User experience over time –
A longitudinal study with iPhone users

Master's Thesis of:

David Tschapnizky, B.Sc.
d.tschapnizky@stud.unibas.ch
Faculty of Psychology
Department of Cognitive Psychology and Methodology
University of Basel

November 2011

First Reviewer: Silvia Heinz, M.Sc.
Department of Cognitive Psychology and Methodology, University of Basel

Second Reviewer: Prof. Dr. Klaus Opwis
Department of Cognitive Psychology and Methodology, University of Basel

Abstract

Human-Computer Interaction (HCI) recently focused more on a holistic perspective of user experience, which goes beyond merely instrumental needs to understand what is relevant for the evaluation of interactive products. Eight first time iPhone users were investigated over 8 weeks after the purchase using an online survey to analyse how peoples' experiences and evaluations of interactive products develop. We could show that the evaluated goodness of an interactive product is related to pragmatic as well as to hedonic aspects, whereas beauty was only related to hedonic aspects. The study findings underline the importance of hedonic aspects of interactive products, which seem to play a key role for outstanding experiences.

User experience over time –

A longitudinal study with iPhone users

Improving usability was the main factor for the commercial success of software in the mid 90s (Butler, 1996). Usability refers to the effectiveness, efficiency, and satisfaction, allowing users to achieve defined goals in a specified context (ISO 9241, 1998). Then as now, the usability of software (Karat 1992; Kjeldskov, Skov, & Fiegel 2008) and websites (Hornbbæk, 2006) was evaluated on the basis of how well a user could achieve defined tasks (i.e. fill a table with numbers). Thereafter HCI also acknowledged self related aspects, such as pleasure (Jordan, 1998), fun (Draper, 1999), beauty (Tractinsky, Katz, & Ikar 2000) and hedonics (Hassenzahl, Platz, Burmester, & Lehner 2000), which demonstrated their relevance for the acceptance and appeal of interactive products and which were often placed in context of natural human needs. Technology has developed continuously and rapidly (Moore, 1965; Mack 2011) and interactive products of today like smartphones allow a user to complete more than just defined tasks (i.e. save a phone number). Smartphones have become a part of every day life whether at work or in private life (Karapanos, Forlizzi, & Martens, 2010) and more than 300 million smartphones were sold in 2010 (Pettey, 2011). People keep in touch with family and friends by ringing, texting or e-mailing them. Furthermore people use their smartphones for surfing the Internet, playing games or passing by at the online store, where one can get any application for almost every need (Karapanos et al., 2010).

The holistic view of the use of a product, which acknowledges the importance of aspects beyond the task-oriented, is commonly known under the expression user experience (UX) in the research field of human-computer

interaction (HCI) (Diefenbach & Hassenzahl, 2011). The distinction between task-oriented, instrumental and self-oriented, non instrumental attributes of interactive products is widespread in the available models of UX but a final definition of UX is absent and it is not clear how to take all these self oriented aspects into account for building overall evaluative judgments of an interactive product also because of its wide variety of meanings (Bargas-Avila & Hornbbæk, 2011). According to Hassenzahl and Tractinscy (2006) the research field of UX reveals three major perspectives, which are connected closely and often overlap. The first stresses the importance of non-instrumental aspects (i.e aesthetics) (Norman, 2004) the second deals with affective and emotional aspects of the interaction (Scheier, Fernandez, Klein, & Picard, 2002) and the third is the experiential perspective of UX, which emphasizes the situation and its temporality of the technology use (Forlizzi & Battarbee, 2004). We will concentrate mostly on the non-instrumental aspects because the presented study focuses on this major perspective. The beginning of academic discourse on the relation between usability and aesthetic can be set in 2000 (Tractinsky et al., 2000). According to Tuch, Hornbaeck, Opwis & Bargas-Avila (2011) aesthetics is one of the most frequently researched dimension in the field of UX today. Tractinsky et al. (2000) examined the association between aesthetics and usability of automatic teller machines (ATM). The participants had to rate user interfaces, which were different according to their aesthetics. They found a significant relation between perceived aesthetics and usability, and summarized their findings with the catch phrase “what is beautiful is usable”. The authors stated that this relation is believed to be similar to the “what is beautiful is good” stereotype well known in social psychology (Dion, Berscheid, & Walster, 1972).

This stereotype describes a cognitive bias whereby the perception of one attribute of a person or an object is influenced by the perception of another attribute of that person or object (i.e. judging a good-looking person as more intelligent) (Eagly, Ashmore, Makhijani, & Longo, 1991). However, nowadays this statement is controversial, as well as some methodological aspects of the study by Tractinsky et al. (2000). Firstly, the constructs aesthetic and usability were measured with a single self-generated item. Asking participants about general ease-of-use with a single item may not capture the usability but rather an overall judgment of the ATMs (Hassenzahl, 2004). Secondly, the manner in which they varied the aesthetic of the ATMs is questionable. It is possible that, the manipulation of the aesthetic factors through moving or changing certain interface-elements changed the usability of the interface (Tractinsky et al. 2000). Therefore, the correlation found by Tractinsky (1997, 2000) may rather reflect a distribution of attributes in the study object pool, than the perceptions of the participants. Others found no relationship between usability and aesthetics (e.g., Hassenzahl, 2004; van Schaik & Ling, 2008). Hassenzahl and colleagues (2000) developed the questionnaire AttrakDiff2 with the main aim to separate instrumental and non-instrumental product aspects, which they named pragmatic (PQ) and hedonic qualities (HQ). They investigated the relation of appeal to PQ and HQ and could show that these two quality aspects can be perceived independently from each other. Three years later they developed the second version of the questionnaire, the AttrakDiff2, because their results showed that the hedonic quality of a product is related to two different aspects (Hassenzahl, Burmester, & Koller 2003). This resulted in hedonic quality stimulation (HQ-S) and hedonic quality identification (HQ-I).

HQ-S refers to the natural human need to evolve constantly (i.e the improvement of knowledge and skills) (Csikszentmihalyi, 1975; Maslow 1943). Products can support this development, whilst they are stimulating. For example, novel, interesting and exciting features, the content, interaction and presentation styles can increase attention, reduce motivation problems or facilitate finding new solutions to existing problems (Hassenzahl et al. 2003). Typical HQ-S attributes are “creative”, “courageous” or “innovative” (Hassenzahl et al., 2003).

HQ-I stresses the natural human need to communicate information about our self-identity in social context (Belk, 1988). People want to be perceived by relevant others in a specific manner. A product can support this by communicating a desired identity (Prentice, 1987). HQ-I is the property of a product to satisfy the natural need of the people to bring their self-expression through objects. Typical HQ-I attributes are “presentable”, “integrating” or “professional” (Hassenzahl et al., 2003).

PQ is the sum of the different attributes of an interactive product, which are connected to the user needs to accomplish behavioural goals and was used often as measurement of perceived usability (Alex et al., (2011); Mahlke, 2007). A high pragmatic quality of a product will be perceived if it holds effective and efficient ways to achieve different tasks successfully (Hassenzahl et al., 2003). Typical pragmatic attributes are ‘‘efficient’’, ‘‘practical’’, ‘‘useful’’ or ‘‘necessary’’. The influence of the perceived pragmatic and hedonic qualities on overall evaluative judgments (i.e. goodness or beauty) was investigated recently in a few studies (Hassenzahl 2004; Hassenzahl & Tractinsky 2006; Mahlke, 2007, Van Schaik & Ling 2008). In two studies by Hassenzahl 2004, they investigated

the influence of pragmatic and hedonic quality on the appeal of different mp3 player surfaces (skins) before (Study 1 and 2) and after their use (only Study 2) with the AttrakDiff2 questionnaire. They measured the perceived PQ and HQ-I & HQ-S as well as the overall evaluative judgments of beauty and goodness of the different skins in both studies. Goodness is a user's evaluation of the overall product quality and beauty is a user's evaluation of the appearance of a product (Hassenzahl, 2004). The results of study 1 showed a strong correlation before use between beauty and HQ-I and also between goodness and HQ-I. The second study, which took measures before and after the use of the mp3 player software, also showed a strong correlation between HQ-I and goodness and no correlation between goodness and PQ before the use. After the use, goodness became strongly correlated to PQ and the correlation between HQ-I and goodness disappeared. The correlation between goodness and the quality aspects HQ-I and PQ seemed to be more influenced by use of the system than the correlations between beauty and PQ, HQ-S and HQ-I. Van Schaik and Ling (2008) used a short version of the AttrakDiff2 questionnaire and showed that judgment about beauty was not related to the manipulations of the usability of websites before and shortly after the use of a website. In contrast, beauty was correlated by hedonic attributes HQ-I and HQ-S. Goodness was correlated to HQ-I, HQ-S and PQ before and after use. Summarized the perceived hedonic qualities were more stable than pragmatic quality and beauty was more stable than goodness with experience of website use in this study.

The systematic change of experience over time is a mainly overlooked aspect in this research field (Diefenbach & Hassenzahl, 2011; Hornbbæk, 2006). The perceived qualities and also the importance we attach to these different

qualities may change over time (Karapanos, Hassenzahl, Forlizzi, & Martens, 2009). According to Karapanos and Forlozzi (2010) users actually face learnability problems primarily in the first week of use. After this first week usability seems to sharply increase over time. A newly bought smartphone exemplifies this statement. At the beginning the user is excited about the novelty of his new smartphone and he may focus on its usability because he has to learn how to operate it (Mendoza, 2005). After a while the user eventually changes his perception of its usability. After using it for some time, the user is able to handle it and focuses more on new functionality because the user is less worried about its usability (Hazlett, 2003).

The only way to be sure that such changes happen over time is to collect data continuously but such longitudinal studies are rare because they are very resource consuming (money, time, maintenance)(Bortz, 2005). Hassenzahl and Tractinsky (2006) used the interview method CORPUS to investigate retrospectively two pragmatic (utility, usability,) and three hedonic dimensions (stimulation, identification and beauty) of mobile phones over 20 months. In this study the perceived pragmatic dimensions remained stable or improved over time, while hedonic aspects decreased or remained stable. The explanation of the authors was that, further familiarity led to a better understanding of the product and thus, to a better perceived usability (Mendoza, 2005). At the same time, however, the product lost its power to surprise and stimulate. Note, that this does not necessarily imply a causal relation between both dimensions. Very different features of a product may make it usable or stimulating. In another study Hassenzahl et al. (2008) used the AttrakDiff2 to rate a TV pointing device at two different times. In this explorative study the objective was to take a closer

look at the relation between the perceived product qualities (PQ, HQ-I and HQ-S), and the overall evaluative judgments of beauty and goodness after 1 week of use and after 4 weeks of use. During the first week, beauty judgments related mostly to attributes of HQ-S then HQ-I. PQ was not related to beauty judgements at all. The relation between HQ-S and beauty seemed to weaken after 4 weeks of use. HQ-I was the most related quality to beauty judgments at the end. They confirmed earlier findings (Hassenzahl, 2004; Mahlke 2007) with their results, which showed that beauty was strongly correlated with HQ-I. Furthermore HQ-S was even more prominent than HQ-I in the beginning, which might be contributed by the novelty of the device (Hassenzahl et al., 2008). Judgments of goodness related mostly to PQ during the first week and to one attribute reflecting HQ-I (i.e. presentable). After four weeks of use, goodness appeared to be related mostly to HQ-I while a weaker relation can be observed with attributes relating to HQ-S (i.e. creative, courageous, original, creative). PQ lost its influence on the judgment of goodness after 4 weeks. Karapanos et al. (2009) conducted a longitudinal study to examine the UX of six individuals beginning at the purchase of the Apple iPhone. They followed six individuals for a period of 4 weeks. Each day, the participants were asked, to mentally reconstruct their daily activities and experiences that were somehow related to the iPhone. Afterwards the participants had to report the three most impactful experiences of the day in the form of one-paragraph essays. Additionally, they had to fill out a short version of the AttrakDiff2 questionnaire for each reported experience. In the short version of the AttrakDiff2 questionnaire PQ is divided into two aspects, namely the ease of use and usefulness of a product. These two pragmatic aspects as well as HQ-I, HQ-S and the overall evaluative judgments were each measured

with just one single item. The results of Karapanos et al. (2009) showed that goodness was strongly correlated with ease of use und HQ-S in the beginning. The correlation with ease of use was stable over time, but the correlation between stimulation and goodness became weaker and after 4 weeks there was no significant correlation anymore. Identification was not correlated to goodness at the beginning but over time it correlated more and more. Beauty was always highly related to HQ-I and partially to HQ-S. There was no significant correlation between the two usability measurements and beauty.

The AttracDiff2 is a widely accepted and validated questionnaire in the research field of UX and has already been used to take a closer look at different interactive products but there is a lack of longitudinal studies (Hassenzahl, 2004; Mahlke, 2007; Van Schaik & Ling, 2008). There are a few longitudinal studies, which have actually assessed pragmatic and hedonic qualities continuously and not retrospectively with the AttracDiff2 but with limitations: either, not the whole AttracDiff2 was used (Karapanos et al., 2009) or one did not measure from the first interactions (Karapanos et al., 2008). In addition the pragmatic and hedonic aspects, as well as the overall evaluative judgments and their relationship to each other, have never been measured from the first interactions over a period longer than 4 weeks (Karapanos, 2009).

Summary of the study and hypothesis

The present study investigates how different perceived product qualities and overall evaluation judgments of an interactive product, the iPhone 4, develop over time. In contrast to previous studies data was assessed not retrospectively but using on going measurements over 8 weeks. To capture changes in perception of iPhone 4 users, participants were examined five times

beginning at the purchase of the smartphone till the 8 weeks after the purchase. Additionally we took a closer look at how different perceived product qualities (PQ, HQ-I & HQ-S) affect overall evaluation judgments (beauty & goodness) over this time span. A longitudinal study, where each participant had to fill out the AttrakDiff2 questionnaire five times within 8 weeks was developed. Following hypotheses were proposed:

*Development-hypotheses***.** Based on the theory above we expect (H_1) the perceived pragmatic quality to increase over time. Furthermore we also expect (H_2) the evaluation of goodness to increase over time. The ratings of HQ-I (H_3), HQ-S (H_4) and beauty (H_5) are expected to be stable over 8 weeks.

*Correlation-hypotheses***.** We assume (H_6) there is a significant correlation between PQ and goodness from the beginning, which lasts over the entire time span of 8 weeks. Additionally we expect (H_7) a significant correlation between HQ-S and goodness, which should weaken over time. In contrast we expect (H_8) an increasing correlation between HQ-I and goodness. Furthermore, it is expected that beauty and PQ do not correlate at all (H_9). On the contrary the hedonic qualities have always played a central role in the evaluation of beauty. Thus it is assumed that (H_{10}) HQ-I and (H_{11}) HQ-S are correlated to beauty across all times of measurement.

Method

Design

We used a within subject design for this longitudinal study. The participants had to fill out a survey five times over a time span of 8 weeks. The independent variable chosen was the factor time of measurement. The dependent variables were pragmatic quality (PQ), hedonic quality-identification

(HQ-I), hedonic quality-stimulation (HQ-S) and the overall evaluations of beauty and goodness.

Materials

The study was conducted online with the survey tool Unipark (EFS Survey 8.0). Participants took part in the study with his or her newly bought iPhone 4.

Measurements

At each of the five points of measurements the participants had to fill out the exact same survey that consisted of different validated questionnaires. In this paper we focussed mainly on one questionnaire in the survey, namely the questionnaire AttrakDiff2 developed and validated by Hassenzahl et al. (2003). The AttrakDiff2 consists of twenty-one 7-point items with semantic differential (i.e., *impractical –practical*) – seven items for each quality. Additionally, the overall evaluative judgments goodness (*bad – good*) and beauty (*ugly – beautiful*) were each measured with a single 7-point differential item. All items are presented in Appendix A. The respective item values of PQ, HQ-S and used HQ-I scores were averaged per participant to assess the scores of the different qualities (Hassenzahl et al., 2003). We conducted Cronbach alpha to check the internal consistency for the dependent variables PQ, HQ-S & HQ-I. Two values – HQ-S and HQ-I did not meet the criterion "acceptable" ($\alpha > 0.7$) at the first time of measurement. Table 1 summarizes all Cronbach alpha values.

Table 1

Cronbach alpha values PQ, HQ-I and HQ-S

	Week 0	Week 1	Week 3	Week 5	Week 8
PQ	.69*	.84	.85	.78	.88
HQ-I	.64*	.85	.90	.92	.94
HQ-S	.81	.92	.80	.84	.79

To complete the picture, the following research variables were also assessed in the survey. Self- Efficacy was assessed using the 33-item scale developed and validated by Compeau and Higgins (1995) and need for cognition was measured with the questionnaire proposed by Bless, Wänke, Bohner, Fellhauer and Schwarz (1994). Perceived usefulness and ease of use was assessed from those who developed and repeatedly validated these scales (Davis, 1989). Cognitive Absorption was measured using the variables of Agarwal and Karahannal (2000) while personal innovativeness was measured using the items suggested by Agarwal and Prasad (1998). Finally we used Webster and Martocchio's (1990) seven-item scale to assess computer playfulness.

Procedure

The participants were individually invited per E-mail to the first survey after they contacted us via E-mail. Therefore they could fill out the survey on any computer with Internet access. Besides formal information about the study, the invitation included a password and a personalised hyperlink, which carried the participants to the login screen of the survey. Each participant could only take part in the study if he or she filled out the first survey within the first 72 hours after the purchase of the smartphone. The first survey was followed by four more. The hyperlink for the second survey was sent one week after the purchase. Between the second and the third and between the third and the fourth survey two weeks went by. Finally, the fifth survey was sent three weeks after the fourth, respectively eight weeks after the purchase of the smartphone. They all had to fill out each survey within 72 hours after receiving the invitation per E-mail to participate further in the study.

Participants

The participants were mainly recruited at six Swisscom mobile shops and also by using the participant database of the Department of Psychology Basel, which consisted of a random population sample. The participants were either given an iPod Shuffle or 60 CHF for participating in the study at the end of the last questionnaire.

In total 32 people filled out the first survey. Overall, there were 17 participants who filled out all five surveys within the time limit and were rewarded. The smartphone models and their operation systems (OS) being tested varied across participants. With the purpose of eliminating the variability in our results, which might be attributed to this diversity of the smartphone configurations we decided to choose only the most frequently model and OS. We choose the iPhone 4 because it was the most frequent one and the preinstalled OS of the iPhone 4 is always the same. Twelve iPhone 4 (70.59%) users could be identified. They were further analysed to check how many of them were actually "first time users". A participant was categorized as a first time user if he or she was using the operations system (OS) of Apple (iOS) for the first time no matter which version. This was ensured by looking at the mobile history of the participants. Finally this selection lead to a sample size of n = 8. The sample consisted of 6 females (75%) and 2 males (25%). They ranged in age from 23 to 53 (M = 31.75, *SD* = 9.22). They rated their computer and Internet skills on a 7-point Likert-scale where 1 represented the lowest and 7 the highest skill. The average self-rated computer skill was 5.5 (*SD* = .76) and the average self-rated Internet skill was 5.375 (*SD* = .92). Summarized, these participants rated their

computer and Internet knowledge as good. The data suggest that these participants were well experienced in handling the Internet and a computer.

Results

An alpha level of .05 was used for all statistical tests. Additionally all data were checked to see if they meet the required conditions for the statistical tests. We included 8 people in our analyses using SPSS 20 as program for computing our analyses. No data was missing. We scanned the data as recommended by Tabachnik and Fidell (2006) for outliers by transforming the data into Z-scores. None of the outliers were excluded because they were not more than 3 standard deviations apart from the mean. Furthermore we tested if the distribution of the data was normal across the different depend variables. As the distribution of goodness and beauty showed non normal distribution and severe negative skewness we decided to use non-parametric statistical procedure for these two variables.

Subjective Measures

The means and standard deviations of the depended variables across all times of measurements were conducted. The descriptive data is shown in Table 2. Additionally the development of the means was plotted and can be found in the Appendix B. The means of beauty and goodness were near the maximum of 7 points over 8 weeks and as mentioned before the distribution of goodness and beaut*y* showed non-normal distribution and severe negative skewness. PQ, HQ-I and HQ-S were normally distributed and in average one point below in comparison to the overall evaluative judgments. The perceived HQ-I and PQ showed a slight decline in the first week. whereas HQ-S showed a slight increase in the same period.

Table 2

Anova for perceived beauty, goodness, PQ, HQ-I and HQ-I over 8 weeks

	Week 0	Week 1	Week 3	Week 5	Week 8	*F*	η_p^2	*p*
Beauty[1]								
M	6.63	6.50	6.63	6.50	6.63	.34	.05	.85
(SD)	(.74)	(.76)	(.74)	(.76)	(.74)			
Goodness[1]								
M	6.38	6.38	6.25	6.5	6.25	.50	.08	.75
(SD)	(.74)	(.52)	(.71)	(.76)	(.71)			
PQ								
M	5.55	5.29	5.57	5.43	5.57	1.41	.17	.26
(SD)	(.62)	(.91)	(.82)	(.77)	(.82)			
HQ-I								
M	5.86	5.59	5.55	5.64	5.55	.97	.14	.44
(SD)	(.69)	(.90)	(1.18)	(1.05)	(1.18)			
HQ-S								
M	5.51	5.71	5.38	5.54	5.38	.70	.09	.60
(SD)	(.63)	(.66)	(.86)	(.66)	(.86)			

Note. [1]Data not normal distributed, for better comparability ANOVA results are shown. Friedmann test showed the same results.

Development of subjective measures

To check if the values of the different dependent variables changed significantly over time we ran the Friedman test for the variables goodness & beauty and calculated one-way ANOVAs for the variables PQ, HQ-S & HQ-I. The Friedmann test for related samples with the factors goodness as dependent variable and the factor time of measurement was conducted. Contrary to our assumption there were no significant differences between the times of measurements of goodness ($\chi^2(5) = 2.214$, $p = .696$). Supporting our hypothesis, the values for beauty did not significantly differ between the times of measurements ($\chi^2(5) = 1.50$, $p = .817$). The Mauchly's test of sphericity was calculated for the dependent variables PQ, HQ-S and HQ-I showed no significances therefore the variances for each set of difference scores were equal. An one-way ANOVA for related samples with PQ as the dependent

variable was calculated. Contrary to our predictions there were no significant differences between the different times of measurements ($F(1, 7) = 1.41$, $p = .26$, $\eta_p^2 = .17$). In contrast the one-way ANOVAs for related samples with HQ-I and HQ-S as the dependent variables supported our hypotheses. No significant differences between the different times of measurements of HQ-I ($F(1, 7) = .97$, $p = .44$, $\eta_p^2 = .14$) respectively of HQ-S ($F(1,7) = .70$, $p = .597$ $\eta_p^2 = .09$) were found.

Development of correlations between subjective measures

Kendall's tau (τ) was used to calculate the correlations between the different dependent variables because the distribution of goodness and beauty showed non normal distribution and severe negative skewness. We chose Kendall's tau instead of Spearman's rho because we had a small data set with a large number of tied ranks (Field, 2009). The correlations between the overall evaluations beauty and goodness with the three product qualities PQ, HQ-S and HQ-I are shown in Table 3 & 4. Furthermore the developments of the correlation coefficients are plotted in Figure 1 & 2. The results pertaining to the correlations between goodness and PQ, HQ-I and HQ-S are presented first.

Table. 3

Kendall's tau (τ) correlations between goodness and PQ, HQ-I and HQ-S

	Week 0	Week 1	Week 3	Week 5	Week 8
PQ - goodness	.13	0.34	0.30	0.32	0.57*
p	(.34)	(.15)	(.17)	(.16)	(.04)
HQ-I - goodness	0.05	0.5	0.75*	0.60*	0.77*
p	(.45)	(.07)	(.01)	(.03)	(.01)
HQ-S - goodness	0.09	-0.05	0.74*	0.61*	0.78**
p	(.40)	(.44)	(.01)	(.03)	(.01)

*Sign p < 0,05 (one-tailed) **Sign p < 0,01 (one-tailed)

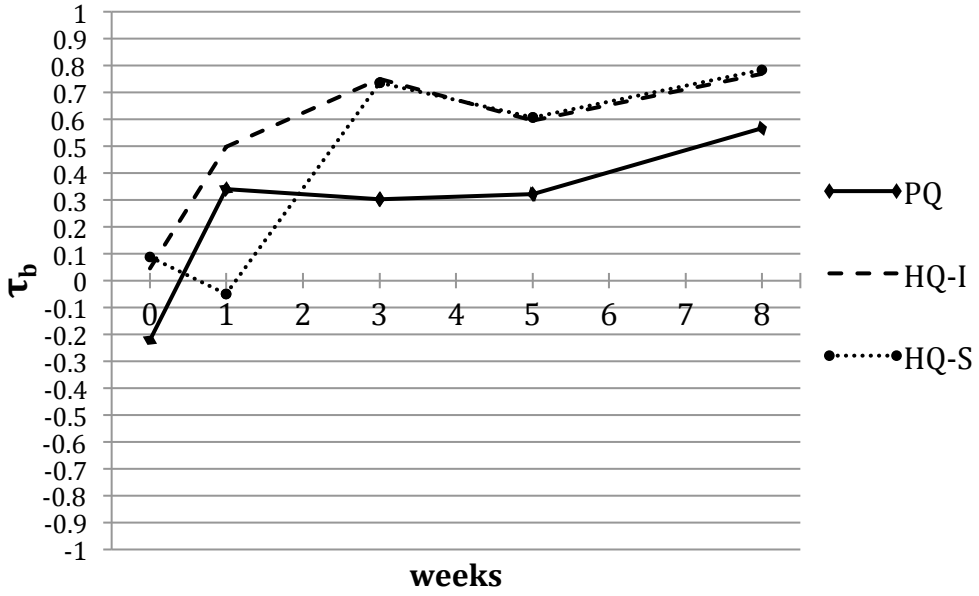

Figure 1. Development of the correlations (τ_b) between goodness and PQ, HQ-I and HQ-S

Contrary to our predictions the data revealed only one significant correlation between PQ and goodness at the fifth time of measurement that was strong (τb(8) = .57). Partially supporting our assumptions the correlations between HQ-I and goodness were significant for the third (τb(8) = .75), the fourth (τb(8) = .60) and the fifth survey (τb(8) = .77). Additionally the correlations between HQ-S and goodness were significant for the third (τb (8) = .75), the forth (τb(8) = .60) and the fifth survey (τb(8) = .77) and gave additional support to our predictions. All significant correlations were very strong (τb > .5).

The results of the analysis of the correlations between beauty and PQ, HQ-I and HQ-S are presented next. As expected the data showed no significant correlation between PQ and beauty at any time of measurement. Furthermore the correlations between HQ-I and beauty were significant for the second, (τ_b (8) = .70), third (τ_b (8) = .59), the forth (τ_b (8) = .60) and the fifth survey (τ_b (8) = .77). Additionally the correlations between HQ-S and beauty showed

significance for the third (τ_b (8) = .68), the forth (τ_b (8) = .7) and the fifth survey (τ_b (8) = .65). All significant correlations were very strong (τ_b > .5).

Table. 4

Kendall's tau (τ) correlations between beauty and PQ, HQ-I and HQ-S

	Week 0	Week 1	Week 3	Week 5	Week 8
PQ - beauty	0.11	0.14	0.16	0.23	0.16
p	(.37)	(.33)	(.31)	(.23)	(.31)
HQ-I - beauty	0.11	0.7*	0.59*	0.69*	0.59*
p	(.37)	(.01)	(.03)	(.02)	(.03)
HQ-S - beauty	0.37	0.28	0.68*	0.7*	0.65*
p	(.12)	(.19)	(.02)	(.01)	(.02)

*Sign p < 0,05 (one-tailed) **Sign p < 0,01 (one-tailed)

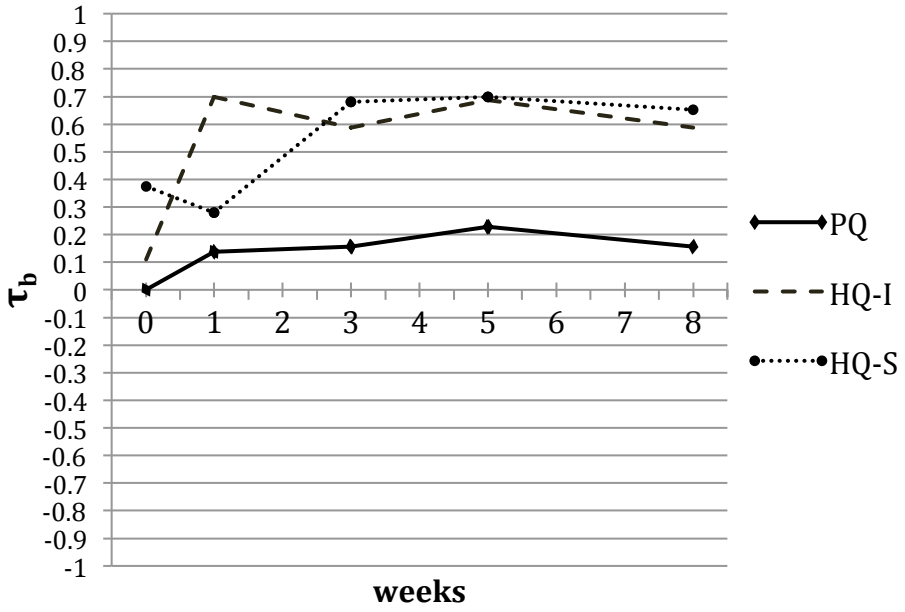

Figure 2. Development of the correlations (τ_b) between beauty and PQ, HQ-I and HQ-S

Discussion

The study showed that dependent variables were stable and that different correlations occurred between the overall evaluative judgments and the product qualities over 8 weeks. Contrary to our predictions and to recent studies, which indicated that the perception of usability normally increases

because of the learning process (Hassenzahl & Tractinsky 2006, Karapanos et al., 2008), PG remained stable and did not differ between the surveys (H_1). The participants reported a high and stable perceived usability from the beginning till the end of the study. The stable perception of PQ can be explained by the small sample size (n=8) in this study, which results in a smaller power (Bortz, 2005) to detect possible differences.

Additionally we observed that the participants, who received their new phone directly in a mobile shop, received an introduction by the shop employees. They transferred the contacts from the old phones to the new and gave the customers some essential information on how to handle the new iPhone (i.e. they explained the lock screen). This may have led to a decrease of potential usability problems in the beginning. Contrary to our assumptions (H_2) the evaluation of goodness, which was correlated with PG in early studies (Karapanos et al., 2009), was high and stable over 8 weeks. If the participants experienced major usability problems during this study, these did not influence the overall judgment of goodness. This is in line with van Schaik & Ling (2011) who stated that perceptions of usability problems can influence overall evaluation judgments but do not have to (van Schaik & Ling 2011). The perceived beauty, HQ-S and HQ-I remained stable over the time which is in line with our assumptions (H_3), (H_4), & (H_5) and earlier mentioned theories (i.e Van Schaik & Ling, 2008). Partially in opposition to our assumptions (H_6) and also previous research (Hassenzahl, 2004; Van Schaik, & Ling 2011) goodness and PQ were only significantly correlated after 8 weeks and not all the time. There was no correlation at all between goodness and PQ in the first survey and from the second till the fourth survey the correlations between PQ and goodness were

moderately strong but not significant (r > .3). Only the last survey, after 8 weeks, showed a significantly strong correlation between goodness and PG. PG clearly became a more prominent influence on the overall evaluative judgment of the iPhone 4 over time, which is in line with Karapanos et al., (2009), who stated that long-term usability plays a key role concerning satisfying and dissatisfying experiences over time. But how can the lack of correlation between PG and goodness in the beginning be explained?

One has to be aware of the fact, that the iPhone is very expensive and that possible experienced usability problems in the beginning might have generated cognitive dissonance. Cognitive dissonance is a discomfort caused by holding conflicting ideas simultaneously (Festinger, 1957). In our study this would be the high price and the eventual experienced usability problems. According to the dissonance theory human beings strive for reducing possible dissonance by changing their beliefs. In other words, they might ignore usability problems at the beginning and just hold a high judgment of goodness because of the high price. Contrary to our expectations and previous research (Van Schaik & Ling, 2008; Karapanos et al., 2009), where stimulation was a prominent aspect for beauty at the beginning, HQ-S (H_7) did not correlate with goodness until the third week. Besides the possible dissonance in the beginning we can only speculate why HQ-S began to correlate with goodness after the third week. One speculation is that the participants needed some time to find and download applications, which had a stimulating effect on them. This seems reasonable due to the fact that there are more than half a million applications available for the iPhone at the moment (Grothaus,2011). According to our assumptions and previous findings (Karapanos et al., 2008) HQ-I (H_8) showed an increasing

correlation with goodness, which became significant after the third week till the end of the study. HQ-I and HQ-S seemed to be increasingly important factors in the overall evaluation of goodness of the iPhone 4. In line with our assumptions PQ (H9) was never correlated with the overall evaluation of beauty, which is consistent with earlier findings (Hassenzahl, 2004; Van Schaik & Ling 2009). Never the less it is important to note, that the relationship between PQ and beauty has shown to be dependent on the characteristic of PQ under some circumstances. Tuch et al. (2011) recently showed that frustration of poor usability may lower ratings on perceived aesthetics. Contrary to our expectations and recent research neither (Van Schaik & Ling, 2011) (H_{10}) HQ-I nor HQ-S (H_{11}) showed a significant correlation with beauty in the first survey. This might be due to earlier mentioned reasons concerning the dissonance theory (Festinger, 1957). After the first survey HQ-I seemed to become and remain an important aspect of beauty as mentioned in previous findings (Karapanos et al., 2008). However on the other hand HQ-S was still not correlated with beauty in the second survey. In opposition to previous findings (Karapanos et al., 2009) HQ-S did not become a prominent aspect of beauty until the third week. In the study of Karapanos et al. (2009) stimulation was only measured with a single item, which was selected by the fact that this item had had the highest loading on the construct in a prior study (Karapanos et al., 2008). Therefore it is unclear if this item captured the whole essence of the stimulation or mainly one aspect (i.e. novelty).

Limitations

The small sample size limited us in the statistical analyses, for example it did not allow us to include co-variables in our analysis (like the need for

cognition or self-efficacy) to make more differentiated statements, which would be of interest. Another limitation is the low external validity because of the fact that we have only tested one specific model of a smartphone. It is not clear if the findings could be adapted to other smartphones. Additionally we realised that it would have been helpful to collect information about the usage of the smartphone for example by asking which applications they used most. It is also important for future studies to collect information about the use of other products of the same brand and the attitude to this brand because this could have had a major influence, which we did not take into account. In our case, Apple is a company with strict usability guidelines for all its products. It could be that the participants had possessed an iPod Touch before the study and therefore were well experienced and familiarized with the usability (i.e. navigation) of the iPhone before they actually bought it. We also did not consider the context in which the participant filled out the surveys (i.e. momentary affective state), which could have had a major influence on the ratings.

Future work

Further research is needed to achieve clearer results. Longitudinal studies in the future should use a larger sample size and also more variability of products to improve external validity. This could be achieved by stronger cooperation with the points of sale. We also tried to work with different service providers but unfortunately were only successful with one. Closer collaboration with points of sale might also have positive financial consequences regarding the rewards of the participants, which may result in a lower dropout rate during the surveys. Additionally other aspects like the affective state of the participants

should be considered in future research. Tuch, Hornbæck, Opwis & Bargas-Avila (2011) recently stressed the influence that affective aspects could have on judgments.

Conclusion

This study shows that different quality aspects of a product play an important role for the overall judgments at certain points in time. It became clear that interactive products of today, which are part of every day life like the iPhone, have to serve more than just instrumental goals. The findings underline the importance of hedonic aspects of interactive products and the key roll they play for outstanding quality experiences.

References

Agarwal, R., & Karahanna, E. (2000). Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage. *Mis Quarterly, 24*(4), 665-694.

Agarwal, R., & Prasad, J. (1998). A conceptual and operational definition of personal innovativeness in the domain of information technology. *Information Systems Research, 9*(2), 204-215.

Bargas-Avila, J. A., Hornbbæk, K., (2011). Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In: *Proceedings of the 2011 annual conference on Human factors in computing systems*., pp. 2689–2698. Vancouver, ACM

Batra, R., & Ahtola, O. T. (1991). Measuring the hedonic and utilitarian sources of consumer attitudes. *Marketing Letters, 2*(2), 159-170.

Belk, R. W. (1988). Possessions and the Extended Self. *Journal of Consumer Research, 15*(2), 139-168.

Bless, H., Wänke, M., Bohner, G., Fellhauer, R. F., & Schwarz, N. (1994). Need for Cognition - a Scale Measuring Engagement and Happiness in Cognitive Tasks. *Zeitschrift für Sozialpsychologie, 25*(2), 147-154.

Bortz, J. (2005). *Statistik: Für Human- und Sozialwissenschaftler.* (4th ed.) Berlin: Springer

Butler, K. A. (1996). Usability engineering turns 10. *interactions, 3*(1), 58-75.

Compeau, D. R., & Higgins, C. A. (1995). Computer self-efficacy: development of a measure and initial test. *MIS Q., 19*(2), 189-211.

Csikszentmihalyi, M. (1975). Beyond Boredom and Anxiety. *San Francisco: Jossey-Bass, 2*(3), 1.

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *Mis Quarterly, 13*(3), 319-340.

Diefenbach, S., & Hassenzahl, M. (2011). The Dilemma of the Hedonic-appreciated, but hard to justify. *Interacting with Computers*, *23*(5), 461-472.

Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology, 24*(3), 285-290.

Draper, S. W. (1999). Analysing fun as a candidate software requirement. *Personal and Ubiquitous Computing, 3*(3), 117-122.

Eagly, A. H., Ashmore, R. D., Makhijani, M. G., & Longo, L. C. (1991). What is beautiful is good but...: A meta-analytic review of research on the physical attractiveness stereotype. *Psychological Bulletin, 110*(1), 109-128.

Festinger, L. (1957). *A theory of cognitive dissonance*: Stanford University Press.

Field, A. (2009). *Discovering statistics using SPSS.* New York, NY: State University of New York Sage Publications.

Forlizzi, J., & Battarbee, K. (2004). Understanding experience in interactive systems. In: *Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques.,* pp. 261-268. Cambridge, MA, USA.

Grothaus, M. (2011, October 4). More than 18 billion apps downloaded from App Store. *TUAW*. Retrieved from http://www.tuaw.com/2011/10/04/more-than-18-billion-apps-downloaded-from-app-store/

Hassenzahl, M. (2001). The effect of perceived hedonic quality on product appealingness. *International Journal of Human-Computer Interaction, 13*(4), 481-499.

Hassenzahl, M. (2004). The interplay of beauty, goodness, and usability in interactive products. *Hum.-Comput. Interact., 19*(4), 319-349.

Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. *Mensch & Computer 2003. Interaktion in Bewegung* (pp. 187–196). Stuttgart, Leipzig: B. G. Teubner

Hassenzahl, M., Platz, A., Burmester, M., & Lehner, K. (2000). Hedonic and ergonomic quality aspects determine a software's appeal. In: *Proceeding of the SIGCHI conference on Human factors in computing systems.* (pp. 201-208). The Hague, ACM

Hassenzahl, M., & Tractinsky, N. (2006). User experience - A research agenda. *Behaviour & Information Technology, 25*(2), 91-97.

Hazlett, R. (2003). Measurement of user frustration: a biologic approach. In: *Proceeding of the CHI '03 Extended Abstracts on Human Factors in Computing Systems.* (pp. 734-735). Florida, USA: ACM

Hornbbæk, K., (2006). Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies, 64*(2), 79-102.

ISO, (1998). *Ergonomic requirements for office work with visual display terminals (VDTs) part 11: guidance on usability*. International Standards Organization, Geneva, Switzerland.

Jordan, P. W. (1998). Human factors for pleasure in product use. *Applied ergonomics, 29*(1), 25-33.

Karapanos, E., Hassenzahl, M., & Martens, B. (2008). User experience over time. In: *Proceeding of the CHI '08 extended abstracts on Human factors in computing systems.* (pp. 3561-3566). Florence, Italy: ACM

Karapanos, E., Zimmerman, J., Forlizzi, J., & Martens, J. B. (2009). User experience over time: an initial framework. In: *Proceedings of the 27th international conference on Human factors in computing systems*. (pp. 729-738). Boston, USA: ACM

Karapanos, E., Zimmerman, J., Forlizzi, J., & Martens, J. B. (2010). Measuring the dynamics of remembered experience over time. *Interacting with Computers, 22*(5), 328-335.

Karat, C.-M., Campbell, R., & Fiegel, T. (1992). Comparison of empirical testing and walkthrough methods in user interface evaluation. In: *Proceeding of the SIGCHI conference on Human factors in computing systems*. (pp. 397-404). Monterey, California, USA: ACM

Kjeldskov, J., Skov, M. B., & Stage, J. (2010). A longitudinal study of usability in health care: Does time heal? *International Journal of Medical Informatics, 79*(6), 135-143.

Mack, C. A. (2011). Fifty Years of MooreE 14 s Law. *IEEE Transactions on Semiconductor Manufacturing, 24*(2), 202-207.

Mahlke, S. (2007). Aesthetic and Symbolic Qualities as Antecedents of Overall Judgements of Interactive Products. In N. Bryan-Kinns, A. Blanford, P. Curzon & L. Nigay (Eds.), *People and Computer* (pp. 57-64). London: Springer

Maslow, A. H. (1943). A theory of human motivation. *Psychological Review, 50*(4), 370-396.

Mendoza, V., & Novick, D. G. (2005). Usability over time. In: *Proceedings of the 23rd annual international conference on Design of communication.* (pp. 151-158). Coventry, United Kingdom: ACM

Moore, G. E. (1965). Cramming More Components onto Integrated Circuits. *Electronics*, *38*(8), 114–117.

Norman, D. A. (2004). *Emotional design : why we love (or hate) everyday things*. New York: Basic Books.

Pettey, C. (2011, February 9). The power of birth order. *Gartner*. Retrieved from http://www.gartner.com/it/page.jsp?id=1543014

Prentice, D. A. (1987). Psychological correspondence of possessions, attitudes, and values. *Journal of Personality and Social Psychology, 53*(6), 993-1003.

Scheirer, J., Fernandez, R., Klein, J., & Picard, R. W. (2002). Frustrating the user on purpose: a step toward building an affective computer. *Interacting with Computers, 14*(2), 93-118.

Tabachnik, B. G, & Fidell, L. S. (2001). Cleaning up your act. Screening data prior to analysis. In B. G. Tabachnick & L. S. Fidell (Eds.), *Using Multivariate Statistics* (pp. 56–110). 4th edition. Boston: Allyn and Bacon.

Tractinsky, N. (1997). Aesthetics and apparent usability: empirically assessing cultural and methodological issues. In: *Proceedings of the SIGCHI conference on Human factors in computing systems.* (pp. 115-122). Atlanta, Georgia, United States: ACM

Tractinsky, N., Katz, A. S., & Ikar, D. (2000). What is beautiful is usable. *Interacting with Computers, 13*(2), 127-145.

Tuch, A. N., Roth, S. P., Hornbaek, K., Opwis, K., Bargas-Avila, J. (2011). *Working Towards Understanding the Relation Between Usability and Aesthetics in HCI: Mediation by Affect.* Manuscript submitted for publication.

Unipark EFS Survey (Version 8.0) [Online-survey tool]. Cologne-Huerth, Germany. Retrieved from http://www.unipark.info/

van Schaik, P., & Ling, J. (2008). Modelling user experience with web sites: Usability, hedonic value, beauty and goodness. *Interacting with Computers, 20*(3), 419-432.

van Schaik, P., & Ling, J. (2011). An integrated model of interaction experience for information retrieval in a Web-based encyclopaedia. *Interacting with Computers, 23*(1), 18-32.

von Wilamowitz-Moellendorff, M., Hassenzahl, M., & Platz, A. (2006). Dynamics of user experience: How the perceived quality of mobile phones changes over time *User Experience - Towards a unified view, Workshop at the 4th Nordic Conference on Human-Computer Interaction* (pp. 74-78).

Webster, J. (1990). The Relationship between Playfulness of Computer Interactions and Employee Productivity. *Desktop Information Technology*, 357-372.

Appendix A

Bipolar Verbal Anchors for each attribute group, beauty, and goodness

Scale	Used Anchors	Translated Anchors*
Hedonic quality-identification (HQ-I)		
HQ-I_1	Isolierend—verbindend	Isolating—integrating
HQ-I_2	Laienhaft—fachmännisch	Amateurish—professional
HQ-I_3	Stillos—stilvoll	Gaudy—classy
HQ-I_4	Minderwertig—wertvoll	Cheap—valuable
HQ-I_5	Ausgrenzend—einbeziehend	Noninclusive—inclusive
HQ-I_6	trennt mich von Leuten—bringt mich den Leuten näher	Takes me distant from people— brings me closer to people
HQ-I_7	Nicht vorzeigbar—vorzeigbar	Unpresentable—presentable
Hedonic quality-stimulation (HQ-S)		
HQ-S_1	Konventionell—originell	Typical—original
HQ-S_2	Phantasielos—kreativ	Standard—creative
HQ-S_3	Vorsichtig—mutig	Cautious—courageous
HQ-S_4	Konservativ—innovativ	Conservative—innovative
HQ-S_5	Lahm—fesselnd	Lame—exciting
HQ-S_6	Harmlos—herausfordernd	Easy—challenging
HQ-S_7	Herkömmlich—neuartig	Commonplace—new
Pragmatic quality (PQ)		
PQ_1	Technisch—menschlich	Technical—human
PQ_2	Kompliziert—einfach	Complicated—simple
PQ_3	Unpraktisch—praktisch	Impractical—practical
PQ_4	Umständlich—direkt	Cumbersome—direct
PQ_5	Unberechenbar—voraussagbar	Unpredictable—predictable
PQ_6	Verwirrend—übersichtlich	Confusing—clear
PQ_7	Widerspenstig—handhabbar	Unruly—manageable
Evaluative judgments		
Beauty	Hässlich—schön	Ugly—beautiful
Goodness	Schlecht—gut	Bad—good

Note. *English anchors were translated by Hassenzahl (2004).

Appendix B

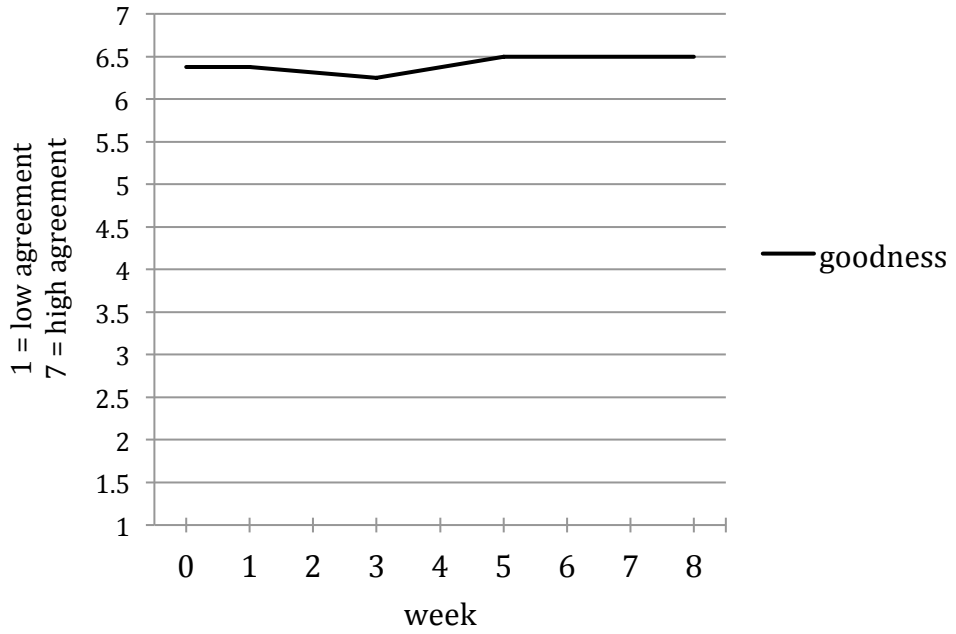

Figure 3. Development of the means of goodness.

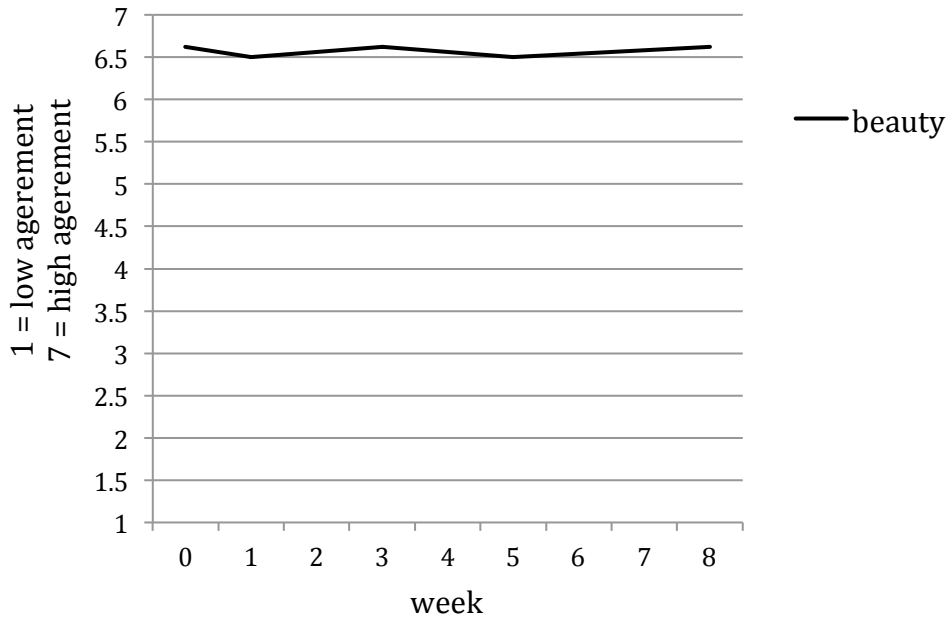

Figure 4. Development of the means of beauty.

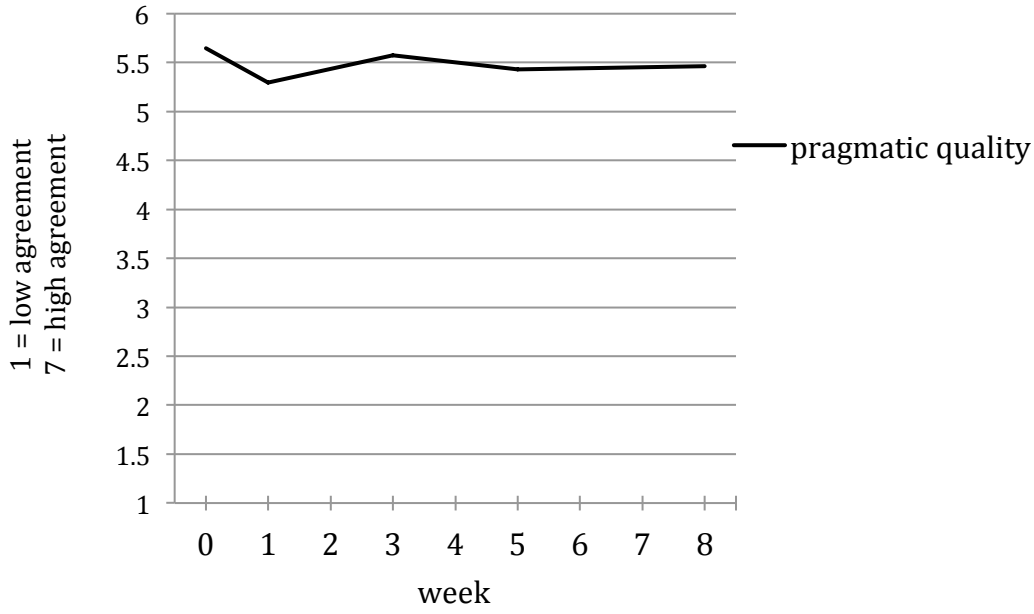

Figure 5. Development of the means of pragmatic quality.

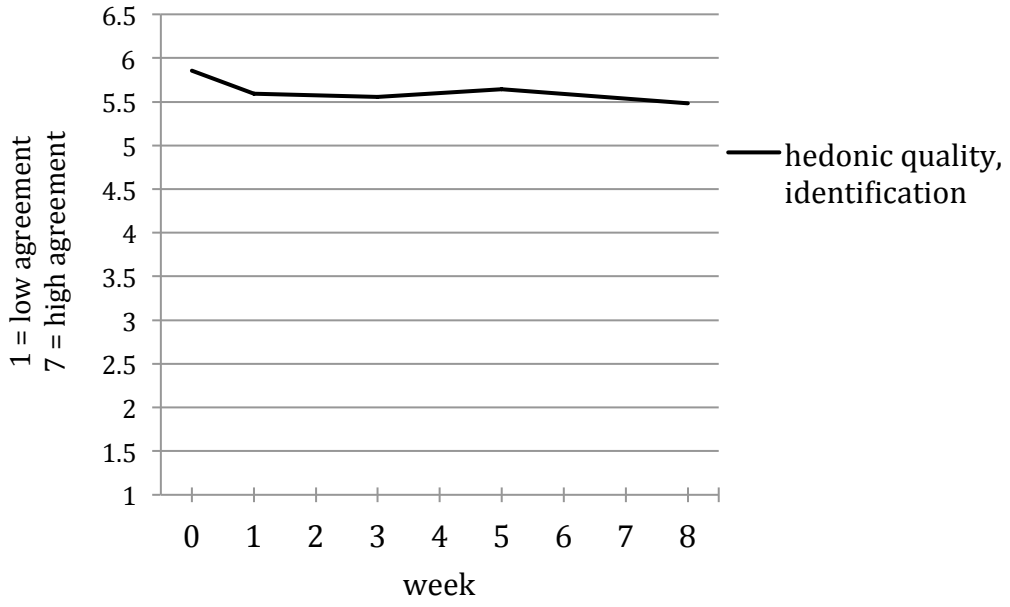

Figure 6. Development of the means of hedonic quality, identification.

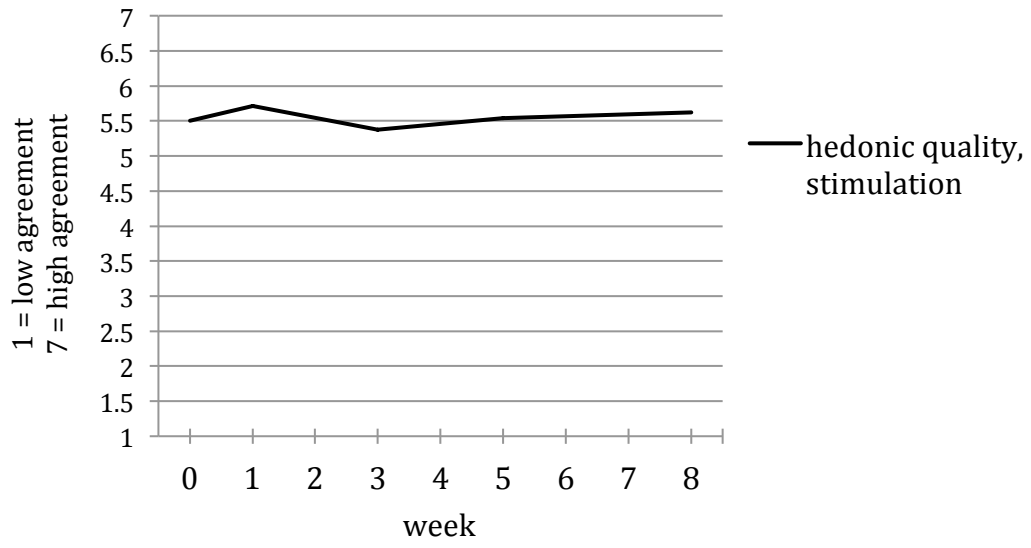

Figure 7. Development of the means of hedonic quality, stimulation.