives of the user and expert reviews for each of the two questionnaires. With respect to the users, it becomes apparent that that only three of the seven factors from the meCUE show the expected tendency in higher mean

ratings for the UX than for the nUX prototype: ‘aesthetics’, ‘use intention’ and the ‘overall’ rating of the meCUE. The remaining factors of that questionnaire show a reversed tendency, as such that the nUX receives higher mean ratings than the UX prototype. This latter notion is even supported by means of the obtained statistical results (see Table X). The Wilcoxon signed-rank test supposes a significant difference between the two prototypes on the factor ‘usability’, $Z = -2.041$, $p = .041$, as well as ‘negative emotions’, $Z = -2.032$, $p = .042$. For the UEQ, also three of seven factors reveal higher mean ratings for the UX prototype: ‘attractiveness’, ‘novelty’ and ‘stimulation’, however, an inspection of the statistical analysis reveals no significant tendencies among these factors. Likewise to the meCUE, the remaining factors show higher mean ratings for the nUX prototype, but also, none of these reaches significance (see Table 12).

Table A. Descriptives UEQ/meCUE for the User and Expert Reviews

Test/Factor	User Review				UX Experts			
	Mean		SD		Mean		SD	
	UX	nUX	UX	nUX	UX	nUX	UX	nUX
<u>meCUE</u>								
Usability	5.80	6.60	.84	.43	5.38	5.75	.48	.50
Aesthetics	<u>3.40</u>	3.07	.60	.64	3.42	4.66	.50	.72
Usefulness	3.27	3.73	1.88	1.50	3.67	4.25	1.09	.32
Use Intention	<u>2.50</u>	2.30	1.37	1.35	<u>2.88</u>	2.50	1.38	.58
Pos. emotions	2.87	2.90	1.36	1.07	<u>3.38</u>	3.33	1.11	.95
Neg. emotions*	5.20	4.33	1.02	1.50	4.84	4.65	1.35	1.25
Overall	<u>3.84</u>	3.82	.69	.34	3.93	4.19	.45	.43
<u>UEQ</u>								
Attractiveness	<u>3.90</u>	3.70	.81	1.05	4.29	4.96	.53	.76
Perspicuity	5.60	6.05	.82	.94	3.88	5.50	1.09	.54
Novelty	<u>3.80</u>	3.30	1.04	.82	<u>4.13</u>	3.81	.48	.43
Stimulation	<u>3.95</u>	3.85	.82	1.07	4.31	4.69	.24	.94
Dependability	5.00	5.70	.92	.74	4.00	5.13	1.20	.78
Efficiency	4.95	5.10	.69	1.47	3.71	6.00	1.23	.35
Overall	4.54	4.62	.43	.69	4.05	5.01	.47	.40

*Lower values are better for this factor

With respect to the experts, the meCUE reveals the expected higher mean ratings for the UX prototype only on two of the seven factors: ‘use intention’ and ‘positive emotions’, see Table 13. However, especially the difference in means on the latter factor, is only marginal, as also supported by the great p-value on that factor advocating for a clear insignificance of that difference: $Z = -.365$, $p = .72$. Of the remaining other factors, neither of the differences between the mean ratings of the two prototypes reaches significance. Likewise, on the UEQ, only one factor ‘novelty’ shows the expected higher mean rating for the UX prototype, but, however, does not reach statistical significance. All other

factors advocate for a better rating of the nUX prototype, but also, none of these differences between the two prototypes is significant.

Table B. Z-Scores and p-Values for the UEQ/meCUE for the User and Expert Reviews

	Users		Experts	
	Z-score	p-Value	Z-Score	p-Value
<u>meCUE</u>				
Usability	-2.041	.04*	-1.089	.28
Aesthetics	-.552	.58	-1.064	.11
Usefulness	-.674	.50	-.921	.36
Use Intention	-.535	.59	-.816	.41
Pos. emotions	-.412	.68	-.365	.72
Neg. emotions*	-2.032	.04*	-.184	.85
<u>UEQ</u>				
Attractiveness	-.406	.68	-1.841	.66
Perspiciuity	-.944	.35	-1.604	.11
Novelty	-1.604	.11	-1.069	.29
Stimulation	-.365	.72	-.730	.47
Dependability	-1.890	.06	-1.300	.19
Efficiency	-.141	.89	1.826	.07
Overall	-.405	.69	-1.826	.07

*Lower values are better for this factor

3.3.2 Qualitative Analysis

User Review. The notes from the user review were collected in an Excel sheets. The qualitative analysis was accomplished as proposed in the overview of qualitative content analyses by Mayring (1988). First, all important facts mentioned were listed. Additionally, for each problem, it was tried to note a cause, a breakdown, an outcome and a design change, if possible. Then, they were sorted according to the similarities of their content (e.g. same page of the prototype, similar functionalities, corresponding context, etc.). Afterwards, headers for each of these categories were found and the urgency of each adjustment proposition was marked. The results of the user reviews can be found in Table 14. It shows a short description of the problem, the participants had, how many of the participants encountered the problem, a comment on the context and an adjustment (or proposal) that should be changed after the interaction or in future prototypes. Besides the fact that in every of the user review interactions errors occurred, there were two findings of major importance taken from the qualitative reviews. First, the nature of the appraisal pop-ups formed a severe interruption for the interaction of the users with the UX prototype. Participant 3 was massively interrupted in the generation process of the ideas, as the she kept clicking into the text field in order to be able to begin to type in an idea, which however, only led to the repeated appearance of the pop-up. Likewise, participant 1 and 2 were annoyed by the pop-ups and found their functionality irritating as it was blocking the text field. Second, it became apparent that the number of topics included in both

prototypes was too much. The participants were expected to deal with each of the prototypes an approximate duration of 15 to 20 minutes. However, it took all participants longer to finish the topics and interact with the prototypes. Therefore, for the main study, it was chosen to drop 3 of the topics per prototype yielding a total amount of two topics each in the final study. For the analysis of the expert review, the same procedure was followed.

Table C. Problems encountered by the users, potential solutions and adjustments.

Problem	Participants					Comments	Adjustments of the UX Prototype
	1	2	3	4	5		
Functional errors in the prototype	X	X	X	X	X	Not being connected to subsequent pages ill-working buttons, etc.	Adjustments after each PP: removing usability errors
Number of topics too much	X	X	X	X	X	PP1: 60 minutes; PP2: 130 minutes; PP3: 65 minutes; PP4: 45 minutes; PP5: 35 minutes	Topics were reduced: After PP2: 3 (UX) & 5 (nUX) topics; After PP4: 3(UX/nUX); After PP4: idea reduction to 6-9
Topics	X			X		PP1: Term ‘DIY’ was not understood; PP4: Some topics = more difficult (e.g. changing tires)	Term was changed into ‘explanatory’
Valence of topics evenly distributed	X			X	X	PP1: Topic ‘weekend trip’ = similar to ‘birthday party’; PP4: Emotion valence = unevenly distributed: ‘planning vacation’ = more positive than ‘job application’; PP5: UX had easier topics	
Pop-ups	X	X	X			PP1: Pop-ups were annoying and irritating; PP2: Pop-ups were annoying; PP3: The motivation pop-ups interrupted the interaction	Changed after PP3. Pop-ups = restructured into appraisal pages after the interaction
Pop-up content			X	X		PP3: opposite of motivation → frightening + pressurizing; PP4: Pressurized by content: It took me quite some time to come up with five ideas and the pop-up showed me that I still had to find almost the same amount again	Pop-up content was changed completely: motivators came at the end of the interaction page and thus displayed overall instead of specific appraisal
Profile picture unseen	X	X	X	X	X	‘Change your profile picture’ was not even seen by all users	Exchange for a profile picture that did not require any changes
Welcome/ Thank you page	X	X				PP1: Greeting ≠ displayed completely, → people also entered surname PP2: Privacy violated → ‘unique ideas’ implied that ideas had been compared + rated according to quality without permission	After PP1, hint was changed into ‘enter your first name’ Other: content was kept

Improvements to the Prototypes. Taking these experiences, those listed below and the results of the quantitative analysis into account, it becomes apparent that the UX experienced during the interaction with the two prototypes contradicted the UX that was expected: the nUX prototype yielded fewer usability problems and critique compared to the UX prototype. Likewise, the UX of the latter prototype is also not perceived as positive, but rather as negative. Though, the statistical results of the analysis have to be interpreted carefully due to the small sample size, they mirror the exact tendency of the participants to comment on the prototypes during the qualitative analysis. On the basis of the retrieved results, the adjustments, displayed in Table 3, were accomplished.

Expert Review. Like the notes from the user review, the comments of the expert reviews were also sorted according to their content. In total, for the UX prototype, five grouping factors of remarkable facts were found: themes, interaction pages, motivations, greeting page and functionality of the prototype. For the nUX prototype, only three factors were found: interaction page, outer appearance and functionality of the prototype. Afterwards, the comments were highlighted in three colors, green (not at all), orange and red (very much) according to their urgency to be changed in order to separate serious functionality errors from comments of mainly aesthetical nature. As the reviews revealed a wide range of different improvement remarks, especially with respect to the UX prototype, only the most important remark for the UX prototype is elaborated on at this point, see Table 15. The most important remark of the UX experts came from reviewer 4. He noticed that the text in combination with the buttons of the ‘quit’-pop-up were misleading: He understood the text as such that ‘yes’ implied that he could return to his current map in order to generate more ideas. However, contradictory to his expectations, the button ‘yes’ directed him to the feedback page right after the interaction page and left no possibility to return to his map and the ideas. This led to the creation of negative emotions, such as incomprehension as well as sadness. Therefore, the text of the pop-up was changed carefully after the reviews by means of clarifying the text in combination with the buttons.

Table D. Problems encountered by the UX experts, potential solutions and adjustments.

Problem	UX Expert				Comments	Adjustments
	1	2	3	4		
Topic has to be selected manually	X				Why selection of topic necessary, if only one works?	Only display one topic at a time and disable the selection-need
Topic buttons and clickable buttons looked alike		X	X	X	Clear difference between buttons and title displays	Outer appearance of the buttons was changed
Name receives a whole column	X				Why?	Added: Standard profile picture + adjective to describe the participant’s style of working
Save and quit buttons ≠ labelled + put somewhere invisible	X				Label buttons + place them somewhere more prominent	Buttons = placed at top right corner of interaction page; Symbols = explained in tutorial
Yes and No of Quit-Pop up = in wrong order	X		X		Change to Yes & No	Accomplished
‘Done’ button makes no sense, text is still editable	X			X	Disable text editing if ‘Done’ is selected	Checkboxes = deleted from interface → functionality = misinterpreted by all
Button of save popup should be on right side of the popup		X			Put button to the right (feels more natural)	Accomplished
Text on quit pop-up was misleading			X	X	Clarify the text	Accomplished
Appraisal untrustworthy (2x same text)	X				Increase credibility	Users = only praised once
Personal texts’ and button labels’ friendliness = incongruent	X				Personalize it	Changed into ‘Begin’
Why is profile picture not kept throughout all topics?			X		Keep profile picture	Profile picture = deleted + standard picture included

Experience categories. With respect to the experience categories, the UX experts had to indicate per prototype, it becomes apparent that the indicated categories are almost the same per prototype and therefore replicate the findings of the quantitative analysis above. With respect to the design of the UX prototype, the included design categories can be found in Section 3.6.2. The results of this analysis can be found in Table 16, with an exception of UX expert 2. This reviewer did not delete any of the experience categories listed and is thus excluded from the analysis due to not completing the task. The categories included in the UX prototype are underlined in the table. For the nUX prototype, only ‘finishing a task’ was assumed to be included during the design phase. According to the results displayed, it becomes apparent that the experts have found the categories 1, 4 and 8 accordingly to our expectation in the prototype. However, the categories 2, 3, 5 and 7 were found only partly: all of these were only partly found or even also found for the nUX prototype, such as 5. One of the categories was even not found in any of the two prototypes: ‘stimulating experiences’.

Table E. Experience Categories found by the UX Expert per prototype

	UX1		UX3		UX4	
	UX	nUX	UX	nUX	UX	nUX
<u>Receiving feedback</u>	X		X		X	
Giving feedback						
<u>Appreciation</u>	X			X	X	
Rising to a challenge			X		X	X
Being given a challenge	X				X	X
Helping others						
Receiving help	X					
Teaching others						
Solving a problem			X	X	X	X
Experiencing creativity			X	X	X	X
<u>Finishing a task</u>	X	X	X	X	X	X
<u>Keeping track of things</u>			X	X	X	X
<u>Prioritizing</u>	X		X		X	
Connecting with others						
<u>Exchanging ideas</u>	X				X	X
<u>Stimulating experiences</u>						
Creating something together						
Contributing to something greater						
Competition	X		X			
Earnings						
<u>Aesthetics</u>				X	X	X

3.4 Discussion

The main aim of this pre-study was to examine whether the interactions with the two prototypes evoked the intended emotions: a more positive UX in the UX than in the nUX prototype, and a similar estimation for their usability. For this, five users and four UX experts rated the functionality, the appearance and the incorporated experience categories also by means of the UEQ and the meCUE (experts).

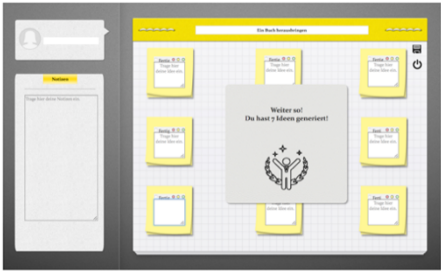
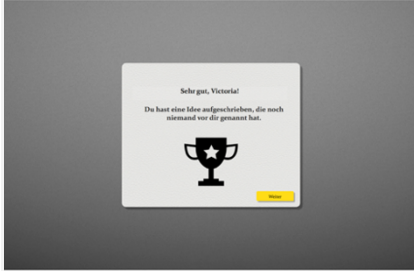


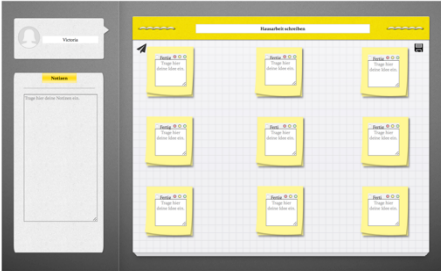

The quantitative analyses of the data revealed contradictory results with respect to the hypotheses. Most of the factors of both questionnaires for both groups implied a higher and thus more positive rating for the nUX prototype than for the UX prototype. This implies that the manipulation of the prototypes has not been successful. Likewise, the qualitative analyses of the data show a similar tendency. The UX prototype revealed most problems during the interaction, while there are only few reported concerning the nUX prototype. This fact, in turn, leads the users and experts to experience more positive emotions with this latter prototype, as fewer negative emotions evolve during the interaction. This explanation is also supported and elaborated on by Hassenzahl (2008): he states that pragmatic qualities “refer [...] to the product’s perceived ability to support the achievement of ‘do-goals’, such as ‘making a telephone call [...]’” (p. 12). Further, he stated that pragmatic qualities refer to the usability and utility of a product which facilitate the fulfillment of be-goals, which are known as hedonic qualities of UX. As mentioned in section 3.1, the hedonic qualities incorporate goals like “‘being competent’, ‘being related to others’ [or] ‘being special’” (Hassenzahl, 2008, p. 12). Additionally, Hassenzahl (2008) claimed that the pragmatic qualities are a pre-requisite or facilitator for hedonic qualities and that a “lack of usability might impose a barrier to the fulfillment of [...] be-goals” (p. 12). Based on this definition, one could argue that the UX prototype received a more negative evaluation, because it showed a greater number of usability problems which left no room for the hedonic qualities to evolve and be fulfilled, thereby preventing the generation of a positive UX. Therefore, it is expected that the adjustment of the recommended design changes by the UX experts as well as the changes retrieved from the user reviews shall lead to a more positive UX in the UX prototype.

Conclusion.

What can be concluded from this pre-study is that with reference to the quantitative and qualitative analyses of the users and experts, it becomes apparent that the manipulation of the two prototypes has not worked as expected: the UX prototype receives generally lower scores on UX than the nUX prototype. Likewise, especially the qualitative analysis revealed that the interactions with the UX prototype incorporated many more problems than the nUX prototype. Therefore, it is expected that an adjustment of the recommended changes will lead to the confirmation of the investigated hypothesis. Accordingly, this hypothesis must be tested again in the main study.

4 Adjustments of UX Prototype

Table A. Most important adjustments of the UX prototype

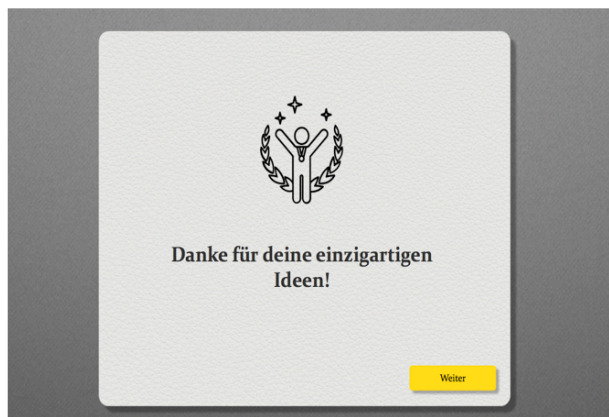
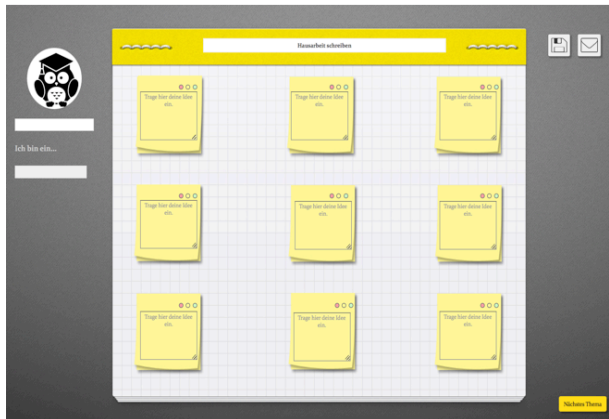
Where/What was the Problem?	How did it look like?	What was the Adjustment	How does it look like now?
<u>Both</u> Too many topics	<u>UX topics:</u> (1) Writing an essay, (2) Preparing to move somewhere, (3) Organizing a journey, (4) Setting up an IKEA shelf, (5) Publish a book <u>nUX topics:</u> (1) Making a tutorial video, (2) Preparing a job interview, (3) Preparing for a pet, (4) Planning a birthday party, (5) Organizing a weekend trip	Reduction from 5 to 2 topics	<u>UX topics:</u> (1) Writing an essay, (2) Preparing to move somewhere <u>nUX topics:</u> (1) Making a tutorial video (2) Preparing a job interview
<u>UX</u> Confusion due to pop-ups		Pop-ups = deleted and placed at end of each topic	
<u>UX</u> Profile picture was unseen		Standard profile picture included	
<u>UX</u> Note area = useless 'Save' and 'quit' were difficult to find/see		Note are = removed; buttons moved + explained in tutorial	

5 Prototype Main Study

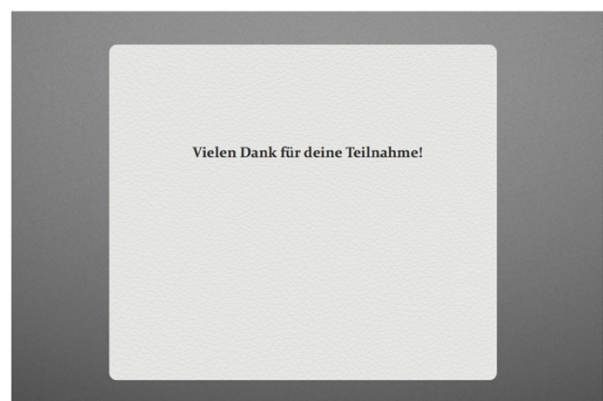
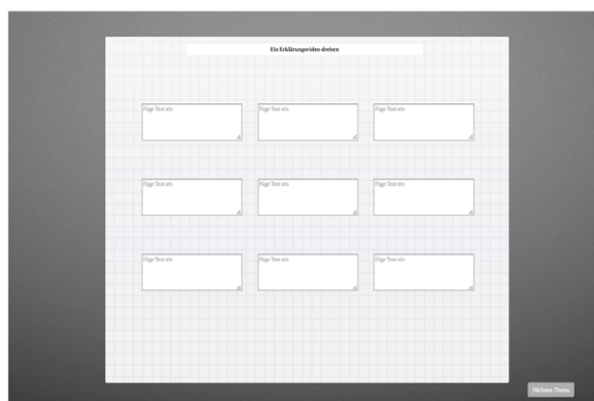
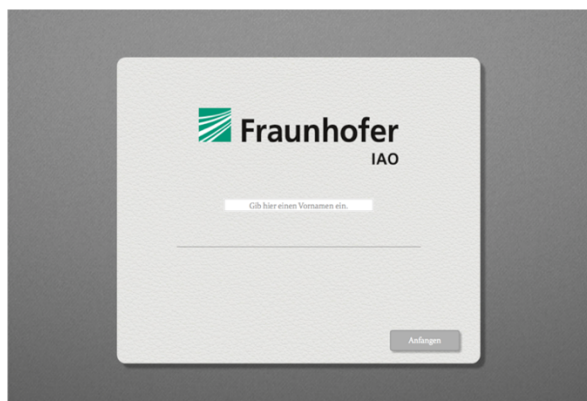
This prototype has been derived from the analyses of the pre-study as well as the feedback given by the participants of the pre-study. The nUX prototype has not been changed. The pre-study version has also been utilized during the main study.

5.1 UX Prototype





5.2 nUX Prototype



5.3 Stimulus Material per Prototype

Table A. Overview of the Screenshots utilized

<u>Picture Content</u>	<u>Prototype</u>	
	<u>nUX</u>	<u>UX</u>
‘Home’ page	X	X
‘Greeting’ page		X
‘Tutorial’ pages		X
‘Topic’ page	X (1-3)	X (1)
‘Interaction’ page	X (1-3)	X (1)
‘Text Area’ close-up	X (Text box)	X (Post-it)
‘Save’ pop-up		X
‘Saved’ pop-up		X
‘Send’ pop-up		X
‘Sent’ pop-up		X
‘Quit’ pop-up		X
‘Appraisal’ page 1		X
‘Appraisal’ page 2		X
‘End page’	X	X

6 Main Study

6.1 Assumptions

6.1.1 Prototype Assumption

As in the pre-study, the UEQ and meCUE are used to shed light on the prototype assumption from the pre-study a second time. Differently to the pre-study, a merged questionnaire resulting from the combination of items from the UEQ and meCUE is used now in order to assess the mean ratings of both prototypes. The merged questionnaire has been established as some of the UX experts noted the inapplicability of some of the items from the meCUE. As the two questionnaires, UEQ and meCUE, both make use of the same 7-point Likert scale response format, the questionnaires could easily be merged together into one questionnaire.

Any 7-point Likert scale incorporates a neutral rating position of ‘4’ at its center. Generally, it is expected that the nUX prototype receives a neutral mean rating, while the UX prototype is assumed to yield a positive mean rating. With respect to this, it is assumed that the mean rating of the nUX prototype centers around ‘4’. With regard to the mean rating of the UX prototype, it is important to take the central tendency bias into account, which, according to Bertram (2007), states that people answering a Likert scale tend to avoid the extremes of both directions. Instead, they make use of the ratings that are positioned around the central position of the Likert scale. Therefore, it is expected that the UX prototype yields a mean rating of ‘5.5’, as this marks the central position between the neutral ‘4’ and the positive extreme ‘7’. With respect to the usability of both prototypes, it is assumed that both prototypes do not differ from each other and achieve both a neutral rating of ‘4’.

Manipulation check – Prototypes. The UX prototype yields an approximate mean rating of 5.5 while the nUX prototype yields an approximate rating of 4. The difference between the two mean ratings is at least 1.5 ratings. Regarding the usability, both prototypes shall not differ from each other and achieve a mean rating of approximately ‘4’.

6.1.2 SAM-ratings

Similarly to the UEQ-meCUE, the SAM makes use of a Likert scale rating format, but instead of a 7-point format, this test utilizes a 9-point rating format. It is expected that participants rate the UX prototype more positively on the ‘valence scale’ than the nUX prototype.

On the basis of the arguments described in section 3.1.1, it is also expected that the nUX prototype receives a mean rating that centers around the neutral position of the Likert scale: ‘5’. For the UX prototype, a higher rating is assumed, a mean rating of ‘7’, as this marks the central position between the neutral rating of ‘5’ and the extreme value of ‘9’.

Manipulation check – SAM. The UX prototype yields an approximate mean rating of 7 while the nUX prototype yields an approximate rating of 5. The difference between the two mean ratings is at least 2 ratings.

6.1.3 AAT

With respect to the AAT, it is expected that participants approach stimuli of the UX prototype and avoid nUX stimuli faster, while they approach nUX stimuli and avoid UX stimuli slower. Therefore, an interaction effect of prototype (UX, nUX) vs. response category (approach, avoid). In order to compare the prototypes, the mean difference scores are calculated by subtracting the means of the Push-condition from the Pull-condition for both prototypes separately. The following is assumed:

Manipulation check – AAT. For the nUX prototype, a positive mean difference score is expected, as this is due to larger RTs on ‘Pull’ than on ‘Push’, while the mean difference score of the UX prototype is assumed to be negative due to smaller RTs on ‘Pull’ than on ‘Push’. According to Chan and Chan (2009), the difference between incompatible and compatible trials is assumed to be approximately 95ms. Therefore, the difference scores for the two prototypes are expected to incorporate a difference of approximately 95ms.

6.1.4 AMP

The studies utilizing the AMP have conducted the test incorporating an extreme dichotomous type of stimulus material: positive vs. negative, or pleasant vs. unpleasant. However, with respect to the generated prototypes and the derived stimulus material, the valence of this study’s material ranges

only from neutral to positive. A neutral valence of a prototype is achieved by rating 50% of the stimulus material for that prototype as 'pleasant' and 50% as 'unpleasant'. Similarly to the procedure followed with the SAM ratings and the UEQ-meCUE, a positive rating is expected to center around 75% 'pleasant' ratings for the stimulus material of a prototype, as this marks the central tendency between 100% and 50% 'pleasant' ratings.

Manipulation check – AMP. It is assumed that the UX prototype yields a proportion of 'pleasant' ratings of .75, while the nUX prototype yields only .50 'pleasant' ratings. The expected mean difference in proportion is thus .25.

6.2 Scenario

Herzlich Willkommen zu unserer Studie und vielen Dank für Ihre Teilnahme.

Stellen Sie sich für die erste Aufgabe die folgende Situation vor:

Für ein großes Projekt wurden Sie und Ihr Team in mehrere Kleingruppen eingeteilt.

Jede Gruppe soll zu verschiedenen Themen Ideen sammeln. Ihre Gruppe besteht aus Ihren zwei Kollegen, Herrn Michael Schmidt (michael.schmidt@fraunhofer.de) und Frau Greta Jakobs (greta.jakobs@fraunhofer.de) und Ihnen. Sie haben sich in der Gruppe entschieden, dass jeder sich zuerst alleine über die Themen Gedanken machen soll. Um möglichst viele Ideen zu erhalten, soll jeder mindestens sechs, maximal aber neun Ideen pro Thema aufschreiben und diese durch Clustering oder Gruppieren sortieren. Zum Abschluss jedes Themas sollen die Ideen gespeichert werden. Ihre Ideen werden danach automatisch mit den Ideen der anderen verglichen. Teilen Sie danach Ihre Ideen mit Ihrer Kollegin, die die Ideen für Ihre gemeinsame Besprechung zusammenträgt. Dann beginnen Sie mit dem nächsten Thema.

Beispiel:

Thema „Ein IKEA Regal aufbauen“

Stellen Sie sich vor, Sie möchten nächste Woche ein neues Regal aufbauen. Was müssen Sie hierfür bedenken?

- ☐ Messen, wie viel Platz für ein Regal zur Verfügung steht
- ☐ Regal kaufen
- ☐ ...

Wir stellen Ihnen für Ihre Überlegungen nacheinander zwei Programme zur Verfügung. Sie können zu jedem Thema aufschreiben, was Ihnen dazu einfällt. Es gibt keine richtigen oder falschen Antworten.

Der Ablauf ist wie folgt:

Nachdem Sie die ersten zwei Themen abgeschlossen haben, bekommen Sie einen Fragebogen zum Ausfüllen. Danach gibt es eine Pause von ca. 2 Minuten, in der Sie sich entspannen können. Nach der Pause werden Sie die nächsten zwei Themen bearbeiten. Hiernach gibt es ebenfalls einen Fragebogen und eine Pause.

Im Anschluss an die Pause beginnen Sie mit dem ersten der drei Tests. Für alle drei Tests gibt es eine separate Einleitung vom Versuchsleiter.

6.3 R – Code

6.3.1 Analyses

Importantly, for the creation of the data sets, two dummy variables are created in order to account for the categorical variables of the data: ‘prototype’ and ‘group’. For both variables, a value of ‘0’ was chosen for nUX, and ‘1’ for UX.

#Import Data

```
setwd("~/Documents")
myData <- read.table("AnalysisData.csv", header = T, sep=";")

#LM Model AAT + Coefficients
AAT.model= lmer (AATDiff ~ Prototype + Group + (1|Participant), data = myData)
summary(AAT.model)
coef(AAT.model)
confint(AAT.model, level = 0.9, data = myData)
mean (myData$AATDiff[myData$Prototype ==0],na.rm=TRUE)
mean (myData$AATDiff[myData$Prototype ==1],na.rm=TRUE)
sd (myData$AATDiff[myData$Prototype ==0],na.rm=TRUE)
sd (myData$AATDiff[myData$Prototype ==1],na.rm=TRUE)
```

#LM Model SAM + Coefficients

```
SAM.model= lmer (SAM ~ Prototype + Group + (1|Participant), data = myData)
summary(SAM.model)
coef(SAM.model)
confint(SAM.model, level = 0.9, data = myData)
mean (myData$SAM[myData$Prototype ==0],na.rm=TRUE)
mean (myData$SAM[myData$Prototype ==1],na.rm=TRUE)
sd (myData$SAM[myData$Prototype ==0],na.rm=TRUE)
sd (myData$SAM[myData$Prototype ==1],na.rm=TRUE)
```

#LM Model AMP + Coefficients

```
AMP.model= lmer (AMP ~ Prototype + Group + (1|Participant), data = myData)
summary(AMP.model)
coef(AMP.model)
confint(AMP.model, level = 0.9, data = myData)
mean (myData$AMP[myData$Prototype ==0],na.rm=TRUE)
mean (myData$AMP[myData$Prototype ==1],na.rm=TRUE)
sd (myData$AMP[myData$Prototype ==0],na.rm=TRUE)
sd (myData$AMP[myData$Prototype ==1],na.rm=TRUE)
```

#LM Model Hedonic Qualities + Coefficients

```
Hedonic.model= lmer (Hedonic.Qualities ~ Prototype + Group + (1|Participant), data
= myData)
summary(Hedonic.model)
coef(Hedonic.model)
confint (Hedonic.model, level=0.9, data = myData)
mean (myData$Hedonic[myData$Prototype ==0],na.rm=TRUE)
mean (myData$Hedonic[myData$Prototype ==1],na.rm=TRUE)
sd (myData$Hedonic[myData$Prototype ==0],na.rm=TRUE)
sd (myData$Hedonic[myData$Prototype ==1],na.rm=TRUE)
```

#LM Model Pragmatic Qualities + Coefficients

```
Pragmatic.model= lmer (Pragmatic.Qualities ~ Prototype + (1|Participant) + Group,
data = myData)
summary(Pragmatic.model)
coef(Pragmatic.model)
confint (Pragmatic.model, level=0.9, data = myData)
mean (myData$Pragmatic[myData$Prototype ==0],na.rm=TRUE)
mean (myData$Pragmatic[myData$Prototype ==1],na.rm=TRUE)
sd (myData$Pragmatic[myData$Prototype ==0],na.rm=TRUE)
sd (myData$Pragmatic[myData$Prototype ==1],na.rm=TRUE)
```

6.3.2 Plots

#Import data

```
setwd("~/Documents")
myData <- read.table("PlotData.csv", header = T, sep=";")
```

#Hedonic Qualities

```
#Distribution of answers per prototype - Overlay plots
a <- hist(myData$Hedonic.Qualities[myData$Prototype==0], xlim=c(1,7), ylim=c(0,15),
xlab="Rating", col = rgb(1,0.6,0,alpha=0.7))
b <- hist(myData$Hedonic.Qualities[myData$Prototype==1], xlim=c(1,7), ylim=c(0,15),
xlab="Rating", col = rgb(0,0,1,alpha=0.5), add=T)
labels <- c("UX Prototype", "nUX Prototype")
legend ("topleft", inset =.05, title = "Legend", labels, lwd = 2, col = c("orange",
"blue"))
```

#Boxplot per prototype

```
c <- boxplot(myData$Hedonic.Qualities~myData$Prototype, xlab="Prototype",
ylim=c(2,7), ylab="Mean Ratings", names = c("UX", "nUX"), col = c("orange",
"blue"))
```

#Boxplot per prototype & group

```
d <- boxplot(myData$Hedonic.Qualities~myData$Group*myData$Prototype,
xlab="Prototype - Group", ylim=c(2,7), ylab="Mean Ratings", names = c("UX-UX", "UX-
nUX", "nUX-UX", "nUX-nUX"), col = c("orange", "orange", "blue", "blue"))
```

#Spaghetti Plot (SPSS)

```
* Chart Builder.
```

```
GGRAPH
```

```
  /GRAPHDATASET NAME="graphdataset" VARIABLES=Prototyp
```

```
MEAN(Hedonic_Qualities)[name="MEAN_Hedonic_Qualities"] Participant MISSING=LISTWISE
REPORTMISSING=NO
```

```
  /GRAPHSPEC SOURCE=INLINE.
```

```
BEGIN GPL
```

```
  SOURCE: s=userSource(id("graphdataset"))
```

```

DATA: Prototyp=col(source(s), name("Prototyp"), unit.category())
DATA: MEAN_Hedonic_Qualities=col(source(s), name("MEAN_Hedonic_Qualities"))
DATA: Participant=col(source(s), name("Participant"), unit.category())
GUIDE: axis(dim(1), label("Prototyp"))
GUIDE: axis(dim(2), label("Mean Hedonic_Qualities"))
GUIDE: legend(aesthetic(aesthetic.color.interior), label("Participant"))
SCALE: cat(dim(1), include("1,00", "2,00"))
SCALE: linear(dim(2), include(0))
ELEMENT: line(position(Prototyp*MEAN_Hedonic_Qualities),
color.interior(Participant), missing.wings())
END GPL.

```

Pragmatic Qualities

```

#Distribution of answers per prototype - Overlay plots
a <- hist(myData$Pragmatic.Qualities[myData$Prototyp==0],
xlim=c(1,7),ylim=c(0,15), xlab="Rating", col = rgb(1,0.6,0,alpha=0.7))
b <- hist(myData$Pragmatic.Qualities[myData$Prototyp==1],
xlim=c(1,7),ylim=c(0,15), xlab="Rating", col = rgb(0,0,1,alpha=0.5), add=T)
labels <- c("UX Prototype", "nUX Prototype")
legend("topleft", inset=.05, title="Legend", labels, lwd=2, col=c("orange",
"blue"))

```

```

#Boxplot per prototype
c <- boxplot(myData$Pragmatic.Qualities~myData$Prototyp,
xlab="Prototyp",ylim=c(2,7), ylab="Mean Ratings", names=c("UX", "nUX"), col =
c("orange", "blue"))

```

```

#Boxplot per prototype & group
d <- boxplot(myData$Pragmatic.Qualities~myData$Group*myData$Prototyp,
xlab="Prototyp - Group", ylim=c(2,7), ylab="Mean Ratings", names=c("UX-UX", "UX-
nUX", "nUX-UX", "nUX-nUX"), col=c("orange", "orange", "blue", "blue"))

```

#Spaghetti Plot (SPSS)

* Chart Builder.

GGRAPH

```

/GRAPHDATASET NAME="graphdataset" VARIABLES=Prototyp
MEAN(Pragmatic_Qualities)[name="MEAN_Pragmatic_Qualities"] Participant
MISSING=LISTWISE REPORTMISSING=NO
/GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
SOURCE: s=userSource(id("graphdataset"))
DATA: Prototyp=col(source(s), name("Prototyp"), unit.category())
DATA: MEAN_Pragmatic_Qualities=col(source(s), name("MEAN_Pragmatic_Qualities"))
DATA: Participant=col(source(s), name("Participant"), unit.category())
GUIDE: axis(dim(1), label("Prototyp"))
GUIDE: axis(dim(2), label("Mean Pragmatic_Qualities"))
GUIDE: legend(aesthetic(aesthetic.color.interior), label("Participant"))
SCALE: cat(dim(1), include("1,00", "2,00"))
SCALE: linear(dim(2), include(0))
ELEMENT: line(position(Prototyp*MEAN_Pragmatic_Qualities),
color.interior(Participant), missing.wings())
END GPL.

```

SAM

```

#Distribution of answers per prototype - Overlay plots
a <- hist(myData$SAM[myData$Prototyp==0], xlim=c(0,10), ylim=c(0,15), xlab="Mean
Rating", col = rgb(1,0.6,0,alpha=0.7))
b <- hist(myData$SAM[myData$Prototyp==1], xlim=c(0,10), ylim=c(0,15),
xlab="Rating", col = rgb(0,0,1,alpha=0.5), add=T)
labels <- c("UX Prototype", "nUX Prototype")

```

```

legend ("topleft", inset =.05, title = "Legend", labels, lwd = 2, col = c("orange",
"blue"))

#Boxplot per prototype
c <- boxplot(myData$SAM~myData$Prototype, xlab="Prototype",ylim=c(2,8), ylab="Mean
Ratings", names = c("UX", "nUX"), col = c("orange", "blue"))

#Boxplot per prototype & group
d <- boxplot(myData$SAM~myData$Group*myData$Prototype, xlab="Prototype - Group",
ylim=c(2,8), ylab="Mean Ratings", names = c("UX-UX", "UX-nUX", "nUX-UX", "nUX-
nUX"), col = c("orange", "orange", "blue", "blue"))

#Spaghetti Plot (SPSS)
* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=Prototype
MEAN(SAM_rating)[name="MEAN_SAM_rating"] Participant MISSING=LISTWISE
REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: Prototype=col(source(s), name("Prototype"), unit.category())
  DATA: MEAN_SAM_rating=col(source(s), name("MEAN_SAM_rating"))
  DATA: Participant=col(source(s), name("Participant"), unit.category())
  GUIDE: axis(dim(1), label("Prototype"))
  GUIDE: axis(dim(2), label("Mean SAM_rating"))
  GUIDE: legend(aesthetic(aesthetic.color.interior), label("Participant"))
  SCALE: cat(dim(1), include(",00", "1,00"))
  SCALE: linear(dim(2), include(0))
  ELEMENT: line(position(Prototype*MEAN_SAM_rating), color.interior(Participant),
missing.wings())

```

AAT

```

#Distribution of answers per prototype - Overlay plots
a <- hist(myData$AATDiff[myData$Prototype==0], xlim=c(-300,300), ylim=c(0,15),
xlab="Difference Score in ms", col = rgb(1,0.6,0,alpha=0.7))
b <- hist(myData$AATDiff[myData$Prototype==1], xlim=c(-300,300), ylim=c(0,15),
xlab="Rating", col = rgb(0,0,1,alpha=0.5), add=T)
labels <- c("UX Prototype", "nUX Prototype")
legend ("topleft", inset =.05, title = "Legend", labels, lwd = 2, col = c("orange",
"blue"))

#Boxplot per prototype
c <- boxplot(myData$AATDiff~myData$Prototype, xlab="Prototype", ylab="Difference
Score", names = c("UX", "nUX"), col = c("orange", "blue"))

#Boxplot per prototype & group
d <- boxplot(myData$AATDiff~myData$Group*myData$Prototype, xlab="Prototype-Group",
ylab="Difference Score", names = c("UX-UX", "UX-nUX", "nUX-UX", "nUX-nUX"), col =
c("orange", "orange", "blue", "blue"))

#Spaghetti Plot (SPSS)
* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=Prototype
MEAN(Difference_Score_AAT)[name="MEAN_Difference_Score_AAT"] Participant
MISSING=LISTWISE REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: Prototype=col(source(s), name("Prototype"), unit.category())
  DATA: MEAN_Difference_Score_AAT=col(source(s), name("MEAN_Difference_Score_AAT"))
  DATA: Participant=col(source(s), name("Participant"), unit.category())

```



```

GUIDE: axis(dim(1), label("Prototype"))
GUIDE: axis(dim(2), label("Mean Difference_Score_AAT"))
GUIDE: legend(aesthetic(aesthetic.color.interior), label("Participant"))
SCALE: cat(dim(1), include(".", "00", "1,00"))
SCALE: linear(dim(2), include(0))
ELEMENT: line(position(Prototype*MEAN_Difference_Score_AAT),
color.interior(Participant), missing.wings())
END GPL.

```

AMP

```

#Distribution of answers per prototype - Overlay plots
a <- hist(myData$AMP[myData$Prototype==0], xlim=c(0,1), ylim=c(0,8),
xlab="Proportion of Positive Ratings", col = rgb(1,0.6,0,alpha=0.7))
b <- hist(myData$AMP[myData$Prototype==1], xlim=c(0,1), ylim=c(0,8), xlab="
Proportion of Positive Ratings ", col = rgb(0,0,1,alpha=0.5), add=T)
labels <- c("UX Prototype", "nUX Prototype")
legend ("topleft", inset =.05, title = "Legend", labels, lwd = 2, col = c("orange",
"blue"))

#Boxplot per prototype
c <- boxplot(myData$AMP~myData$Prototype, ylim=c(0,1), ylab="Proportion of Positive
Ratings", names = c("UX", "nUX"), col = c("orange", "blue"))

#Boxplot per prototype & group
d <- boxplot(myData$AMP~myData$Group*myData$Prototype, ylim=c(0,1),
xlab="Prototype-Group", ylab="Mean Proportion", names = c("UX-UX", "UX-nUX", "nUX-
UX", "nUX-nUX"), col = c("orange", "orange", "blue", "blue"))

#Spaghetti Plot (SPSS)
* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=Prototype
MEAN(AMP_positive)[name="MEAN_AMP_positive"] Participant MISSING=LISTWISE
REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: Prototype=col(source(s), name("Prototype"), unit.category())
  DATA: MEAN_AMP_positive=col(source(s), name("MEAN_AMP_positive"))
  DATA: Participant=col(source(s), name("Participant"), unit.category())
  GUIDE: axis(dim(1), label("Prototype"))
  GUIDE: axis(dim(2), label("Mean AMP_positive"))
  GUIDE: legend(aesthetic(aesthetic.color.interior), label("Participant"))
  SCALE: cat(dim(1), include(".", "00", "1.00"))
  SCALE: linear(dim(2), include(0))
  ELEMENT: line(position(Prototype*MEAN_AMP_positive), color.interior(Participant),
missing.wings())
END GPL.

```

7 Results of the Main Study

7.1 Validation of merged UEQ-meCUE questionnaire

For the analysis of the UEQ and meCUE, an exploratory factor analysis has been conducted in order to obtain clusters of variables among the two merged questionnaires. The principal component analysis was performed with a “Direct Oblimin rotation” and a “Kaiser Normalization”. Eigenvalues equal to or greater than 1 were extracted and yielded six factors accounting for the 32 variables of both

questionnaires. The six factors accounted for 72.65% of the total variance, while each factor explained 40.24%, 15.57%, 4.94%, 4.35%, 4.00% and 3.55%, respectively, of the total variance.

Table A. Assignment of Items to the Six Factors Resulting from the Exploratory Factor Analysis

Items	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
<u>meCUE items</u>						
Exhilarates	.806	-.017	-.105	-.094	-.068	-.288
Annoys	.598	.457	.084	.051	-.190	-.131
Frustrates	.547	.317	-.081	.048	-.294	.095
Euphoric	.826	-.032	-.086	-.215	.107	-.036
Cheerful	.802	-.192	-.094	-.342	-.058	-.046
Angers	.629	.343	.094	.176	-.465	.201
<u>UEQ items</u>						
Activitiy1	.826	.047	-.116	-.259	.097	.115
Perspicuity1	.180	.539	.410	-.019	.112	.122
Novelty1	.621	-.366	.263	.253	-.033	.292
Perspicuity2	-.016	.480	.320	.033	.383	.153
Stimulation1	.782	-.027	.210	.073	.142	.179
Stimulation2	.834	-.363	.066	.080	.024	.020
Stimulation3	.858	-.261	-.009	.041	.056	.070
Dependability1	.127	.545	-.584	.157	-.030	.175
Efficiency1	.080	.692	-.066	.215	-.198	-.092
Novelty2	.702	-.552	.178	.159	.013	-.011
Dependability2	.717	.383	.006	-.045	.134	-.015
Attractivity2	.900	.125	.038	.015	.034	.061
Perspicuity3	-.157	.599	.260	-.338	-.478	.027
Attractivity3	.770	.260	-.087	.009	.006	-.149
Novelty3	.668	-.545	.142	.273	.124	-.032
Attractivity4	.788	.352	-.207	.039	-.032	-.160
Dependability3	.390	.438	.044	.622	-.122	.048
Stimulation4	.846	-.250	.050	.024	-.041	.076
Dependability4	.375	.316	-.303	-.152	.258	.604
Efficiency2	.495	.332	-.086	.281	.210	-.435
Perspicuity4	.295	.557	.276	-.086	.247	.115
Efficiency3	.214	.490	-.150	.035	.445	-.286
Efficiency4	-.047	.552	.556	-.276	.056	-.172
Attractivity5	.847	-.155	-.041	-.269	-.022	-.045
Attractivity6	.832	-.030	-.174	-.318	-.133	-.048
Novelty4	.621	-.474	.260	.017	-.132	-.155

However, an analysis of the component matrix revealed, displayed in Table 8, that most variables loaded on the first and second factor. Only two variables, ‘dependability4’ and ‘efficiency4’ loaded on the factors 6 and 3, respectively. However, ‘efficiency4’ loaded .556 on factor 3 and .552 on factor 2, and can therefore also be assigned to factor 2. Though, ‘dependability4’ loads high on factor 6 (.604), it has also a loading of .375 on factor 1, which implies to attribute this item to factor 1. The resulting two factors were named accordingly to the contextual groupings of items: ‘hedonic qualities’ and

‘pragmatic qualities’. The assignment of the 32 items to the two factors, hedonic and pragmatic qualities, can be seen in Table 7.A.

7.2 Exploratory Data Analysis

For the analysis of the subsequent four individual tests, an analysis structure was defined, as follows: first, an exploratory data analysis (EDA) was conducted in order to examine the data visually. Herein, the distributions of the mean ratings, difference scores or proportions of positive scores was analyzed, the data was scanned for outliers and potential order effects. Secondly, the statistical analyses followed in order to examine the differences in means across the prototypes statistically. As all tests of this study incorporate a repeated measures design on the factor ‘prototype’, one of the basic assumptions of the standard repeated measures Analysis of Variance (ANOVA) is violated: for an ANOVA, it has to be ensured that the variables are independent and identically distributed. However, a repeated measures design indicates that multiple data points are retrieved from the same subject, which, therefore, advocates against this assumption. A linear mixed effects (LME) model accounts for this violation of data independence by adding random factors to the models. Thus, for the analyses of the tests, four LME models were chosen and constructed with two fixed effects, ‘prototype’ and ‘ordering’ as well as one random effect, ‘participant’, as each participant gave two ratings, one for UX and one for nUX. The model is as follows: **Dependent Variable ~ Prototype + Ordering + (1|Participant)**

7.2.1 UEQ-meCUE

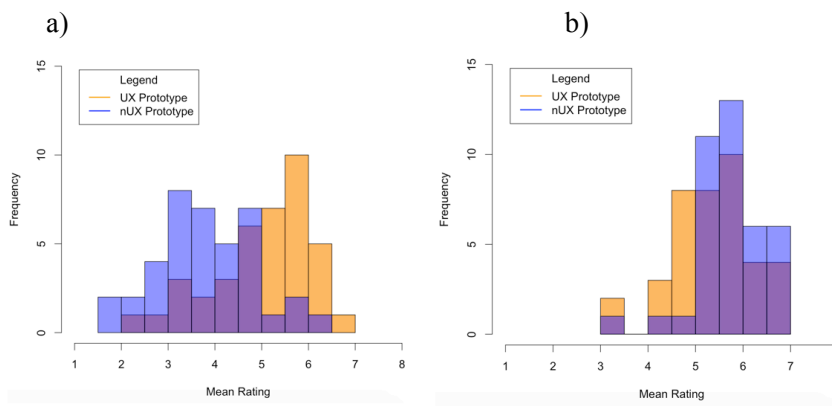


Figure A: Histogram plot displaying mean ratings of the two factors 'hedonic qualities' and 'pragmatic qualities'

Figure A shows two histogram plots displaying the frequencies of mean ratings for each of the two factors, ‘hedonic qualities’ and ‘pragmatic qualities’ retrieved during the exploratory factor analysis of the UEQ-meCUE for $n = 39$. With respect to plot (a), it becomes apparent that the mean ratings for the UX prototype on the factor ‘hedonic qualities’ are on average higher than the mean ratings for the nUX prototype. The former range from ‘2’ to ‘7’, while the highest frequency is reported for a mean rating of ‘5.5’. The latter range from ‘1.5’ to ‘6.5’ and highest frequencies are reported for a mean

rating of '3'. The plot also shows the overlap between the ratings for the two prototypes, which is displayed in dark violet. In plot (b), a substantial overlap between the ratings of the two prototypes can be seen, which is in accordance with the expected assumption: the prototypes do not differ in the estimation of their usability. The mean ratings of both range from '3' to '7' with highest frequencies around '5.5', however, the nUX prototype received higher frequencies for mean ratings above '5' than the UX prototype. In order to investigate this phenomenon, a boxplot diagram was constructed for the two factors, which is displayed in Figure B.

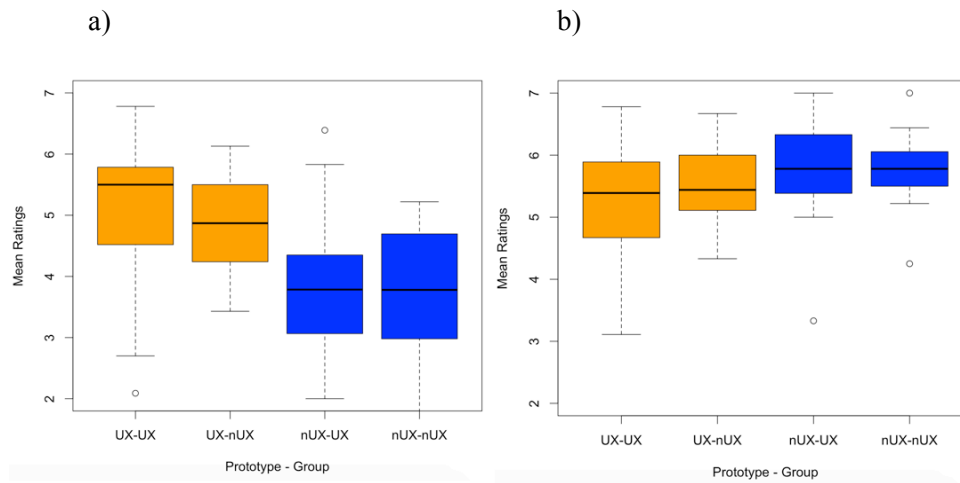


Figure B: Boxplot diagram displaying mean ratings of the two factors 'hedonic qualities' and 'pragmatic qualities' per prototype and group

The boxplot reveals that the estimation of the participant's mean ratings per prototype are reversed across the two factors: the UX prototype reaches apparently higher mean ratings than the nUX prototype on the factor 'hedonic qualities'. However, on the factor 'pragmatic qualities', the nUX prototype reaches slightly higher mean ratings. Furthermore, in order to control for a potential effect of the order, in which the participants received the prototypes, the UX- and nUX-boxplots were displayed for each of the two conditions: UX-first and nUX-first. No order effect of the prototypes were found on the two factors, however, the factor 'hedonic qualities' shows that the UX prototype was rated slightly better in the UX-first group than in the nUX-first group.

Additionally, the plots also show outliers: plot (a) reveals two outliers for the UX-first group. One rating for the UX prototype is unexpectedly low, while another for the nUX prototype unexpectedly high. However, these outliers were not excluded from the analysis, as they had normal interactions with the prototypes. Plot (b) displays three outliers for the mean ratings of the nUX prototypes. However, these, as well as the outliers described on the factor 'hedonic qualities', were not excluded from the analysis, as they were still within the range of inclusion described above.

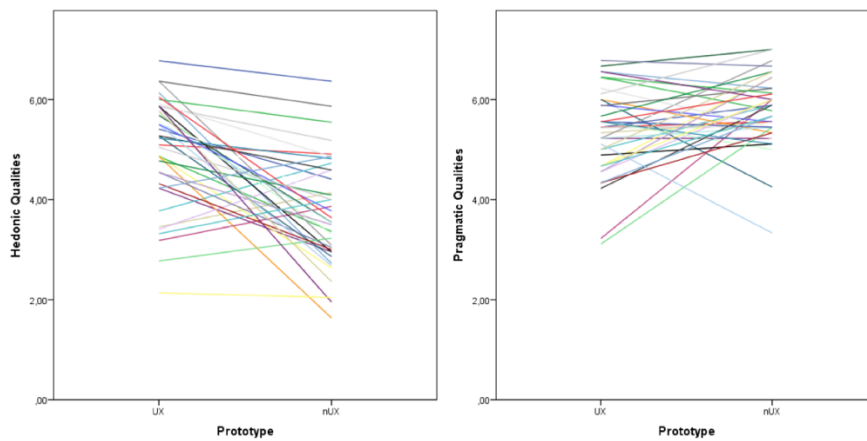


Figure C: Spaghetti plot displaying the distribution of ratings of the two factors 'hedonic qualities' and 'pragmatic qualities' per participant and prototype

The spaghetti plot, displayed in Figure C, confirms the results of the histograms and boxplots above: For the 'hedonic qualities', there is a general tendency to rate the UX prototype higher than the nUX prototype and only few seem to contradict this tendency. However, interestingly, many of the ratings appear to follow a parallel trend, implying a participant intercept random effect that should be investigated during the statistical analysis. For the factor 'pragmatic qualities', there is an overall tendency to rate the nUX prototype slightly higher, while there are also few that contradict this tendency.

7.2.2 SAM

The data included $n = 1297$ retrieved answers on the SAM and 9055 missing values from $n = 25$ participants. There was a slight imbalance of the answers retrieved per prototype due to the large number of missing values: $n = 603$ answers were obtained for the UX prototype as well as $n = 694$ for the nUX prototype. From these answers, the mean ratings per prototype of each participant and the corresponding standard deviations were calculated. From this data set, the mean ratings and standard deviations per prototype were calculated by averaging across all participants. Figure D displays the distribution of mean ratings per prototype. With respect to this figure, it becomes visible that the UX prototype's mean ratings range from ratings around '3' to a maximum rating of '8', while those of the nUX prototype range from '2' to a maximum of '6'. The UX prototype shows the highest frequencies of mean ratings around a value '6' with a frequency of approximately 11 times, while nUX prototype shows highest frequencies around a value of '4' with a frequency of approximately 9 times. In order to inspect the

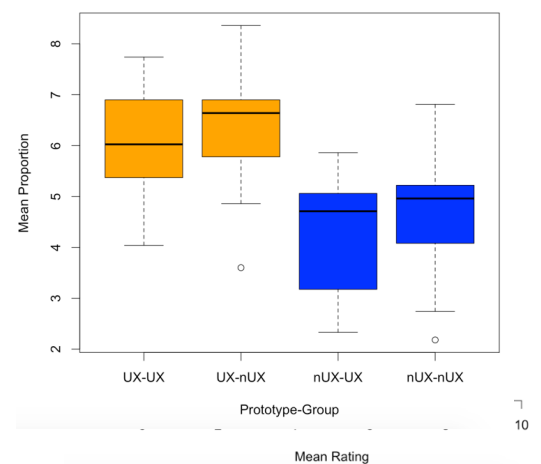


Figure D: Distribution of the mean ratings on the SAM valence scale

outliers of the SAM as well as a potential effect of the prototype order, a boxplot diagram, displayed in Figure 14, has been established. The plot shows a major difference between the mean ratings of the two prototype: the mean rating for the UX prototype revolves around a rating of '6 - 6.5', while that of the nUX prototype centers around '4.5 - 5'. The variances of the boxplots are approximately symmetrically distributed towards the lowest and highest ratings and match each other in their size. However, it becomes apparent that the nUX-first group had the tendency to rate the both prototypes slightly better than the UX-first group.

Furthermore, the plot also displays two outliers with respect to the nUX-group, one for each prototype. Both times, the prototypes were rated remarkably low compared to the rest of the ratings. Additionally, a Spaghetti plot been established (see Figure 8, section 3.3.1) in order to investigate the random effects of the SAM.

7.2.3 AAT

Before the initial analysis of the AAT reaction times (RT), the descriptive statistics and frequencies per movement and prototype were analyzed. Trials incorporating wrong answers or changes in the direction of the movement were displayed by means of a filter in SPSS. Likewise, trials with $200\text{ms} < \text{RT} < 2 \cdot \text{SD}$ were also marked with that filter and later excluded from the analysis. Additionally, the total error percentage per participant was calculated. Any participant yielding an error percentage of more than 20% were excluded from the analysis, which resulted in the exclusion of participant 9, 35 and 40 as they had error percentages of 37.5%, 23.5% and 21.3%, respectively. The exclusion of these resulted in $n = 36$ participants and led to the descriptives displayed in Table 7.B.

Table B. Descriptives of the AAT

Condition	N		Mean		SD	
	UX	nUX	UX	nUX	UX	nUX
Pull	1029	1200	979.83	967.87	285.23	274.79
Push	1086	1178	970.65	957.83	283.48	283.28
Missing values	4640					
Total	9133					

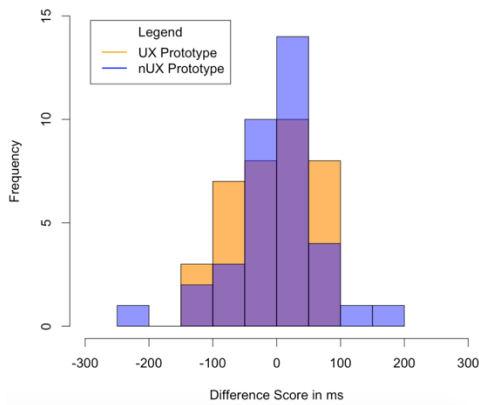


Figure E: Distribution of difference scores per prototype

was conducted, displayed below.

It becomes apparent that the highest frequencies for both prototypes are reported for difference scores that center around zero. Figure E also shows that the difference scores of the UX prototype are more distributed than those of the nUX prototype: the UX prototype also incorporates difference scores of about -250ms as well as scores of about +150ms. In contrast to this, the distribution of the nUX difference scores ranges only from -150ms to 100ms. However, the expected pattern of the UX prototype receiving merely negative difference scores and the nUX prototype to receiving merely positive difference scores is not found here. In order to investigate the differences in scores between the prototypes, outliers as well as potential order effects, a boxplot was established (see Figure F).

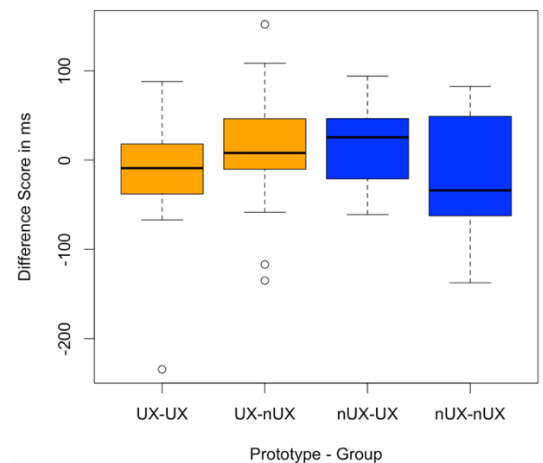


Figure F: Boxplot on the difference scores per prototype and group

With respect to Figure F, it becomes apparent that the

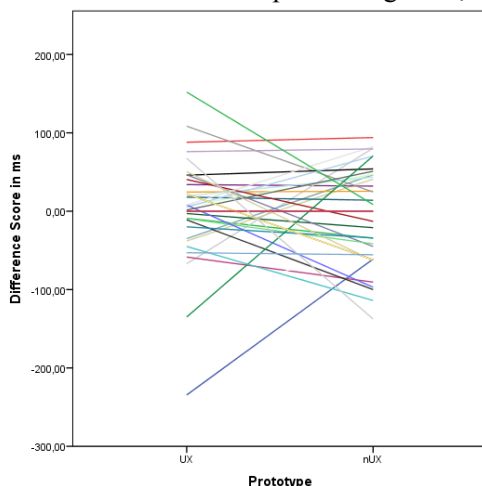


Figure G: Spaghettiplot on the difference scores per participant and prototype

difference scores are, indeed, influenced by the order of the prototypes. The UX-first group shows results that are in accordance with the hypothesis: For the UX prototype the mean difference score is approximately -25ms, caused by greater RTs in congruent than in the incongruent condition. The mean difference score for the nUX prototype of this group is approximately +20ms, due to greater RTs in the pull than the push condition. However, the scores of the nUX-first group, show the exact opposite pattern: their mean difference score for the UX prototype is approximately +15ms, while the score for the nUX prototype is approximately -15ms. These values imply that this group showed an approach tendency

towards the screenshots from the nUX prototype and an avoidance tendency towards the UX prototype pictures, which contradicts the established hypothesis.

An investigation on the ‘participant’ level by means of a spaghetti plot, displayed in Figure G, shows this exact pattern: the participants’ responses are contradicting each other, with some scoring positively on the UX prototype and some scoring negative on this prototype and same on the nUX prototype.

7.2.4 AMP

For the analysis of the AMP, the proportion of positive ratings per participant ($n = 14$) was calculated. By this, participant 34 was excluded from the analysis as this participant utilized the answer

‘unpleasant’ for both prototypes in ‘Run 2’, thus achieving 100% unpleasant ratings for both prototypes in this run. These results are assumed to be due to systematic behavior of the participant, which advocates for the exclusion of the data in order to avoid biasing the results. For the visual inspection, a histogram has been established. Figure H displays a large overlap of the distributions of the positive proportions of the two prototype overlap largely: Both distributions begin at a positive proportion of .3 and range to .7 and .8, for the nUX and UX prototype, respectively. Highest frequencies for the UX prototypes’ positive proportions center around values of .3 and .6, while highest frequencies for nUX center around .4 and .65. This plot is not

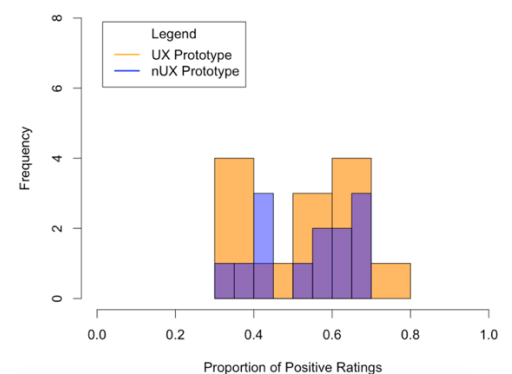


Figure H: Histogram plot on the distribution of positive proportions per prototype

sufficient to draw a clear conclusion on how the proportions differ for the two prototypes. In order to

investigate this question, a boxplot was generated for the positive response rates for the two prototypes (see Figure H).

The boxplot, (Figure I), displays an interesting phenomenon: The boxplots of the UX-first group for both prototypes appear to have no variances and the lowest and highest proportions are incorporated within the interquartile distance. This can be explained, by the fact that this group has too few observations for the proper construction of the boxplots. This was due to the large amounts of missing values in this test condition. However, the medians of these

Figure I: Boxplot on the distribution of positive proportions per prototype and group

two boxplots, indeed, show the expected tendency of the UX prototype to receive higher proportions than the nUX prototype. In contrast to this, the two boxplots of the nUX-group contain large interquartile distances that overlap to great extent. Similarly, the medians also overlap at an approximately proportion of ‘.55’ and do not show the expected tendency, described above. Instead,

the proportion of both prototypes is only slightly above the level that both types of stimuli received the same amounts of positive ratings. chance level of receiving either of the responses, pleasant or unpleasant.

An investigation of the ratings per participant, displayed in the Spaghetti-plot in Figure J, showed that there were only few participants that represented the assumed tendency of higher positive proportions for the UX than the nUX prototype. With respect to the retrieved results, it can be advocated that the manipulation check of the AMP was not successful, as no significant difference between the mean proportion of positive ratings on two prototypes was found.

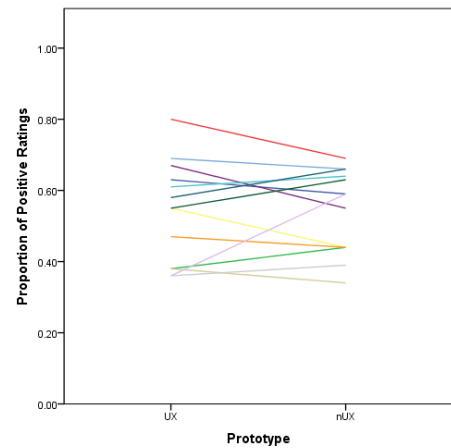


Figure J: Spaghettiplot on the proportion of positive ratings per participant and prototype

8 References – Pictures

8.1 Owls

Owl end page: Macarranza (2016). “Owl”. Retrieved from: <https://thenounproject.com/carranzamaria/collection/owl/?oq=owl&cid=4&i=147404>

Pre-Study UX Owl greeting page (Appendix 8.1): Mcarranza (2016). “Owls”. Retrieved from: <https://thenounproject.com/carranzamaria/collection/owl/?oq=owl&cid=4&i=147406>

Main Study – UX Owl (Interaction Page): Parkijsun (2016). “Academic”. Retrieved from: <https://thenounproject.com/search/?q=owl+academic&i=301757>

8.2 Figure 5&6 – Keyboard

Retrieved from: http://javitas.info/image.php?pic=https://upload.wikimedia.org/wikipedia/commons/thumb/e/e4/Keyboard-icon_Wikipedians.svg/2000px-Keybaord-icon_Wikipedians.svg.png

8.3 Figure 4 – Joystick

Conrad (2016). “Flugsimulator-Joystick Thrustmaster T-Flight Stick X USB PC, PlayStation® 3 Schwarz”. Retrieved from: <https://www.conrad.de/de/flugsimulator-joystick-thrustmaster-t-flight-stick-x-usb-pc-playstation-3-schwarz-906477.html>

8.4 Figure 6 – SAM (Valence/Arousal)

Valence: Irtel, H. (2007). PXLab: The Psychological Experiments Laboratory [Online]. Version 2.1.11. Mannheim: University of Mannheim. Available at: <http://www.pxlab.de> [Accessed: 11 March 2009].

Arousal: AdSAM®'s Empirical Foundations (2016). Retrived from: <http://adsam.com/adsam-empirical-foundations.php>

8.5 UX Prototype – Pre-Study (Appendix 1)

8.5.1 Quit + Save (Interaction Page)

Quit: Amos (2016). “Power”. Retrieved from:
<https://thenounproject.com/search/?q=on+off&i=9957>

Save: Draiman, H. (2016). “Save”. Retrieved from:
<https://thenounproject.com/search/?q=save&i=599630>

8.5.2 Appraisals (Appraisal Page)

Lacke, N. (2016). “Trophy”. Retrieved from:
<https://thenounproject.com/search/?q=winner&i=63988>

Parkijsun (2016). “Winner”. Retrieved from:
<https://thenounproject.com/search/?q=winner%20manikin&i=492250>

Prajapati, D. (2016). “Paper Plane”. Retrieved from:
<https://thenounproject.com/search/?q=paper+plane&i=187149>