ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/233864402

The Interplay of Beauty, Goodness, and Usability in Interactive Products

Article *in* Human-Computer Interaction · December 2004

DOI: 10.1207/s15327051hci1904_2

CITATIONS
1,115

READS
8,729

1 author:

Marc Hassenzahl
University of Siegen

308 PUBLICATIONS **20,094** CITATIONS

SEE PROFILE

All content following this page was uploaded by Marc Hassenzahl on 31 May 2014.

The user has requested enhancement of the downloaded file.

HUMAN-COMPUTER INTERACTION, 2004, Volume 19, pp. 319–349
Copyright © 2004, Lawrence Erlbaum Associates, Inc.

The Interplay of Beauty, Goodness, and Usability in Interactive Products

Marc Hassenzahl
Darmstadt University of Technology

ABSTRACT

Two studies considered the interplay between user-perceived usability (i.e., pragmatic attributes), hedonic attributes (e.g., stimulation, identification), goodness (i.e., satisfaction), and beauty of 4 different MP3-player skins. As long as beauty and goodness stress the subjective valuation of a product, both were related to each other. However, the nature of goodness and beauty was found to differ. Goodness depended on both perceived usability and hedonic attributes. Especially after using the skins, perceived usability became a strong determinant of goodness. In contrast, beauty largely depended on identification; a hedonic attribute group, which captures the product's ability to communicate important personal values to relevant others. Perceived usability as well as goodness was affected by experience (i.e., actual usability, usability problems), whereas hedonic attributes and beauty remained stable over time. All in all, the nature of beauty is rather self-oriented than goal-oriented, whereas goodness relates to both.

Marc Hassenzahl is a psychologist with an interest in applying research on social cognition, judgment, and choice to human–computer interaction; he is a researcher and lecturer in the Institute of Psychology, Social Psychology and Decision-Making, of the Darmstadt University of Technology in Germany.

CONTENTS

1. INTRODUCTION

Beauty—its nature and manifestations—has fascinated writers for ever. From the great Greek philosophers to 20th-century thinkers: Beauty mattered to them as it still matters to every human. And as long as humans are essential elements in the study of human–computer interaction (HCI), to better understand beauty must be an important endeavor of the field.

But far from it: Until recently, HCI was, at best, suspicious about beauty. “If it is pretty, it won’t work,” summarizes one of the common prejudices, and sometimes a pretty product is still accused of hiding “harm behind its beauty" (Russo & De Moraes, 2003, p. 146). Nevertheless, in the last 2 decades several attempts were made to incorporate nonutilitarian aspects, such as beauty, enjoyment, or fun, into HCI in general and usability engineering specifically (e.g., Draper, 1999; Gaver & Martin, 2000; Hassenzahl, 2002; Jordan, 2000; Monk & Frohlich, 1999). Albeit different in detail, these approaches have three aspects in common: a focus on the subjective side of usability, namely user perceptions and experiences; on the positive sides of using products (instead of simply avoiding usability problems), and on human needs as a whole. Especially the last—the holistic perspective on users—requires taking all human needs seriously and not only task-related ones, such as efficiency or effectiveness.

Although the existing literature about beauty in HCI is substantial (e.g., Norman, 2004), empirical research is only in its beginnings (e.g., Burmester, Platz, Rudolph, & Wild, 1999; Kurosu & Kashimura, 1995; Lavie & Tractinsky, 2004; Roberts et al., 2003; Schenkman & Jönsson, 2000; Tractinsky, 1997; Tractinsky, Katz, & Ikar, 2000; Tractinsky & Zmiri, inpress; Wilson, 2002). In general those studies demonstrate beauty to be a good (often the best) predictor of a product's overall impression or general user satisfaction. Moreover, a strong correlation between beauty and usability repeatedly emerged. This relation is believed to resemble the "what is beautiful is good" stereotype well known in social psychology (e.g., Dion, Berscheid, & Walster, 1972). People, for example, assume beautiful individuals to be more successful in their jobs and to be better parents solely on the basis of their looks. A meta-analysis (Eagly, Ashmore, Makhijani, & Longo, 1991) showed people to believe that a person's beauty is positively related to social competence, adjustment, potency, intellectual competence, and general "goodness." In a similar vein, beauty in an interactive product might also indicate increased usability.

Most of the empirical studies on beauty in the field of HCI suffer from a lack of a guiding research model. The consequential rather casual definition and measurement of key variables leads to methodological problems that may even cast doubts on some of the available findings and conclusions. This article's aim is to take a fresh look on the relationship between usability, beauty, and other important product attributes. Hassenzahl's (2003) model of the user–product relationship is employed to define key concepts and to serve as a basis for predicting their relation.

2. BACKGROUND

Hassenzahl's model (2003) assumes users to construct *product attributes* by combining the product's features (e.g., presentation, content, functionality, interaction) with personal expectations or standards. For example, the color and layout (i.e., product features) of a particular Web site may be new to a user and thus perceived as novel (i.e., a product attribute). A different user may perceive the same presentational style as amateurish. A *product character* is a bundle of attributes, such as innovative, comprehensible, professional. It can be understood as a cognitive structure that integrates product attributes and their covariation. On one hand, beliefs about covariation between attributes (e.g., if it looks clear, it will be easy to handle) allow for inferences beyond the merely perceived. On the other hand, those beliefs may also serve as a basis for inappropriate generalizations (i.e., stereotyping).

The model assumes that two distinct attribute groups, namely pragmatic and hedonic attributes, can describe product characters. Pragmatic attributes are connected to the users' need to achieve behavioral goals. Above all, goal achievement requires utility and usability. In this sense, a product that allows for effective and efficient goal-achievement is perceived as pragmatic (or possesses perceived pragmatic quality). In contrast, hedonic attributes are primarily related to the users' self. They can be further subdivided into *stimulation* and *identification*[1] (see also Leventhal et al., 1996; Logan, Augaitis, & Renk, 1994). Stimulation, novelty, and challenge are a prerequisite of personal development (i.e., the proliferation of knowledge and development of skills), which in turn is a basic human need (e.g., Berlyne, 1968; Csikszentmihalyi, 1975; Schwartz & Bilsky, 1987). Identification addresses the human need to express one's self through objects. This self-presentational function of products is entirely social; individuals want to be seen in specific ways by relevant others (e.g., Prentice, 1987; Wicklund & Gollwitzer, 1982). Using and possessing a product is a means to a desired self-presentation. To summarize, a product can be perceived as pragmatic because it provides effective and efficient ways to achieve behavioral goals. Moreover, it can be perceived as hedonic because it provides stimulation by its challenging and novel character or identification by communicating important personal values to relevant others.

Using a product with a particular product character in a particular situation will lead to consequences, such as emotions (e.g., satisfaction, pleasure), explicit evaluations (i.e., judgments of appeal, beauty, goodness), or overt behavior (i.e., approach, avoidance). The separation of the perception of attributes from their evaluation allows for the fact that individuals may find a product novel (an attribute) but not necessarily like it (an evaluation). In other words, perceptions of hedonic or pragmatic attributes can *potentially* lead to a

[1] In Hassenzahl (2003), I presented a third component, namely evocation. *Hedonic quality–evocation* is given, if a product is able to provoke memories and, thus, to act as a symbol of the past. However, I refrain from further discussing this component because this study did not address evocation empirically.

positive evaluation but they must not necessarily do so. In the terms of the model, usability (i.e., pragmatic quality) is understood as a bundle of low-level product attributes (e.g., clear, supporting, predictable) and beauty as a higher level evaluative construct, comparable to (but not identical with) other evaluative constructs, such as goodness or pleasantness. This notion resembles a position in the philosophy of aesthetics that views beauty as "verdictive" (i.e., an expression of authoritative judgment) and other attributes, such as elegance, as "substantive" (i.e., relating to the essence or substance; see Zangwill, 2003). Elegance, for example, describes merely one way of being beautiful. Things can be elegant without being beautiful. Moreover, what is considered as beautiful may change. In the same vein, I understand beauty as a high-level evaluative construct and perceived usability as one of its potential low-level determinants. To be usable may be a way of being beautiful but not necessarily the only one.

This distinction is supported by some empirical studies. Schenkman and Jönsson (2000) asked participants to rate several Web pages on the following dimensions: overall impression, beauty, meaningfulness, comprehension, order, legibility, and complexity. Beauty was the closest related to overall impression (an obviously evaluative construct). In addition, a principal components factor analysis revealed two independent components. One component, called Appeal, had high loadings of overall impression, beauty, and meaningfulness. The other component had high loadings of comprehension, order, legibility, and complexity, which are all usability-related product attributes. This tentatively points at the possibility that beauty is more closely related to other abstract evaluative constructs than to specific usability attributes. Similarly, Roberts et al. (2003; see also Wilson, 2002) reported beauty to have the strongest correlation with an overall rating. A reanalysis of the data published in Hassenzahl (2002) revealed beauty to be more related to other high-level constructs (e.g., goodness, pleasantness) than to pragmatic or hedonic attributes. The averaged correlation for beauty with appeal (with beauty removed from the appeal score) was .64, whereas the correlation of beauty with pragmatic attributes was only .18. Interestingly, beauty correlated more with hedonic ($r = .46$) than with pragmatic attributes ($r = .18$). In another study (Hassenzahl, Kekez, & Burmester, 2002), the relation between pragmatic attributes of Web sites and their appeal was found to depend on the participants' instruction. One half of the participants was asked to complete given tasks, the other half was instructed to "just have fun" with the Web sites. In the *task* condition a positive relation between pragmatic attributes and appeal emerged. In the *fun* condition, however, this relation disappeared. Note that only the way pragmatic attributes contributed to appeal differed between both conditions, not the rating of the Web sites' pragmatic attributes.

Study 1 explores the relation between three different groups of low-level product attributes and beauty without actual usage experience. Study 2 examines differences in the relation before and after using the product.

3. STUDY 1

The objective of Study 1 is to explore the relation between the perceived low-level attribute groups *pragmatic*, *hedonic–stimulation*, *hedonic–identification*, and *beauty*. Given the previous results on the relation between beauty and usability, one may expect beautiful products to differ from the ugly at least with regard to pragmatic quality (i.e., usability). In addition, previous research on hedonic quality (e.g., Hassenzahl, 2002; Hassenzahl, Platz, Burmester, & Lehner, 2000) found both hedonic and pragmatic attributes to significantly contribute to the general appeal of products. In those studies, appeal comprised of beauty as well as goodness and other evaluative constructs (e.g., pleasantness). Based on this, hedonic quality (stimulation or identification, or both) is as well expected to be affected by the product's level of beauty. Previously used measures did not explicitly distinguish between stimulation and identification. Thus, a prediction concerning a differential impact of level of beauty on stimulation and identification is difficult to make. However, in a recent study by Tractinsky and Zmiri (submitted) beauty was strongly related to a factor called Symbolism ($r = .72$), which captures aspects of identification. Thus, beauty may be more related to identification than to novelty and stimulation.

To better decide whether the observed relation between hedonic quality–identification (HQI) and beauty is specific for beauty or in general typical for evaluative constructs, results for beauty are further contrasted to additionally obtained goodness judgments (i.e., good–bad).

3.1. Method

Participants

Thirty-three individuals (28 women, 5 men) participated in the study. The sample's median age was 22 years (min. = 20, max. = 40). All participants were 1st- or 2nd-year students of psychology at the Darmstadt University of Technology. The study was carried out at the beginning of a course in social psychology. The participants received no compensation for their participation.

Study Objects–Stimuli

Empirical studies on beauty and usability so far either relied on simple automated teller machine (ATM) layouts (e.g., Tractinsky et al., 2000) or Web sites (Schenkman & Jönsson, 2000) as study objects. Both object types have their difficulties. The ATM layouts may be best described as impoverished. Each layout consisted of the same basic elements known from ATMs (e.g., a number block). Variations in beauty were solely due to variations in the spatial layout. Without variation on other design dimensions, such as form, color, interactional style, and so forth, results cannot easily be generalized to real products. Moreover, the spatial-layout dimension may even have an impact on both beauty and usability. For example, Gestalt psychology's laws of organization (e.g., proximity; see Goldstein, 1989, pp. 192) are regarded as both a general theory of aesthetics and a central way to a "simple and natural dialog" with an interactive product (Nielsen, 1993, p. 118). In contrast to the ATM layouts, a set of Web sites does often substantially vary on multiple design dimensions. However, additional differences in purpose, content, and functionality will further complicate the interpretation of findings.

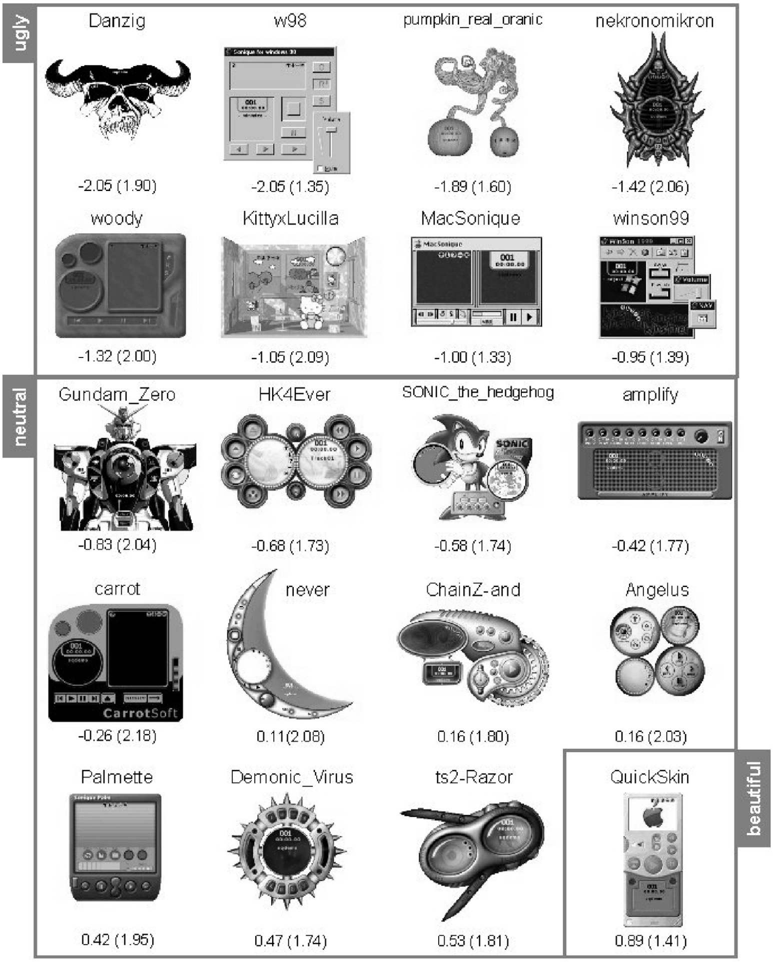
I used MP3-player software (Sonique 1.63, see http://sonique.lycos.com/) with different "skins" as my study object. A skin is a graphic file used to change the appearance of an application's user interface. Sonique skins substantially vary in presentational style and usability (due to, e.g., differences in layout, legibility, positioning of controls), although purpose and functionality remain constant. Figure 1 shows the initial selection of skins used in Study 1.

Nineteen individuals (11 women, 8 men) participated in a pretest to determine which skins from the initial selection were beautiful or ugly. A questionnaire was prepared and distributed by e-mail. On each page, a skin was presented in its original size and color together with a 7-point bipolar scale with the verbal anchors *ugly* and *beautiful.* The presentation order was counterbalanced. Figure 1 shows the mean beauty rating and standard deviation (*SD*) below each skin. Eight out of the 20 skins were rated as ugly (i.e., their mean beauty was rated as being significantly below zero). Eleven were rated as neutral (i.e., no difference from zero) and only 1 skin, namely QuickSkin, was rated as definitely beautiful (i.e., better than zero). Overall, the ratings of the skins tended to be negative, that is, more skins were rated as definitely ugly than definitely beautiful. This may be a mere consequence of the initial selection of skins. However, it is also possible that individuals find it easier to identify ugliness in this context (and express it in a rating) than beauty. In this sense, beauty is regarded as something outstanding and rare, whereas ugliness is more common.

Variables and Measurements

Based on the pretest ratings, the two ugliest (Danzig, w98) and most beautiful (ts2-Razor, QuickSkin) skins were selected to establish the independent factor *beauty* (ugly, beautiful).

Figure 1. **Initial selection of skins and mean beauty ratings (standard deviation) from pretest.**



The AttracDiff 2 questionnaire (Hassenzahl, Burmester, & Koller, 2003) was employed to measure perceived pragmatic quality (PQ), perceived hedonic quality–stimulation (HQS) and perceived hedonic quality-identification (HQI). The questionnaire consists of twenty-one 7-point items with bipolar verbal anchors (i.e., a semantic differential, see Figure 2). It is an advancement of a semantic differential previously used in studies addressing the influence of pragmatic and hedonic quality on a product's appeal (e.g., Hassenzahl, 2002; Hassenzahl et al., 2000; Kunze, 2001).

Figure 2. **Bipolar Verbal Anchors for Each Attribute Group, Beauty, and Goodness.**

Scale	Original Anchors	Translated Anchors
Hedonic quality–identification (HQI)		
HQI_1	Isolierend—verbindend	Isolating—integrating
HQI_2	Laienhaft—fachmännisch	Amateurish—professional
HQI_3	Stillos—stilvoll	Gaudy—classy
HQI_4	Minderwertig—wertvoll	Cheap—valuable
HQI_5	Ausgrenzend—einbeziehend	Noninclusive—inclusive
HQI_6	trennt mich von Leuten—bringt mich den Leuten näher	Takes me distant from people—brings me closer to people
HQI_7	Nicht vorzeigbar—vorzeigbar	Unpresentable—presentable
Hedonic quality–stimulation (HQS)		
HQS_1	Konventionell—originell	Typical—original
HQS_2	Phantasielos—kreativ	Standard—creative
HQS_3	Vorsichtig—mutig	Cautious—courageous
HQS_4	Konservativ—innovativ	Conservative—innovative
HQS_5	Lahm—fesselnd	Lame—exciting
HQS_6	Harmlos—herausfordernd	Easy—challenging
HQS_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
Evaluational constructs		
Beauty	Hässlich—schön	Ugly—beautiful
Goodness	Schlecht—gut	Bad—good

Note. Order and polarity of items was randomized

HQI, HQS and PQ scores were calculated by averaging the respective item values per participant. Internal consistency of the three scores was high (Cronbach's on the pooled values; HQI, $\alpha = .85$; HQS, $\alpha = .95$; PQ, $\alpha = .90$). A high HQI score implies a high perceived capability of communicating identity to others. HQI attributes are primarily social (i.e., outwards). A high HQS score implies a high degree of perceived novelty, stimulation and challenge. HQS attributes are primarily related to personal growth (i.e., inwards). A high PQ score primarily implies high usability.

In addition, the evaluative constructs *beauty* as well as *goodness* was measured with a single 7-point differential item each (see Figure 2).

Procedure

The study was carried out at the beginning of an undergraduate course in social psychology. The participants were split into two groups. The second group had to leave the room. Each participant of the first group received a booklet containing four identical AttracDiff 2 questionnaires, one for each skin. The skins were projected with a data projector onto the wall. The first slide contained all four skins and participants were given the instruction to subsequently rate each skin separately with the questionnaires. After the instruction, each single skin was presented again and participants were given the opportunity to rate them. The presentation order of the skins for the first group was QuickSkin, Danzig, ts2-Razor, and w98. The second group differed from the first only in a reversed presentation order (w98, ts2-Razor, Danzig, QuickSkin) to minimize order effects. The whole study took about 15 min per group.

3.2. Results

Manipulation Check

An analysis of variance (ANOVA) with skin (Danzig, w98, ts2-Razor, QuickSkin) as within-subjects factor and beauty as dependent variable revealed a highly significant main effect of skin, $F(3, 96) = 36.80$, $p < .001$. Repeated contrasts showed ts2-Razor ($M = 0.27$, $SD = 1.64$) to be significantly more beautiful than w98 ($M = -2.00$, $SD = 1.39$), $F(1, 32) = 28.93$, $p < .001$. Neither Danzig ($M = -2.46$, $SD = 1.20$) differed significantly from win 98 ($M = -2.00$, $SD = 1.39$) nor QuickSkin ($M = 0.56$, $SD = 1.49$) from ts2-Razor ($M = 0.27$, $SD = 1.64$). To summarize, skins, pre-rated to be more beautiful (ts2-Razor, QuickSkin) than others (Danzig, w98), were again perceived as significantly different in beauty.

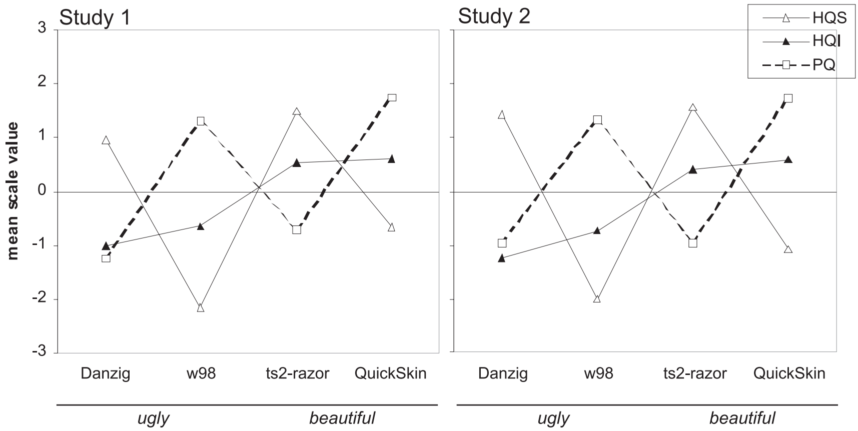
Attribute Groups

Figure 3 (left panel) shows the mean PQ, HQS, and HQI for each skin.

Interestingly, each subgroup, beautiful and ugly, consisted of one predominantly pragmatic (w98 or QuickSkin) and one predominantly stimulating (Danzig or ts2-Razor) skin. Because of this, two separate 2 × 3 ANOVAs with skin and attribute group (PQ, HQS, HQI) as within-subjects factors and score as dependent variable were performed.

The ugly skins analysis (Danzig, w98) revealed a highly significant main effect of attribute group, $F(2, 64) = 26.41$, $p < .001$, which was further qualified by a highly significant interaction of attribute group with skin, $F(2, 64) = 133.06$, $p < .001$. No main effect for skin emerged. Danzig was perceived as

Figure 3. **Mean perceived pragmatic quality (PQ), hedonic quality–stimulation (HQS) and hedonic quality–identification (HQI) for each skin (Study 1 and Study 2).**



significantly more stimulating (HQS) than w98, difference = 3.10, *t*(32) = 13.81, *p* < .01 (Bonferroni-corrected), and w98 was perceived as significantly more pragmatic (PQ) than Danzig, difference = –2.55, *t*(32) = –9.13, *p* < .01 (Bonferroni-corrected). With regard to identification (HQI), no significant difference was found (difference = –0.38).

The beautiful skins analysis (ts2-Razor, QuickSkin) revealed a highly significant interaction of attribute group with skin only, *F*(2, 64) = 152.40, *p* < .001. Neither a main effect for *skin* nor *attribute group* emerged. Ts2-Razor was perceived as significantly more stimulating (HQS) than QuickSkin, difference = 2.14, *t*(32) = 9.83, *p* < .01 (Bonferroni-corrected) and QuickSkin was perceived as significantly more pragmatic (PQ) than ts2-Razor, difference = –2.45, *t*(32) = –9.70, *p* < .01 (Bonferroni-corrected). With regard to identification (HQI), again no significant difference was found (difference = –0.07).

To compare ugly with beautiful skins, mean PQ, HQI, and HQS scores were calculated by averaging across ugly (Danzig, w98) and beautiful skins (ts2-Razor, QuickSkin) per individual. The difference between ugly and beautiful was the most pronounced for HQI, difference = 1.40, *t*(32) = 10.07, *p* < .01 (Bonferroni-corrected), followed by HQS, difference = 1.00, *t*(32) = 5.46, *p* < .01 (Bonferroni-corrected), and PQ, difference = 0.50, *t*(32) = 3.88, *p* < .01 (Bonferroni-corrected).

To summarize, within each group of ugly and beautiful skins, a predominantly stimulating (HQS) and a predominantly pragmatic (PQ) skin was found. Only identification (HQI) mirrored the beauty ratings. On average, beautiful skins were clearly better in communicating identity (HQI), were considered to be more stimulating (HQS) and only slightly, albeit significantly, more pragmatic (PQ).

Relation Between Attribute Groups, Beauty, and Goodness

A correlational analysis was performed to explore the relation between attribute groups (PQ, HQS, HQI), beauty, and goodness. Figure 4 (upper half) shows the partial correlations between beauty and PQ, HQI, and HQS for each skin (i.e., controlling for the remaining attribute groups). Only the mean partial correlation of HQI with beauty was significant and stable across skins, with a minimum of 35% and a maximum of 53% explained variance. A regression analysis of the pooled data showed HQI to be a highly significant predictor of beauty, $\beta = 0.74$, $t = 11.34$, $p < .001$. Both HQS and PQ were only marginally significant: HQS, $\beta = 0.14$, $t = 1.80$, $p < .10$; PQ, $\beta = 0.15$, $t = 1.98$, $p < .10$. All in all, the model explained 68% of the available variance (corrected). Beauty as an evaluative construct seems to be rather related to a product's ability to provide identification than to its pragmatic quality or stimulation.

To decide whether the observed relation between HQI and beauty is specific for beauty or typical for evaluative constructs, a similar analysis was done for the relation between goodness and attribute groups (PQ, HQI, HQS). The mean correlation between beauty and goodness across skins was 0.54, which is low enough (approx. 29% explained variance) to suggest differences between both evaluative constructs. Figure 4 (lower half) summarizes the partial correlations between goodness and PQ, HQI, and HQS. In contrast to beauty, HQI as well as PQ significantly correlated with goodness. A regression analysis of the pooled data showed HQI as well as PQ to be highly significant predictors of beauty: HQI, $\beta = 0.59$, $t = 8.76$, $p < .001$; PQ, $\beta = 0.43$, $t = 5.30$, $p < .001$. HQS failed to reach significance, $\beta = 0.10$, $t = 1.19$, $p < .10$. All in all, the model explained 65% of the available variance (corrected). Whereas goodness seem to be a consequence of the presence of pragmatic and hedonic attributes, beauty solely depends on the product's apparent ability to communicate identity.

3.3. Discussion

The differences in beauty of certain skins (ts2-Razor, QuickSkin) compared to others (Danzig, w98) obtained in the pretest were replicated in this experiment. Specifically, the correlation between pretest measurements of beauty and measurements from Study 1 was 0.99 and no differences in level (i.e., although the rank order remains stable, skins could be systematically rated as being more or less beautiful) were found. The beauty judgments were remarkably stable.

***Figure 4.* Partial Correlations (Percentage of Explained Variance) Between Beauty and Pragmatic Quality (PQ), Hedonic Quality–Stimulation (HQS) and Hedonic Quality–Identification (HQI) for Each Skin.**

Variable	Skin w98	Danzig	ts2-Razor	QuickSkin	M[a]
Partial correlation with beauty[b]					
PQ	.15 (2%)	–.14 (2%)	.30 (9%)	–.04 (0%)	.07 (0%)
HQI	.59 (35%)	.73 (53%)	.46 (21%)	.62 (38%)	.61** (37%)
HQS	.33 (11%)	–.08 (1%)	.20 (4%)	–.04 (0%)	.10 (1%)
Partial correlation with goodness[b]					
PQ	.26 (7%)	.50 (25%)	.45 (20%)	.42 (18%)	.41* (17%)
HQI	.52 (27%)	.44 (19%)	.51 (26%)	.50 (25%)	.49** (24%)
HQS	.25 (6%)	.22 (5%)	.33 (11%)	.11 (1%)	.23 (5%)

Note. $N = 33$.

[a] Correlations were Z-transformed, averaged and retransformed. [b] Partial correlation between attribute group and evaluative construct controlling for the remaining two attribute groups.

* $p < .05$. ** $p < .01$.

Previous research suggested a clear relation between usability (i.e., PQ) and beauty. This hypothesis was not supported. On each level of beauty (ugly, beautiful) a predominantly pragmatic and a predominantly stimulating skin was found. Although the attribute perceptions considerably varied among skins, this variation did not depend on beauty. On average, beautiful skins were perceived to be primarily better in providing identification (HQI) followed by being more stimulating (HQS) and being more pragmatic (PQ). The highest observed correlation between perceived pragmatic quality and beauty was .30 (9% explained variance). The other correlations were lower or even negative. None of the single correlations between PQ and beauty were as high as the one observed by Kurosu and Kashimura (1995) or Tractinsky (1997; Tractinsky et al., 2000). On average, the correlation was close to zero. This can be attributed neither to a lack of validity of the PQ scale nor to a lack of validity of the beauty measurement. The internal consistency of PQ, HQI, and HQS was high, which lends support to the reliability of the scales. In addition, PQ was substantially related to goodness, which would not have been found given a flawed measurement of PQ. In short, all evidence points at one conclusion: What is judged to be more beautiful is not necessarily perceived as more usable (and vice versa).

How can these results be explained in the light of the previous findings? In the most sophisticated study reporting a strong correlation between usability and beauty, namely Tractinsky et al. (2000), participants had to operate an ATM. Beauty (high, neutral, low) and usability (neutral, low) were manipulated. Before using the ATM, each participant was presented with nine different ATM layouts, selected from a previous study (Tractinsky, 1997). Each layout's beauty and usability were rated on a 10-point scale. Depending on the study condition, participants had to use the layout they rated as the most beautiful, neutral, or least beautiful to work through a series of tasks (e.g., to withdraw cash). Usability (neutral, low) was manipulated by deliberately introducing usability problems (e.g., increase in system response time, removal of shortcuts and buttons that did not operate when first pressed). After completing the tasks, the participants had to rate the ATM's beauty and usability. In addition, post-use user satisfaction was measured. Beauty and usability were significantly correlated (pre-use, $r = .66$; post-use, $r = .71$). In general pre- and post-use measures were correlated (beauty pre–post, $r = .62$: usability pre–post, $r = .48$). Post-use satisfaction correlated highly with post-use beauty ($r = .71$) and usability ($r = .87$). In addition, a multivariate analysis of covariance with beauty and usability as independent variables, pre-use usability as covariate and all post-use measurements as dependent variables, revealed a significant impact of beauty on both post-use ratings of beauty and usability. Based on the correlation between beauty and usability, as well as the finding that post-use usability ratings were not affected by actual usability, Tractinsky et al. (2002) concluded that a product's beauty is a stronger indicator for its perceived usability than its actual usability.

There are several explanations for the inconsistencies in findings between previous studies and this study. First, perceived usability as measured by Tractinsky et al. (2000) becomes likely to be treated as a high-level evaluative construct itself. The employed single-item measurement may imply a rather holistic "it seems to work"-judgment. As mentioned, evaluative constructs tend to correlate. If an object is regarded as good, other evaluations, such as desirable or inviting, will be positive too, at least in tendency. In this study, the correlation between goodness and beauty ($r = .54$) was closer to Tractinsky et al.'s (2000) findings than any single observed correlation between PQ and beauty. In addition, Tractinsky et al. (2000) found post-use usability to be highly correlated with post-use user satisfaction ($r = .87$), which supports the conclusion that participants treated the usability construct as a high level construct, close to general satisfaction. The question at hand is whether this approach to the measurement of usability—although common in the context of the Technology Acceptance Literature (e.g., Igbaria, Schiffman, & Wieckowski, 1994)—is adequate. In general, published definitions discuss *usability* as a multi-attribute concept. Part 10 of the DIN EN ISO 9241 (ISO, 1996), for example, defines seven dialog principles, such as task adequacy, self-descriptiveness, or controllability. Typically, a number of items per principle are used to operationalize those principles by related questionnaires (e.g., Gediga, Hamborg, & Düntsch, 1999; see Kirakowski, 1998, for an alternative set of attributes). In these questionnaires abstract principles are translated into concrete, answerable questions. A holistic judgment of the same quality would require participants to summarize and weight their perceptions to construct a general assessment of usability. This seems difficult for an expert and almost impossible for a layperson. Consequently, asking participants about general comfort or ease-of-use with a single question (item) may not capture the essence of usability. In addition, a correlation between such a general evaluation and beauty is not surprising.

A second alternative explanation for the stronger correlation between beauty and usability is a bias in Tractinsky's study object or stimuli pool. In a prior study (Tractinsky, 1997), individuals were presented with 26 different ATM layouts, adapted from Kurosu and Kashimura (1995), 9 of which were used in Tractinsky et al. (2000). Individuals rated beauty and usability. Ratings were then averaged across participants to form a score for each layout. Those values were correlated to support the notion that subjectively perceived usability is highly related to beauty. Correlations in three different experiments ranged from .83 to .92. However, these correlations indicate that the pool of 26 layouts included either ugly *and* unusable or beautiful *and* usable layouts. Nevertheless, it does *not* support the existence of a "what is beautiful is usable" stereotype. A stereotype is better expressed by correlations of beauty and usability within one product. Tractinsky's strategy, however, explicitly eliminated the according variance. In other words, the correlation found by Kurosu and Kashimura (1995) and Tractinsky (1997) may rather reflect a distribution of attributes in the study object pool than individuals' stereotypes. Schenkman and Jönsson's (2000) study of different Web sites demonstrates the strong influence of the actual distribution of attributes on conclusions. Their factor analysis showed beauty- and usability-related attributes to be rather uncorrelated (both loaded on distinct, orthogonal factors). This analysis was also done on ratings averaged across participants.

Tractinsky et al.'s (2000) study does not suffer from the problem of eliminating variance between individuals, because each participant used only one study object. However, it still used a sub-selection of the previously studied layouts, which are obviously biased in the distribution of beauty and perceived usability. This may—at least—have added to the reported correlation between beauty and usability.

One major limitation of Study 1 is that participants did not actually use the different skins. Pragmatic attributes may be primarily triggered by interaction with a product and not by merely looking at it. Especially hedonic attributes of a product's presentation are easy to identify and to judge at first sight, whereas pragmatic attributes of an interaction may become apparent only after attempted (and failed) usage of the player. Thus, without direct experience, beauty may be primarily based on the easier to conceive hedonic attributes, whereas with direct experience (i.e., after using the product), beauty may be more based on actual experience of how well the product worked, that is, pragmatic attributes. In other words, beauty may change its nature.

The following study extended these findings to post-use ratings to study the influence of actual use on the relation between beauty and usability.

4. STUDY 2

Pragmatic quality (i.e., usability) is an experience-based quality perception. Its true importance unfolds in goal-directed interaction with a product. Hassenzahl et al. (2002), for example, found a strong correlation between post-use ratings of PQ and appeal, if the preceding interaction was goal directed (partial correlation = .87). If participants were instructed to just "have fun" with a system, pragmatic quality was irrelevant to appeal (partial correlation = –.10).

Concerning beauty, one may argue that by using the product, beauty will change from being appearance-based to being experience-based (Djajadiningrat, Overbeeke, & Wensveen, 2000). Beauty could therefore change its nature and pragmatic quality (i.e., usability) may become a crucial determinant. In contrast, Tractinsky et al. (2000) found no significant main effect of actual usability on post-use ratings of either usability or beauty. Therefore, they concluded that actual, experienced usability neither has an effect on usability perceptions nor on judgments of beauty.

Study 2's main purpose is to shed light on the impact of usage experience on post-use ratings of the different attribute groups, beauty, and goodness. Pre-use ratings should replicate the findings of Study 1, namely a strong link between beauty and hedonic attributes. The relations among post-use ratings are difficult to predict. If beauty changes its nature from appearance-based to experience-based, pragmatic attributes should become a more important determinant. If not, beauty remains an evaluative construct largely linked to hedonic considerations and perceived pragmatic attributes will not become more influential. Post-use goodness as the most central evaluative construct, however, should be affected by experience and, therefore, show a relation to post-use PQ. In general, ratings of PQ should be more strongly affected by usage experience than ratings of hedonic quality. Moreover, as long as pragmatic quality is expected to be experienced based, pre–post correlations for PQ should be low, or at least lower than for HQI and HQS.

In addition to the variables and measures from Study 1, mental effort was assessed during the experience. Increased mental effort is an indicator for experiencing usability problems (Hassenzahl, 2000). Previous studies found post-use ratings of pragmatic quality to be negatively correlated with mental effort. In contrast, post-use ratings of hedonic quality did not correlate with mental effort (Hassenzahl, 2002). In this study, mental effort serves as an additional criterion for assessing the strength of the link between beauty and usability.

4.1. Method

Participants

Ten individuals (4 women, 6 men) participated in the study. The samples median age was 25 years (min. = 19, max. = 51). The majority of participants were students of the Darmstadt University of Technology. The participants received no compensation for their participation.

Variables and Measurements

Study 2 used the same study objects (i.e., Danzig, w98, ts2-Razor, QuickSkin) and questionnaires (i.e., AttracDiff 2, single-item measurement of beauty and goodness). The internal consistency of HQI, HQS, and PQ scores was high (Cronbach's α on the pooled values: HQI, $\alpha = .86$; HQS, $\alpha = .95$; PQ, $\alpha = .91$). After the initial rating (pre-use), participants were asked to work through two small usage scenarios. Mental effort was measured with the subjective mental effort questionnaire (SMEQ; Zijlstra & van Doorn, 1985; German translation, Eilers, Nachreiner, & Hänecke, 1986; see also Arnold, 1999) immediately after the completion of each scenario. The German version of the SMEQ is a single rating scale ranging from 0 to 220. Different verbal anchors such as *hardly effortful* or *very effortful* facilitate the rating process. A total subjective mental effort (SME_{tot}) was calculated by summing both SMEQ values per individual. After using each skin, its post-use HQI, HQS, PQ, beauty, and goodness was assessed.

Procedure

Participants were led one by one into the laboratory. They were first presented with all four skins and were asked to rate each with the help of four identical AttracDiff 2-questionnaires. All questionnaires were presented electronically. The order of presentation of the skins was varied.

After having rated all skins, participants used the first player in two scenarios. In the first scenario, music was playing while a virtual telephone rang. The participants were asked to reduce the volume (or to pause the play back) to be able to answer the call. After that, play back was to be continued. In the second scenario, participants had to load a pre-prepared play list into the player and play back a number of specific songs. Immediately after finishing each scenario, subjective mental effort was assessed with the SMEQ. Subsequent to the second SMEQ rating, the pre-use ratings of the respective skin were again presented to the participants. They were instructed to revise their former ratings on the basis of the experience gained with the skin. Emphasizing change rather than stability should reduce participants' implicit need for a consistent rating. It also prompts participants to think about the explicit impact of their actual experience on the perception and evaluation of the skin. After the rating, the next skin was used and rated. Each participant worked through all four skins. The sequence was varied. The whole experiment took about 30 min.

4.2. Results

Manipulation Check

A 4 × 2 within-subjects ANOVA, with skin (Danzig, w98, ts2-Razor, QuickSkin) and time (pre-use, post-use) as independent and beauty as dependent variable, revealed a highly significant main effect of skin, $F(2.09, 27) = 10.07$, $p < .01$ (*df* is Greenhouse-Geisser corrected). Neither the main effect of time nor the interaction of skin with time was significant. Repeated contrasts showed Danzig ($M = –2.35$) to be marginally less beautiful than w98 ($M = –1.10$), $F(1, 9) = 3.50$, $p < .10$, and w98 ($M = –1.10$) to be marginally less beautiful than ts2-Razor ($M = 0.60$), $F(1, 9) = 3.76$, $p < .10$. No difference in beauty was apparent for ts2-Razor ($M = 0.60$) and QuickSkin ($M = 0.80$). Again, Danzig was clearly rated as ugly compared to ts2-Razor and QuickSkin. W98, however, was rated as more beautiful than in the prior study, which led to an only marginal difference in beauty to the beautiful skins. Except for this slight deviation, the expected differences in beauty between skins were replicated. In addition, the level of beauty was not significantly affected by actual usage experience.

Attribute Groups

Figure 3 (right panel) shows the PQ, HQS, and HQI of each skin averaged across pre- and post-use measurements.

The obtained results are strikingly similar to the results from Study 1 (see Figure 3, left panel). Again in each subgroup, beautiful and ugly, a predominantly pragmatic (w98, QuickSkin) and a predominantly stimulating (Danzig, ts2-Razor) skin were found. For the sake of brevity, I refrain from separate analyses of ugly and beautiful skins.

Mean PQ, HQI, and HQS values were calculated by averaging across ugly (Danzig, w98) and beautiful skins (ts2-Razor, QuickSkin) per individual. A 2

× 2 × 3 within-subjects ANOVA with beauty (ugly, beautiful), time (pre-use, post-use), and attribute group (PQ, HQI, HQS) as independent variable and the mean score as the dependent variable was performed. It revealed highly significant main effects of beauty, $F(1, 9) = 17.56$, $p < .01$, and attribute group, $F(2, 18) = 6.05$, $p < .05$ (sphericity is not given, significance holds for lower-bound), as well as a highly significant interaction of beauty with attribute group, $F(2, 18) = 12.75$, $p < .01$. No other main effect or interaction approached significance. The difference between ugly and beautiful is the most pronounced for HQI, difference = 1.49, $t(9) = 5.51$, $p < .01$ (Bonferroni-corrected) followed by HQS, difference = 0.54, $t(9) = 3.19$, $p < .05$ (Bonferroni-corrected). The difference for PQ was not significant.

To summarize, within each group of ugly and beautiful skins a predominantly stimulating (HQS) and a predominantly pragmatic skin (PQ) was found. The results are strikingly similar to the results obtained in Study 1. No significant influence of usage experience on post-use ratings was found. On average, beautiful skins were again perceived to be primarily better in providing identification (HQI), followed by being more stimulating (HQS).

Relation Between Attribute Groups, Beauty, and Goodness

Figure 5 (upper half) shows the pre- and post-use partial correlations between beauty and PQ, HQI, and HQS for each skin. In general, the correlations were more inconsistent compared to Study 1. None of the mean correlations were significant and none of the correlations significantly changed from pre- to post-use. On a mere descriptive level, pre-use HQI was again the attribute with the closest relation to beauty, followed by HQS and PQ. Post-use, HQS became more and HQI less prominent. The mean correlation for PQ decreased slightly.

The mean correlation between beauty and goodness across skins pre-use was .53 and post-use was .39, which are both low enough to suggest differences between the constructs (approx. 28% and 15% explained variance). Figure 5 (lower half) summarizes the partial correlations between goodness and PQ, HQI, and HQS. The emerging pattern of the pre-use assessment resembles the pattern of Study 1 with in general slightly higher correlations between all attributes and the evaluative construct for goodness compared to beauty. However, only the mean correlation of HQI with beauty was marginally significant. Concerning the post-use assessment, PQ becomes prominent as a correlate of goodness.

In contrast to Study 1, differences between beauty and goodness before using the skins (pre-use) were not as pronounced. However, an interesting difference between post-use assessments of beauty and goodness emerged. For goodness PQ became relevant but not for beauty.

Figure 5. **Pre- and Post-use Partial Correlations (Percentage of Explained Variance) Between Beauty and Pragmatic Quality (PQ), Hedonic Quality–stimulation (HQS) and Hedonic Quality–identification (HQI) for Each Skin.**

Variables	Skin w98	Danzig	ts2–Razor	QuickSkin	M[a]
Partial correlation with beauty[b]					
pre-use					
PQ	.19 (4%)	.52 (27%)	.02 (0%)	–.22 (5%)	.14 (2%)
HQI	–.04 (0%)	.88 (77%)	.56 (31%)	.01 (0%)	.46 (21%)
HQS	.31 (10%)	–.52 (27%)	.74 (55%)	.29 (8%)	.25 (6%)
post–use					
PQ	.34 (12%)	.47 (22%)	.01 (0%)	–.51 (26%)	.08 (1%)
HQI	–.24 (6%)	.80 (64%)	.38 (14%)	.24 (6%)	.36 (13%)
HQS	.77 (59%)	–.43 (18%)	.60 (36%)	.51 (26%)	.43 (18%)
Partial correlation with goodness[b]					
pre–use					
PQ	.04 (0%)	.43 (18%)	.62 (38%)	–.10 (1%)	.27 (7%)
HQI	.16 (3%)	.80 (64%)	.70 (49%)	.34 (12%)	.55** (30%)
HQS	.43 (18%)	–.03 (0%)	.67 (45%)	–.13 (2%)	.27 (7%)
post–use					
PQ	.73 (53%)	.24 (6%)	.91 (83%)	.64 (41%)	.70* (49%)
HQI	.40 (16%)	.67 (45%)	.06 (0%)	.44 (19%)	.41 (17%)
HQS	.39 (15%)	–.05 (0%)	.32 (10%)	.04 (0%)	.18 (3%)

Note. $N = 11$.

[a] Correlations were Z–transformed. averaged and retransformed. [b] Partial correlation between attribute group and evaluative construct controlling for the remaining two attribute groups.

* $p < .05$. **$p < .01$.

Impact of Experience on Attribute Groups, Beauty, and Goodness

It was hypothesized that pragmatic quality as the more experience-based attribute should be less stable compared to hedonic quality. The mean pre–post correlation of both HQI and HQS were significant, whereas the mean pre–post correlation of PQ was not: HQI, $r = .89$, $t(8) = 5.52$, $p < .001$; HQS, $r = .94$, $t(8) = 7.79$, $p < .001$; PQ, $r = .47$. The mean pre–post correlation of beauty was significant, $r = .87$, $t(8) = 4.99$, $p < .01$, whereas the mean pre–post correlation of goodness was not, $r = .46$.

This seems to contradict findings from the ANOVAs presented earlier, where experience did not change the mean level of PQ (no main effect or interaction with time emerged). However, the low pre–post use correlation of PQ implies that single individuals do change their assessment with experience. In fact, if the absolute mean differences between the pre- and the post-use ratings of PQ were used—which is equivalent to ignoring the direction of change per individual—all differences become significant: Danzig, difference = 0.89, $t(9) = 3.26$, $p < .05$; w98, difference = 0.89, $t(9) = 3.19$, $p < .05$; ts2-Razor, difference = 0.94, $t(9) = 4.59$, $p < .01$; QuickSkin, difference = 0.80, $t(9) = 2.41$, $p < .05$. In other words, the lack of differences in pre- and post-use ratings of PQ are due to a compensatory configuration of values. Participants change their ratings of PQ, however, depending on their actual experience. Some find PQ better than before using the skin, others worse. Devaluation of some participants and up-valuation of others compensated each other, which led to the low correlation between pre- and post-use ratings of pragmatic quality.

In sum, pragmatic quality and goodness were more influenced by using the system than beauty and hedonic quality.

Subjective Mental Effort

The importance of experience for PQ is also reflected by the mean correlation across skins between subjective mental effort and attribute groups. As expected, the mean correlation of the total subjective mental effort (SME_{tot}, i.e., the sum of both SME measurements) with post-use PQ was substantial and negative, $r = -.74$, $t(8) = 3.11$, $p < .05$, whereas the mean correlation for SME_{tot} with HQI, $r = -.29$, and SME_{tot} with HQS, $r = -.17$, were small and not significant. A similar pattern emerged for beauty and goodness. The mean correlation of goodness with SME_{tot} was marginally significant, $r = -0.56$, $t(8) = 1.91$, $p < .10$, whereas the correlation of beauty with SME_{tot} remained nonsignificant, $r = -.17$.

Participants experienced usability problems (indicated by increased mental effort), which directly influenced the post-use assessment of pragmatic quality. In contrast, post-use hedonic quality is not related to the amount of the mental effort spent. This difference was also apparent in the relation of mental effort with goodness and beauty. Goodness is related to pragmatism and, thus, is more influenced by the amount of mental effort spent, although beauty was more tied to hedonic attributes and thus not as closely related to mental effort.

4.3. Discussion

Skins perceived to be more beautiful (ts2-Razor, QuickSkin) than others (Danzig, w98) in the pretest and Study 1 were again rated as more beautiful. Consistent with Study 1, a predominantly pragmatic and a predominantly stimulating skin were found on each level of beauty (ugly, beautiful). Although the attribute perceptions considerably varied among skins again, only the variation of HQI was related to beauty. On average, beautiful skins were perceived to be primarily better in providing identification (HQI), followed by being more stimulating (HQS).

In contrast to Study 2, differences in the correlation patterns of the attribute groups with beauty and goodness before using the skins were not as prominent. However, an interesting difference in the post-use pattern of beauty and goodness emerged. PQ became relevant for judgments of goodness but not for judgments of beauty. One may conclude that goodness is based on both pragmatic and hedonic attributes, whereas beauty is primarily based on hedonic attributes.

Pragmatic attributes as well as goodness were affected by experience (i.e., usability problems), whereas hedonic attributes and beauty remained stable over time. This effect of usability problems experienced during usage on some post-use ratings is also obvious in the relation of mental effort to attribute groups, beauty, and goodness. Experienced mental effort is related to post-use ratings of PQ. As expected, the post-use HQI and HQS were less related to mental effort than the pre-use rating. This difference is mirrored in the correlation of mental effort with goodness and beauty. Post-use goodness is related to pragmatic attributes and is, therefore, more influenced by the amount of mental effort spent. In contrast, post-use beauty is tied to hedonic attributes and therefore not as closely related to mental effort.

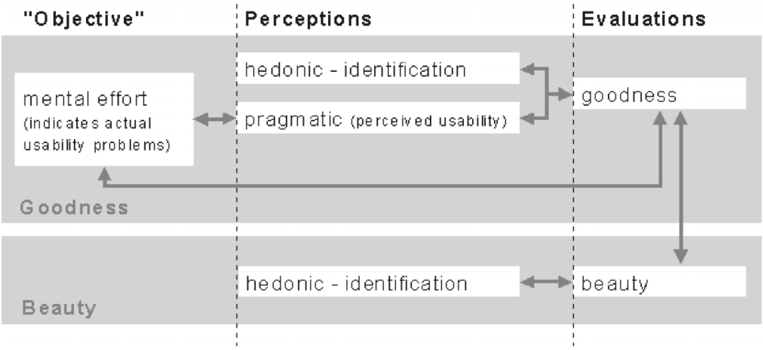
Although internally consistent, these results contradict the results of Tractinsky et al. (2000). They found no significant main effect of experimentally manipulated usability on post-use ratings of usability and beauty. They concluded that actual usability has no effect on perceptions of usability. However, an alternative explanation for their finding is a failure of the manipulation. Tractinsky et al. manipulated usability (normal, low) by, for example, increasing system response time, removing shortcuts, and including buttons that did not operate when first pressed. The manipulation was considered successful because task completion times in the low usability condition were extended. Task completion time is often treated as an efficiency measure (e.g., Wixon & Wilson, 1997). However, slightly increased task completion times are unlikely to have a strong impact on participants in a usability test setting. First, compared to effectiveness (i.e., whether participants are able to successfully complete a task at all), slight differences in efficiency might not seem as important from a user perspective. Second, in Tractinsky et al.'s (2000) experiment, participants lack a standard of comparison, because of the experiment's between-participants design. They might not even become aware of the minor differences in task completion time, as long as they do not directly experience a difference, that is, compare two products. Third, the absence of stress (i.e., low arousal in a task-oriented setting) is viewed as a crucial ingredient of satisfaction. DIN EN ISO 9241–11 (ISO, 1998) defines *satisfaction* as "freedom from discomfort," which is synonymous to absence of stress. Stress (i.e., mental effort) is rather induced by encountering problems (errors) or, more specifically, by extensive problem-handling times than by increased task-completion times (Hassenzahl, 2000). Individuals become aware of a barrier to task completion and the time and effort they must invest into overcoming this barrier. Consequently, mental effort is negatively correlated with perceived post-use usability (Hassenzahl, 2002). In other words, the aspect of usability (i.e., task completion time) manipulated by Tractinsky et al. is unlikely to induce stress, which in turn makes an impact on post-use ratings of usability unlikely. Indeed, Tractinsky et al. reported no effect of the usability manipulation on number of mistakes (errors). Overall, the mean number of errors was low (1.26), supporting the notion that participants were simply not stressed enough by the system to observe substantial impact on ratings of usability.

5. GENERAL CONCLUSION

Two studies showed beauty to be rather related to self-oriented, hedonic attributes of a product than to its goal-oriented, pragmatic attributes. The partial correlations between beauty and pragmatic attributes were only rarely substantial and never as high as one would expect from previous studies. Goodness, however, was related to pragmatic attributes, especially after participants used the respective products. Figure 6 summarizes the empirically found relationship between perceived attribute groups (pragmatic, hedonic), evaluational constructs (goodness, beauty) and objective aspects of the experience (mental effort).

What does this study tell us about beauty as a construct? First, beauty seems to imply a fairly outstanding quality. The initial selection of sample MP3-player skins contained 9 ugly and only 1 beautiful skin. This might have been a consequence of a biased selection of skins, more likely is it a consequence of beauty itself: "Beauty is rare in all nature's works, and in all works of art" (Voltaire, 1764/1984). Second, beauty's strong relation to identification, that is, to the product's ability to communicate a favorable identity to relevant others, makes it social. Beauty is something to be shared, to be approved by others. Otherwise, it would not be an effective statement in the constant dialog that people have with their social environment. This parallels a common finding in social psychology: Individuals in company of a beautiful friend are also perceived more favorably compared to being unaccompanied (an "assimilation effect," see e.g., Meiners & Sheposh, 1977). In the same vein, individuals may expect the beauty of things they own to rub off on them—to make them shine. Third, beauty as a judgment is not strongly affected by experience. Hedonic attributes may be derived from appearance, whereas prag-

Figure 6. **Summary of relationships between attribute groups (pragmatic, hedonic), evaluational constructs (goodness, beauty) and experience (mental effort).**



matic attributes may be derived from experience. As beauty is more strongly related to perceived hedonic attributes than to pragmatic attributes, it is less affected by experience compared to goodness.

What are the implications of the finding that identification is primarily related to beauty? First, beauty should be more relevant for preferences, if the product is to be owned. The beauty of a telephone we are about to buy is more important than the beauty of the telephone in a phone booth. Likewise, the beauty of an ATM might not be socially motivated because users neither own nor are likely to be identified with an ATM. This does not mean that individuals do not acknowledge beauty in things they do not own. On the contrary, beauty may be a driving force to become an owner, if possible. Beauty may then be understood as a prerequisite for a bonding between user and product. Second, the beauty of products may be a more potent purchase criterion if the product is used in social situations, such as mobile phones, laptop computers, and watches are. Only then, the beauty of a product in one's possession becomes a message to relevant others. These are testable hypotheses that should be addressed in future studies.

There are at least four important lessons to be learned from this study. They address the importance of the selection of products to be studied, the importance of an underlying model, general questions of causality, bottom-up versus top-down approaches to beauty, and the general meaning of the term *satisfaction* in HCI.

5.1. Selection of Products

Stereotypes are a consequence of an implicit personality theory (see Eagly et al., 1991) held by an individual. Such a theory specifies the applicability of attributes to other humans and those attributes' covariation. Individuals believe attributes to covary, such as beauty and social competence, although this must not necessarily be true for a particular individual to whom the theory is applied. In other words, the way we approach and treat other individuals is influenced by our—maybe wrong—theories about the relations between directly perceivable and only indirectly accessible attributes. However, the same mechanism allows us to make inferences about others beyond the merely perceived. I assume similar theories to exist for products. To study and understand those theories requires a focus on the covariation of attributes in individuals. A covariation between, for example, beauty and usability is given, if some individuals perceive a product as beautiful and usable at the same time, whereas the other individuals perceive the same product as less beautiful and less usable at the same time. Individuals may possess several implicit "product personality theories," that is, assumed pattern of covariation, depending on the product type or family they are dealing with. For example, the implicit pattern of covariation between attributes may be different for leisure-related (e.g., MP3-players) compared to task-related interactive products (e.g., ATMs). To reduce the impact of the specific product–person interaction, one must study a number of different products. At this point, the heterogeneity of the product sample becomes crucial. Products must vary on as many features as possible: layout, colors, form, symbols, and signs. Only if covariation between two attributes remains stable for very different products (as within-person variance) may one truly speak of a stereotype held by a significant number of users.

5.2. The Importance of a Model

An adequate heterogeneity of the product sample requires an idea of which features and attributes may be important. Tractinsky et al. (2000), for example, planned their study with the implicit notion of a link between beauty and usability. Accordingly, the product sample varies on usability related features. Variations in usability may or may not have an impact on beauty, however, such a study cannot assess the relative importance of usability as substance of beauty compared to other attributes because other attributes are simply not addressed. Another instructive example is the Technology Acceptance literature. Usefulness was long seen as the major technology-related antecedent of technology acceptance. However, the introduction of perceived enjoyment explained a substantial amount of additional variance, thus altering our understanding of technology acceptance (e.g., Davis, Bagozzi, & Warshaw, 1992; Igbaria et al., 1994). Rich models of user experience seem necessary as a starting point for any study of implicit product theories. Only if empirical studies address as many potentially relevant attribute dimensions as possible can resulting inter-attribute relations be safely generalized.

5.3. Causality

In studies addressing the beauty of objects, two different questions can be asked: What impact has beauty and what makes something beautiful? Both questions are important and both imply opposite directions of causality. Tractinsky et al.'s (2000) claim, "what is beautiful is usable," for example, assumes beauty to have an impact on judgments of usability, whereas my present model assumes beauty to be an evaluative consequence of attribute perception, that is, an appraisal process. The only scientific method to test causal relations is an experiment. Its important features are: the deliberate manipulation of the presence or absence of the potential cause (a so-called factor) and the random assignment of participants to conditions (factor levels) where the cause is present or absent. In Tractinsky et al.'s study, beauty is not treated as a factor. Beauty had three levels (low, neutral, high). Participants were randomly assigned to one of those levels. However, the actual interface layout was then selected on the basis of the participants' own ratings. This self-selection prevents a strict causal interpretation, because one cannot be sure that a participant's beauty judgments are not confounded with other dimensions. The manipulation of beauty in this study avoided this problem by using skins pretested to be less and more beautiful. However, a strict test for causal relations of my model's variables would have required treating beauty as dependent rather than as an independent variable. To conclude, the assumption that beauty is a high-level construct, which integrates various low-level perceptions, does not necessarily imply strong causalities. Particular perceptions can lead to a specific judgment and attributes can be inferred from a judgment. Future research will certainly address questions of causality in a more convincing way than either this study or Tractinsky and colleagues' studies did.

5.4. Bottom-up Versus Top-down Approaches to Beauty

In all experimental studies that try to manipulate beauty by selecting pre-tested extreme stimuli, the question of what really was manipulated remains open. Stimuli judged to be more or less beautiful are often complex stimuli, which differ in a variety of additional ways. Many studies of beauty suffer from this problem. Burmester and colleagues (1999), for example, found differences in perceptions and evaluations of a functionally equivalent user interface in a standard graphical user interface and a designed version. Besides the fact that different experts worked on the interface (computer scientists versus graphic designer), the genuine difference between both versions remained unclear. Others seek to disentangle determinants of beauty by defining sets of operationalizable characteristics, such as balance, regularity, proportion, or rhythm (e.g., Ngo & Byrne, 2001) or presence or absence of colors (e.g., Mundorf, Westin, & Dholakia, 1993). All in all, one may distinguish between research in aesthetics that explores how objective, perceptual features of objects cause beauty (bottom-up) and research that describes how the subjective meaning of objects contributes to beauty (top-down).

5.5. Satisfaction

In HCI the term *satisfaction* is often used synonymously with perceived usability or at least with the overall evaluation of a product. I neither agree with the former nor with the latter use of satisfaction. I understand satisfaction as an emotional consequence of goal-directed product use (Hassenzahl, 2003). If we expectedly achieve a self-relevant goal, we are satisfied with *ourselves.* If we further attribute the achievement, at least in part, to the help of a product, we may value the product. The difference between a general evaluation (i.e., goodness) and this view of satisfaction is obvious. One cannot be satisfied with the product itself, because satisfaction is tied to a goal and according expectations about goal-achievement. Statisfaction can only rub off on the product if users attribute their achievement to the product. That ties satisfaction to actual product usage. The question "How satisfied are you with product X" just does not make sense without having used it earlier. If usage experience is available, the individual will search for instances of goal-achievement. Based on that, the importance or centrality of the product for those instances of goal-achievement will be assessed. In contrast, a general evaluation can be made on other grounds than mere usage, as shown in Study 1. Goodness, as addressed in this study, comes closest to the understanding of satisfaction as a general evaluation, which is widespread in HCI. I showed that goodness is by no means identical with perceived usability. Actual usability (as expressed by mental effort) and perceived usability (i.e., pragmatic attributes) can be important ingredients of goodness. Nevertheless, there are additional sources for goodness, such as the communication of a desired identity. The finding that beauty is related to goodness parallels Tractinsky et al.'s (2000) findings. But goodness is not identical with beauty and, indeed, no substantial relation between actual or perceived usability and beauty was found.

To conclude, despite HCI efforts in defining core constructs (e.g., satisfaction) there is no accepted working model available that would be able to describe the key elements of user experience. Widely accepted definitions of usability, such as ISO 9241–11 (ISO, 1998), frequently neither sufficiently address the subjective side of user experience nor take additional, hedonic attributes (i.e., stimulation, identification) into account. Alternative approaches, such as Jordan's (2000) pleasure model or Logan et al.'s (1994) emotional versus behavioral usability, seem too simplistic to derive meaningful and empirically testable hypotheses. The lack of agreement in definition of key elements (e.g., "is satisfaction just the subjective side of usability?") makes building up empirical knowledge about what constitutes user experience difficult. However, without proper knowledge about what user experience actually *is*, designing for experience seems a daunting mission. Future research must aim at unifying approaches to user experience. Its major objectives will be the selection of key constructs and a better understanding of their interplay.

NOTES

Acknowledgments. I would like to thank Boriana Gotcheva, Romanof Kroh, Reinhard Leichner, Don Norman, Uta Sailer, Noam Tractinsky, and two anonymous reviewers for their helpful comments on an earlier draft of this article, as well as Bettina Heindel and Björn Schölling for helping me with collecting the data.

Author's Present Address. Marc Hassenzahl, Darmstadt University of Technology, Institute of Psychology, Social Psychology and Decision-Making, Steubenplatz 12, 64293 Darmstadt, Germany. E-mail: hassenzahl@psychologie.tu-darmstadt.de

HCI Editorial Record. First manuscript received October 8, 2003. Revisions received February 24, 2004, March 8, 2004, and June 14, 2004. Accepted by Donald Norman. Final manuscript received July 21, 2004. *–Editor*

REFERENCES

Arnold, A. G. (1999). Mental effort and evaluation of user interfaces: a questionnaire approach. In H. -J. Bullinger & J. Ziegler (Eds.), *Proceedings of the HCII 99 international conference on Human–Computer interaction, vol. 1* (pp. 1003–1007). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Berlyne, D. E. (1968). Curiosity and exploration. *Science, 153,* 25–33.

Burmester, M., Platz, A., Rudolph, U., & Wild, B. (1999). Aesthetic design—just an add on? In H. -J. Bullinger & J. Ziegler (Eds.), *Proceedings of the HCII 99 international conference on Human–Computer interaction, vol. 1* (pp. 671–675). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Csikszentmihalyi, M. (1975*). Beyond boredom and anxiety.* San Francisco: Jossey-Bass.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1992). Extrinsic and intrinsic motivation to use computers in the workplace. *Journal of Applied Social Psychology, 22,* 1111–1132.

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

Djajadiningrat, J. P., Overbeeke, C. J., & Wensveen, S. A. G. (2000). Augmenting fun and beauty: A pamphlet. *Proceedings of DARE 2000 Designing Augmented Reality Environment.* New York: ACM.

Draper, S. W. (1999). Analysing fun as a candidate software requirement. *Personal Technology, 3,* 1–6.

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,* 109–128.

Eilers, K., Nachreiner, F., & Hänecke, K. (1986). Entwicklung und Überprüfung einer Skala zur Erfassung subjektiv erlebter Anstrengung [Development and validation of a scale for the measurement of subjective mental effort]. *Zeitschrift für Arbeitswissenschaft, 40,* 215–224.

Gaver, W. W., & Martin, H. (2000). Alternatives. Exploring information appliances through conceptual design proposals. *Proceedings of the CHI 2000 Conference on Human Factors in Computing.* New York: ACM.

Gediga, G., Hamborg, K. -C., & Düntsch, I. (1999). The IsoMetrics usability inventory: An operationalization of ISO 9241–10 supporting summative and formative evaluation of software systems. *Behaviour & Information Technology, 18,* 151–164.

Goldstein, E. B. (1989). *Sensation and perception (3rd ed.).* Belmont, CA: Wadsworth.

Hassenzahl, M. (2000). Prioritising usability problems: Data-driven and judgement-driven severity estimates. *Behaviour & Information Technology, 19,* 29–42.

Hassenzahl, M. (2002). The effect of perceived hedonic quality on product appealingness. *International Journal of Human–Computer Interaction, 13,* 479–497.

Hassenzahl, M. (2003). The thing and I: understanding the relationship between user and product. In M. Blythe, C. Overbeeke, A. F. Monk, & P. C. Wright (Eds.), *Funology: From usability to enjoyment* (pp. 31–42). Dordrecht: Kluwer.

Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität [AttracDiff: A questionnaire to measure perceived hedonic and pragmatic quality]. In J. Ziegler & G. Szwillus (Eds.), *Mensch & Computer 2003. Interaktion in Bewegung* (pp. 187–196). Stuttgart, Leipzig: B. G. Teubner.

Hassenzahl, M., Kekez, R., & Burmester, M. (2002). The importance of a software's pragmatic quality depends on usage modes. In H. Luczak, A. E. Cakir, & G. Cakir (Eds.), *Proceedings of the 6th international conference on Work With Display Units (WWDU 2002;* pp. 275–276). Berlin: ERGONOMIC Institut für Arbeits- und Sozialforschung.

Hassenzahl, M., Platz, A., Burmester, M., & Lehner, K. (2000). Hedonic and ergonomic quality aspects determine a software's appeal. *Proceedings of the CHI 2000 Conference on Human Factors in Computing.* New York: ACM.

Igbaria, M., Schiffman, S. J., & Wieckowski, T. J. (1994). The respective roles of perceived usefulness and perceived fun in the acceptance of microcomputer technology. *Behaviour & Information Technology, 13,* 349–361.

ISO (1996). ISO 9241: *Ergonomic requirements for office work with visual display terminals (VDTs) —Part 10: Dialogue principles.* International Organization for Standardization. Geneva, Switzerland: International Organization for Standardization.

ISO (1998). ISO 9241: *Ergonomic requirements for office work with visual display terminals (VDTs)—Part 11: Guidance on usability.* International Organization for Standardization. Geneva, Switzerland: International Organization for Standardization.

Jordan, P. (2000). *Designing pleasurable products: An introduction to the new human factors.* London: Taylor & Francis.

Kirakowski, J. (1998). *SUMI user handbook.* York University College: Human Factors Research Group.

Kunze, E. -N. (2001). How to get rid of boredom in waiting-time-gaps of terminal-systems. In M. G. Helander, H. M. Khalid, & T. Ming Po (Eds*.), Proceedings of The International Conference on Affective Human Factors Design.* London: Asean Academic Press.

Kurosu, M., & Kashimura, K. (1995). Apparent usability vs. inherent usability. *Proceedings of the CHI 95 Conference on Human Factors in Computing.* New York: ACM.

Lavie, T. & Tractinsky, N. (2004). Assessing dimensions of perceived visual aesthetics of Web sites. *International Journal of Human–Computer Studies, 60,* 269–298.

Leventhal, L., Teasley, B., Blumenthal, B., Instone, K., Stone, D., & Donskoy, M. V. (1996). Assessing user interfaces for diverse user groups: Evaluation strategies and defining characteristics. *Behaviour & Information Technology, 15,* 127–137.

Logan, R. J., Augaitis, S., & Renk, T. (1994). Design of simplified television remote controls: A case for behavioral and emotional usability. In *Proceedings of the 38th Human Factors and Ergonomics Society Annual Meeting* (pp. 365–369). Santa Monica, CA: HFES.

Meiners, M. L., & Sheposh, J. P. (1977). Beauty or brains: Which image for your mate? *Personality & Social Psychology Bulletin, 3,* 262–265.

Monk, A. F., & Frohlich, D. (1999). Computers and fun. *Personal Technology, 3,* 91.

Mundorf, N., Westin, S., & Dholakia, N. (1993). Effects of hedonic components and user's gender on the acceptance of screen-based information services. *Behaviour & Information Technology, 12,* 293–303.

Ngo, D. C. L., & Byrne, J. G. (2001). Application of an aesthetic evaluation model to data entry screens. *Computers in Human Behavior, 17,* 149–185.

Nielsen, J. (1993). *Usability engineering.* Boston: Academic.

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

Ortony, A., Clore, G. L., & Collins, A. (1988). *The cognitive structure of emotions.* Cambridge, MA: Cambridge University Press.

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

Roberts, L., Rankin, L., Moore, D., Plunkett, S., Washburn, D., & Wilch-Ringen, B. (2003). Looks good to me. *Proceedings of the CHI 2002 Conference on Human Factors in Computing.* New York: ACM.

Rozin, P. (2003). Introduction: Evolutionary and cultural perspectives on affect. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective science* (pp. 839–851). New York: Oxford University Press.

Russo, B., & De Moraes, A. (2003). The lack of usability in design icons: An affective case study about *Juicy Salif. Proceedings of the 2003 International Conference on Designing Pleasurable Products and Interfaces.* New York: ACM.

Schenkman, B. N., & Jönsson, F. U. (2000). Aesthetics and preferences of Web pages. *Behaviour & Information Technology, 19,* 367–377.

Schwartz, S. H., & Bilsky, W. (1987). Toward a universal psychological structure of human values. *Journal of Personality and Social Psychology, 3,* 550–562.

Tractinsky, N. (1997). Aesthetics and apparent usability: Empirically assessing cultural and methodological issues. *Proceedings of the CHI 1997 Conference on Human Factors in Computing.* New York: ACM.

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

Tractinsky, N., & Zmiri, D. (inpress). Exploring attributes of skins as potential antecedents of emotion in HCI. In P. Fishwick (Ed.), *Aesthetic computing.* MIT Press.

Voltaire (Arouet, F. M.) (1764/1984). The philosophical dictionary—rare (reprinted by Penguin Books). Retrieved October 18, 2004 from http://users.compaqnet.be/cn111132/Voltaire/volrare.htm

Wicklund, R. A., & Gollwitzer, P. M. (1982). *Symbolic self-completion.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Wilson, K. (2002). Evaluating images of virtual agents. *Proceedings of the CHI 2002 Conference on Human Factors in Computing.* New York: ACM.

Wixon, D., & Wilson, C. (1997). The usability engineering framework for product design and evaluation. In M. G. Helander, T. K. Landauer, & P. Prabhu (Eds.), *Handbook of Human–Computer Interaction (2nd ed.*). Englewood Cliffs, NJ: Elsevier.

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist, 35,* 151–175.

Zangwill, N. (2003). Beauty. In J. Levinson (Ed.), *Oxford handbook of aesthetics*. Oxford: Oxford University Press.

Zijlstra, R., & van Doorn, L. (1985). *The construction of a scale to measure subjective effort (Tech. Rep.).* Delft, Netherlands: Delft University of Technology, Department of Philosophy and Social Sciences.

Copyright of Human-Computer Interaction is the property of Lawrence Erlbaum Associates and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.

View publication stats