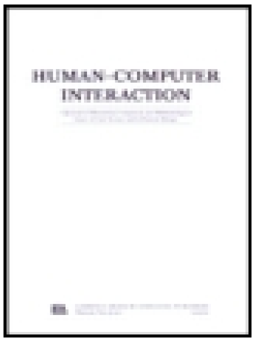


HUMAN–COMPUTER INTERACTION

HUMAN–COMPUTER INTERACTION

ISSN: 0737-0024 (Print) 1532-7051 (Online) Journal homepage: https://www.tandfonline.com/loi/hhci20

The Inference of Perceived Usability From Beauty

Marc Hassenzahl & Andrew Monk

To cite this article: Marc Hassenzahl & Andrew Monk (2010) The Inference of Perceived Usability From Beauty, HUMAN–COMPUTER INTERACTION, 25:3, 235-260, DOI: 10.1080/07370024.2010.500139

To link to this article: https://doi.org/10.1080/07370024.2010.500139

HUMAN–COMPUTER INTERACTION, 2010, Volume 25, pp. 235–260
Copyright © 2010 Taylor & Francis Group, LLC
ISSN: 0737-0024 print / 1532-7051 online
DOI: 10.1080/07370024.2010.500139

Taylor & Francis
Taylor & Francis Group

The Inference of Perceived Usability From Beauty

Marc Hassenzahl[1] **and Andrew Monk**[2]
[1]*Folkwang University, Germany*
[2]*University of York, United Kingdom*

A review of 15 papers reporting 25 independent correlations of perceived beauty with perceived usability showed a remarkably high variability in the reported coefficients. This may be due to methodological inconsistencies. For example, products are often not selected systematically, and statistical tests are rarely performed to test the generality of findings across products. In addition, studies often restrict themselves to simply reporting correlations without further specification of underlying judgmental processes.

The present study's main objective is to re-examine the relation between beauty and usability, that is, the implication that "what is beautiful is usable." To rectify previous methodological shortcomings, both products and participants were sampled in the same way and the data aggregated both by averaging over participants to assess the covariance across ratings of products and by averaging over products to assess the covariance across participants. In addition, we adopted an inference perspective to qualify underlying processes to examine the possibility that, under the circumstances pertaining in most studies of this kind where participants have limited experience of using a website or product, the relationship between beauty and usability is mediated by goodness.

A mediator analysis of the relationship between beauty, the overall evaluation (i.e., "goodness") and pragmatic quality (as operationalization of usability) suggests that the relationship between beauty and usability has been overplayed as the correlation between pragmatic quality and beauty is wholly mediated by goodness. This pattern of relationships was consistent across four different data sets and different ways of data aggregation. Finally, suggestions are made regarding methodologies that could be used in future studies that build on these results.

Marc Hassenzahl is psychologist interested in applying research on social cognition, judgment, choice and decision-making to human–computer interaction; he is Professor for Ergonomics and User Experience at the Design Faculty of the Folkwang University in Essen, Germany. **Andrew Monk** is an HCI researcher with an interest in requirements for technology in the home; he is a Professor in the Department of Psychology of University of York, United Kingdom.

CONTENTS

1. INTRODUCTION

Over the last few years, researchers in human–computer interaction (HCI) have become increasingly interested in aesthetics, that is, beauty (Hassenzahl, 2008; Lindgaard & Whitfield, 2004; Norman, 2004). The study of beauty has now become a part of user experience research, an approach to HCI, which emphasizes subjectively experienced, positive, and noninstrumental outcomes of owning and using interactive products as a complement to the traditional, predominantly task-oriented approach (e.g., Hassenzahl & Tractinsky, 2006).

One particular question concerns the relationship between perceived beauty and perceived usability. In a widely recognized publication, Tractinsky, Katz, and Ikar (2000) found a substantial correlation between ratings of beauty and ratings of usability. They compared their results to findings from studies of social perception where people judged beautiful are found also to be judged good ("what is beautiful is good"; Dion, Berscheid, & Walster, 1972). Accordingly, Tractinsky and colleagues (2000) titled their article, "What Is Beautiful Is Usable."[1]

The notion that "what is beautiful is usable" was influential in the further recognition of aesthetics as an important aspect of HCI. Usability had previously

[1] Note that despite this title, Tractinsky et al. (2000) were careful not to attribute direct causation in their conclusions.

been thought of as an objectively measurable quality assessed by measures such as task completion time. That users would make consistent judgments of usability and that this was related to their judgments of beauty was thus surprising. In his book *Emotional Design*, Norman (2004, p. 18), for example, described this puzzlement by taking up Tractinsky's (1997) quote of Herbert Read (1953): "It requires a somewhat mystical theory of aesthetics to find any necessary connection between beauty and functionality." Intuitively, beauty and usability may make strange bedfellows—the empirical findings nevertheless suggest a solid relationship between both.

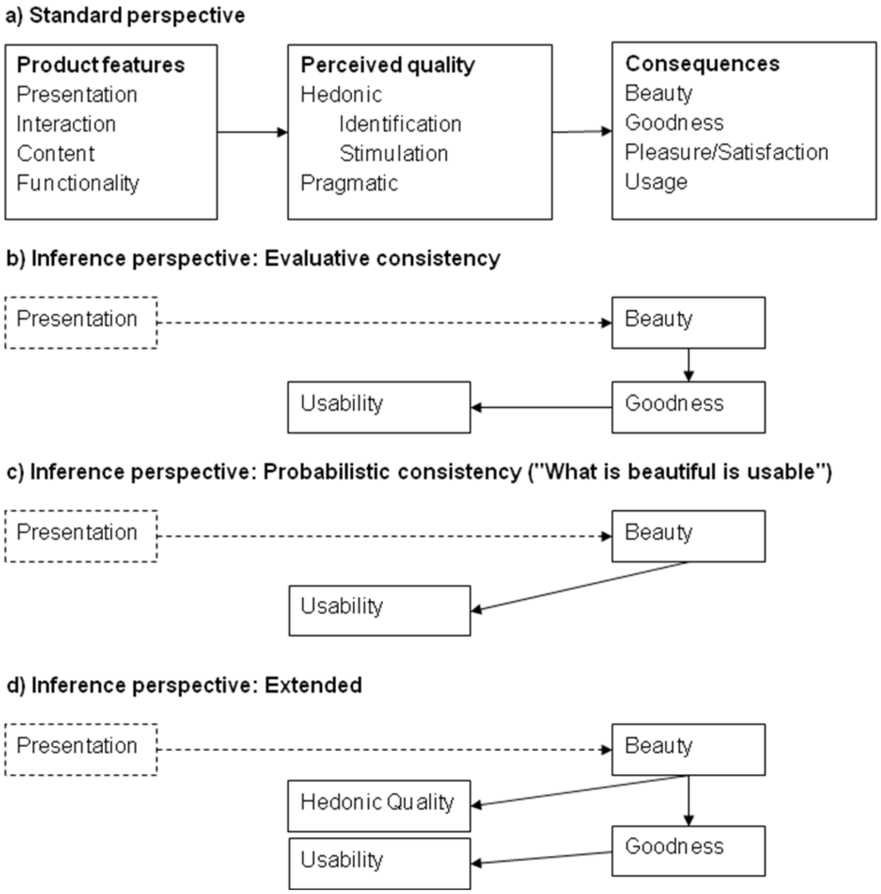
This conceptual dissociation between objective and experiential usability marks a fundamental change in the way HCI researchers consider the goals of their science. It is reflected in the detailed theoretical work of authors such as McCarthy and Wright (2004) that take a holistic approach to acting, sensing, thinking, and feeling. Our approach, and that of the other authors considered in this article, is rather different and draws instead of empirical psychological methods for clarifying psychological constructs using rating data. However, the aim is the same, that is, to refine a set of psychological constructs through which we can understand how users think about usability and beauty and how these constructs affect user behavior and preferences.

However inspiring for the development of the field, the findings of Tractinsky and colleagues (2000) have been challenged. Hassenzahl (2004), for example, studied different MP3 player skins and found no correlation between beauty and perceived usability (see also van Schaik & Ling, 2008, 2009). Stimulated by these contradictory findings, the present article's aim is to further examine and clarify the strength and implications of the correlation between beauty and usability. It starts (Section 2) with a description of four central constructs: beauty, goodness, pragmatic, and hedonic quality. Section 3 presents a quantitative review of studies reporting correlations between beauty (and similar constructs) and usability (and similar constructs). This revealed a huge variation not only in the correlation coefficients obtained but also in the products and analytical approach used. Section 3 further suggests a new, inferential perspective on the relationship between constructs to explain the inconsistency of findings. The remainder of the article presents and discusses data to support this approach (Section 4 and 5).

2. CONSTRUCTS

Figure 1a presents a conventional model based on Hassenzahl's (2003) user experience model. Product *features*, that is, the way the product is presented, what it does, and so on, lead to particular product *perceptions*, such as usability, which in turn have specific *consequences*, such as how much the interactive product is valued or how often it is used. Consequences are dependent on context. For example, usability may be con-

FIGURE 1. Models: (a) a simple causative model based on Hassenzahl's (2003) user experience model; (b) model implied by "evaluative consistency"; (c) model implied by "probabilistic consistency" (what is beautiful is usable); (d) model extended to hedonic quality.



sidered more important, when the product is used for an important and time-critical task and less so in a more relaxed mode of exploration and discovery (e.g., Hassenzahl & Ullrich, 2007). Perceived qualities, on the other hand, are independent of context. For example, the perceived usability of a product is taken to be the same no matter what the exact characteristics of the current context are. This general idea of more or less stable product perceptions leading to context dependent consequences is also common to other, recent models of User Experience (e.g., Hartmann, Sutcliffe, & De Angeli, 2008).

The present article focuses on the interplay between four distinct constructs, namely, goodness, beauty, pragmatic quality, and hedonic quality. Goodness is the overall evaluation (the value) of a product in a given context. For this reason we put it in the rightmost box in Figure 1a as a "consequence." Beauty is defined as "a predominantly affect-driven evaluative response to the visual Gestalt of an object" (Hassenzahl, 2008, p. 291). Thus, in this scheme, beauty is also seen as a consequence, because beauty, like goodness, has strong evaluative connotations. The difference between goodness and beauty is that beauty is more based on the visual, that is, the presentation of a product. Note that other authors, such as Lavie and Tractinsky (2004) and Tractinsky and Zmiri (2006), do not make this distinction between perceptions and consequences. Indeed, the correlations upon which these arguments are made preclude strong conclusions about causative direction. Nevertheless we find this model convenient for thinking about the constructs put forward. The main point to be made here is that constructs must be enumerated and clearly defined in any study of this kind.

Turning to the central box in Figure 1a, two broad groups of "perceived qualities" are listed. This follows from Hassenzahl and colleagues (Hassenzahl, 2001, 2003; Hassenzahl, Platz, Burmester, & Lehner, 2000) who have argued that the perceived qualities of a product contributing to goodness can be roughly divided into instrumental, pragmatic and noninstrumental, self-referential, hedonic aspects (see also Batra & Ahtola, 1990). Specifically, "pragmatic quality" refers to a judgment of a product's potential to support particular "do-goals" (e.g., to make a telephone call). "Hedonic quality" is a judgment with regard to a product's potential to support pleasure in use and ownership, that is, the fulfilment of so-called be-goals (e.g., to be admired, to be stimulated).

Pragmatic quality is akin to a broad understanding of usability as "quality in use," which includes the judgment of the appropriateness of the functionality provided in addition to the ease of access to this functionality. To illustrate this consider the reduced AttrakDiff2 questionnaire used in Section 4. Here the semantic differential items: "confusing–clearly structured," "unpredictable–predictable," "simple–complicated," and "practical–impractical" operationalize pragmatic quality. The latter item, "practical–impractical," implies more than a narrow understanding of usability as ease-of-use.

In summary, pragmatic quality focuses on tasks; it addresses the "what" and "how" of interacting with a product. In contrast, hedonic quality focuses on personal needs and aspirations, the "why" of interaction. In this simple causative model, goodness and beauty are evaluative judgments resulting from the combination of these perceived qualities in a way that is dependent on the context of use. The present article specifically examines the relationship between goodness, beauty, pragmatic, and hedonic quality (Section 4). Before doings so, however, we review previous studies on the relationship between beauty and usability. Note that the constructs and the simple causative model set out in Figure 1a are not without controversy. They are provided here as an explicit statement of a starting point for the evaluation of the data to be presented in Section 4 and as a conceptual structure for interpreting the studies reviewed in Section 3. The statistical data presented in Section 4 suggest that under the circumstances most commonly pertaining in these studies Figure 1a is, in point of fact, not the way people make these ratings. Alternative well developed constructs, notably those of Tractinsky and colleagues are considered in Section 5.2.

3. QUANTITATIVE REVIEW: BEAUTY AND USABILITY

Inspired by the empirical work of Kurosu and Kashimura (1995) and Tractinsky (1997), a number of studies examined the relationship (i.e., correlation) of perceived aesthetics or beauty with perceived usability. We first summarize the reported coefficients and the methods used. Based on this, we point at a number of methodological and theoretical issues emerging from a closer look at these studies. Finally, we suggest an alternative theoretical approach to model the relationship between beauty and usability, which is then used in Section 4.

3.1. Beauty and Usability

A search of the literature revealed 15 papers reporting 25 independent correlations of beauty with usability. The results of these studies are summarized in order of their publication in Figure 2. Column Source references the paper and indicates how beauty and usability were operationalized.

Although the general question of the studies was similar, the labels for the constructs used and the way they were operationalized differed. Lavie and Tractinsky (2004), for example, distinguished between "usability," "classic aesthetics," "expressive aesthetics," "pleasurable interaction," and "service quality." Vilnai-Yavetz, Rafaeli, and Schneider-Yaacov (2005) suggested "instrumentality," "aesthetics," and "symbolism"; Hartmann and colleagues (2008) proposed "usability," "content," "aesthetics," "reputation," and "customization." Although superficially different, a closer look at the content of the constructs, however, reveals some communality (see Section 5.2 for a further discussion). Usability, instrumentality, and content, for example, all refer to instrumental, pragmatic aspects of interactive products, whereas reputation, aesthetics, symbolism refer to noninstrumental, rather self-referential, hedonic aspects.

Figure 2, column *r*, shows the Pearson correlation coefficients (or comparable coefficients) quantifying the relationship between beauty and usability. The coefficients ranged from .92 to .00, with a median of .49 (24% explained variance). Restricting the coefficients to Pearson correlations only—thereby removing smaller "controlled" correlations (e.g., partial correlations, standardized β-weights, path coefficients)—made almost no difference (median *r* = .50, 25% explained variance). In terms of effect size, a median correlation of .5 is a large effect (Cohen, 1992). However, the range (.92–.00) and variability in the obtained coefficients is still striking (25% percentile = .14; 75% = .68).

3.2. Methodological Inconsistency

The variation seen in Figure 2 could be due to a number of factors. Obviously, the way beauty and usability was measured varied considerably (Figure 2, Source column). Furthermore, the products varied (Figure 2, Product column). This is good in

FIGURE 2. Overview of studies reporting a relationship between beauty (and similar constructs) and usability (or similar constructs).

Source		Product	*r*	N_{par} (N_{pro})	Sampling Unit
Kurosu & Kashimura, 1995 (Beauty with ease-of-use)		ATM layouts	.59	252 (26)	Product
Tractinsky, 1997 (Beauty with ease-of-use)	Study 1	ATM layouts	.92	104 (26)	Product
	Study 2	ATM layouts	.83	81 (26)	Product
	Study 3	ATM layouts	.92	108 (26)	Product
Tractinsky et al., 2000 (Aesthetics with ease-of-use)	Preuse	ATM layouts	.66	124 (9)	Pooled
	Postuse	ATM layouts	.71	124 (1)	Participant
Hassenzahl, 2001 (reanalyzed for Hassenzahl, 2004) (Beauty with pragmatic quality)		Monitors	.18	15 (3)	Pooled
van Schaik & Ling, 2003 (Aesthetics with display quality)		Websites	.49	86 (2)	Participant
Lavie & Tractinsky, 2004 (Classic aesthetics with usability)	Initial	Websites	.68	384 (5)	Participant
	cross- validation	Websites	.78	384 (5)	Participant
Hassenzahl, 2004 (Beauty with pragmatic quality)	Study 1	MP3 player skins	.07[a]	33 (4)	Participant
	Study 2, preuse	MP3 player skins	.14[a]	11 (4)	Participant
	Study 2, postuse	MP3 player skins	.08[a]	11 (4)	Participant
Vilnai-Yavetz et al., 2005 (Aesthetics with instrumentality – ability to perform)		Office designs	.65	148 (148)	Combined[c]
Sutcliffe & De Angeli, 2005		Website	.50[b]	25 (1)	Participant
(Classical aesthetics with usability)		Website	.50[b]	25 (1)	Participant
Lindgaard et al., 2006		Websites			
(Visual appeal with "clear – confusing")			.63	31 (50)	Product
(Visual appeal with "simple – complex")			.10	31 (50)	Product
De Angeli, Sutcliffe, & Hartmann, 2006		Website	.38[b]	28 (1)	Participant
(Classic aesthetics with usability)		Website	.49[b]	28 (1)	Participant
Cyr et al., 2006 (Design aesthetics with ease of use)		Mobile Service	.23[a]	60 (1)	Participant
Mahlke, 2006 (Ease of use with beauty)		Digital audio players	.00[a]	30 (4)	Pooled
Hartmann et al., 2007 (Classic aesthetics with usability)		Websites	.43	43 (3)	Pooled
van Schaik & Ling, 2008	Preuse	Websites	.12	111 (4)	Participant
(Beauty with pragmatic quality)	Postuse	Websites	.41	111 (4)	Participant

Note. The table reports Pearson's correlation coefficients except superscripts a and b.

[a]"Controlled" coefficients (e.g., partial correlation coefficients, standardized β-weights, or path coefficient). [b]Original work reports only significance levels; coefficients are reconstructed from significance level and total sample size. [c]Each participant rated their own office; see Section 5 for discussion.

N_{par} = number of participants; N_{pro} = number of products; see text for explanation of "sampling unit."

that it tests the generalizability of the findings to different types of products. It is notable, for example, that the Automated Teller Machine (ATM) layouts used by Tractinsky et al. (2000), and before them Kurosu and Kashimura (1995), tend to produce higher than average correlation coefficients, though there are other differences between these studies and the others.

More seriously, participants and products were treated very differently in the studies reviewed. Whereas participants were sampled randomly, products were often selected arbitrarily or based on pretesting to represent, for example, extreme groups. This is unfortunate given the fact that the question of whether two constructs correlate depends not only on the individuals who judge but also on the objects to be judged (e.g., Monk, 2004). A correlation is a measure of covariance, that is, how much variance in one set of ratings can be explained by the variance in the other. A sample of participants, or products, selected to be homogeneous (low variance) will generally result in lower correlations than that would be found in a more heterogeneous sample.

The N_{par} (N_{pro}) column in Figure 2 gives the number of participants, N_{par}, sampled and the number of products, N_{pro}, sampled. Perhaps because undergraduate participants are easier to obtain than products, in all studies but one, the number of participants exceeded the number of products. The median ratio of participants to products was 14 (i.e., 14 times more participants than products).

This leads to the final point, which is that studies differed in the way data were aggregated before correlating (Figure 2, Sampling Unit column). As Monk (2004) pointed out, one may either have a *materials* or a *subjects* perspective on the constructs and their correlation.

In a materials analysis the correlation reflects covariance across materials—here, products—so the sampling unit is the product. Tractinsky (1997), for example, had 104 participants in his first study who each rated 26 different ATM layouts with respect to usability and beauty. Before correlating, he computed an average usability and beauty rating for each layout, leading to an N_{pro} of 26.

In a subjects analysis, the correlation reflects covariance across subjects, and so the sampling unit is the participant. Where the study used more than one product, ratings should be averaged across products to get a mean for each participant and these means then correlated. None of the available studies followed this procedure. Six studies only used one product making it unnecessary. Others reported separate participants analysis correlations for each product, or in one case, the mean participant analysis correlation computed as a mean across four products (Hassenzahl, 2004). These analyses are all labeled as having the sampling unit "participant" in the Sampling Unit column of Figure 2. Another analysis used was to treat the ratings of different products made by the same participant as independent observations. The studies are labeled as having the sampling unit "pooled" (Figure 2, Sampling Unit column). This method of (non)aggregation confuses variance across the two sampling units products and participants. In statistical terms, using participant as the sampling unit is to treat participant as a "random variable'," so that a test of statistical significance tests the generality of the finding (here a positive correlation) to other samples of participants from the same population. Using product as the sampling unit is to treat product as a "random variable," so that, a test of statistical significance tests the generality of the finding to other samples of products from the same product population. Of course, ratings of different products from the same participant are not likely to be statistically independent, thus, treating ratings of 3 products by 15 participants as 45 independent observations (Hassenzahl, 2001) may not result in a valid test of statistical significance.

The two different methods of data aggregation lead to correlations with seemingly different meaning. A correlation of beauty and usability found by materials analysis implies that products are either both beautiful and usable or ugly and unusable (as rated by an "average" participant). A potential cause might be the design process, which either leads to a product superior on all product attributes or not. In contrast, a correlation found by a subjects analysis shows that people who rate all the products higher than average on beauty also rate all the products high on usability, and people who rate all the products lower than average on beauty rate all the products low on usability. This would be observed if people, in general, consider beauty and usability to be similar concepts.

Although it is possible to construct data sets that exhibit a correlation in a subjects analysis but not in a materials analysis, or vice versa, see Monk (2004) for examples, this ought not to be the normal state of affairs. One would expect that any implicit model of covariance among constructs assumed by people would be based on the covariance they have experienced in products. Thus, if products and people (i.e., participants) are both sampled in such a way that they are representative of the natural variance in their respective pools, both materials and subject analysis should give the same picture.

What comes out of this discussion then is a strong plea for a particular methodological approach to studies addressing the relationship between beauty and usability and other quality aspects (see Clark, 1973, for a similar argument made in the area of psycholinguistics).

1. The method requires treating participants *and* products similarly. It is standard practice to specify the pool from which participants are drawn (e.g., undergraduate volunteers from such and such a university) and then make a case that the sample is representative of this pool because they are sampled randomly or exhaustively. Less commonly, but necessarily if the aforementioned argument is accepted, one should specify the pool from which products are sampled (e.g., health-oriented websites) and then sample randomly or exhaustively from that pool.
2. To claim that a particular pattern of correlations is reliable, it must be demonstrated in both materials and subjects analyses. The former demonstrates that the results apply across the defined product pool and the latter across the participant pool.

The present article takes this methodological approach.

3.3. Theoretical Issues: An Inference Perspective

Besides the methodological inconsistencies discussed in the preceding section, there are also theoretical shortcomings in previous works, which may result in inappropriate interpretations of observed correlations between beauty and usability. As we have seen, most of the available studies focus on simple bivariate correlations between constructs without further specification of underlying processes. However, consumer research suggests that the simple model that underlies most research in this area (depicted in Figure 1a) is limited as it ignores the possibility of inference when information is unavailable at the moment of judgment (see Kardes, Posavac, & Cronley, 2004, for an overview). This inference perspective assumes that when confronted with the need to judge or characterize a product people may use all currently available information and will infer the unavailable. We argue that beauty will play an important role as a starting point of these inference processes, because its primarily sensory nature makes it one of the most immediately available (e.g., Lindgaard, Fernandes, Dudek, & Brown, 2006). A "well-proportioned layout" may be easier to perceive immediately than a "good navigational structure" so that the latter is inferred from the former if no direct experience of the navigational structure is available because there had been no opportunity to use the product yet. Note that an inferential perspective on construct relationships does not necessarily contradict earlier models. A model of the kind proposed in Figure 1a will apply in situations where inference is not needed because information can be gathered, weighted and integrated deliberately into an overall evaluation.

Inference based on beauty may use two distinct mechanisms: "evaluative consistency" (Lingle & Ostrom, 1979) and "probabilistic consistency" (Ford & Smith, 1987). *Evaluative consistency* assumes that individuals infer a general value ("goodness") from all available attributes. Unavailable attributes are then inferred from goodness rather than from any specific available attribute. The original claim "beautiful is good" (Dion et al., 1972) is a case in point. It seemed that value was inferred from beauty and then further spread to conceptually different, conceptually unrelated aspects (e.g., beautiful people were believed to be better parents). Hence, the characterization of evaluative consistency as a *halo effect* evidenced by positively correlated attributes without conceptual similarity (Thorndike, 1920). Figure 1b characterises an explanation of the correlation between usability and beauty as evaluative consistency. In this hypothetical path diagram, the arrows indicate a potential causative link. Goodness is inferred from beauty, and then usability from goodness. A further way of describing this model, that we use to characterize the data presented in Section 4, is that goodness *mediates* the relationship between beauty and usability. An individual infers goodness from beauty, which in turn leads to higher estimates on unavailable or harder to assess attributes, such as usability. A beautiful layout may be used to infer a good navigational structure via the mediating concept of goodness.

Probabilistic consistency in contrast assumes that individuals infer unavailable attributes directly from some specific available attribute believed to be conceptually or even causally linked to the unavailable attribute. The original claim "what is beautiful is usable" (Traktinsky et al., 2000) can be thought of as an example of probabilistic consistency. Note, however, that the authors did not actually exclude the notion of mediating variables. Figure 1c captures this model. Usability is inferred from beauty directly without or at least with only partial mediation by goodness.

The purpose of the statistical analysis of our data presented in Section 4.2 is to falsify the model in Figure 1c, by supporting the model in Figure 1b, that is, to demonstrate that the relationship between beauty and usability is mediated by goodness. Such a result would suggest that unless individuals have substantial firsthand experience with the product to be evaluated usability is inferred (i.e., "guessed") from an overall judgment of goodness (value). As goodness is in turn highly influenced by beauty the result is a correlation between beauty and usability because of evaluative consistency (a halo effect).

One could ask the same question of the relationship between beauty and hedonic quality. Several studies have observed correlations between beauty and more general hedonic product qualities (Traktinsky & Zmiri, 2006; van Schaik & Ling, 2008). Perhaps the clearest statement of this relationship is in Hassenzahl (2004), which demonstrated a correlation between hedonic quality and beauty, especially with hedonic quality–identification. Hassenzahl (2004) argued for a conceptual link between beauty and hedonic quality (i.e., probabilistic consistency). Figure 2d summarizes our suggestion of a mediated link between beauty and usability (i.e., evaluative consistency) and a direct, conceptual link between beauty and hedonic quality (i.e., probabilistic consistency).

In the remainder of the article, we test the suggestions derived from the inference perspective, taking account of the methodological issues discussed earlier, that is, the selection of products and the aggregation of data to form materials and subjects analyses.

4. LABORATORY AND FIELD DATA FROM RATINGS OF WEBSITES AND OTHER INTERACTIVE PRODUCTS

The present study uses four distinct data sets (see Figure 3 for an overview). The first three were collected from undergraduates at the University of York in two studies, Sets 1 and 2 together, then Set 3. They represent a wide range of website genres, each of which was sampled in a systematic manner. For Set 1, participants rated 10 websites. Obtained ratings of beauty, goodness, pragmatic, and hedonic quality were averaged across websites to provide a subjects perspective. Set 2 had participants assess 60 websites each. Ratings were averaged across participants to provide a materials perspective. For Set 3 participants rated 30 websites. This set was aggregated to form a subjects (3a) and a materials perspective (3b). Set 4 consists of ratings gathered via an online questionnaire website (http://www.attrakdiff.de). Rating scales and products were in German. Four hundred thirty individual ratings pertaining to 21 different products were used to provide a subjects perspective (4a) and a materials perspective (4b). The subjects perspective ($N_{\text{par}} = 430$) was computed by centering individual ratings on the product average (i.e., subtracting the product average from the individual

FIGURE 3. Participants and products in the four studies.

Study	Participants	Products
1	60 students (56 female), averaging over websites gives *Data Set 1*	10 websites randomly selected from the 60 described below
2	10 students (5 female)	30 travel companies and 30 shopping sites (including electronic goods, clothes, books, and CDs), averaging over participants gives *Data Set 2*
3	57 students (30 female), averaging over websites gives *Data Set 3a*	10 gadget sites randomly selected from a Google search, 10 women's clothing sites randomly selected from www.shopsafe.co.uk and 10 UK higher education randomly selected from http://www.ucas.com, averaging over participants gives *Data Set 3b*
4	430 individuals (202 female) anonymously recruited, centred by subtracting the product's average value form the participants' individual value (thereby removing variance stemming from the product) gives *Data Set 4a*	21 different websites, averaging over a median of 18 participants (min = 14, max = 40) gives *Data Set 4b*

rating). The materials perspective (N_{pro} = 21) was computed by averaging a median number of 18 ratings for each product (min = 14, max = 40). The former removes all variance stemming from differences across products; the latter removes all variance stemming from differences across participants.

The data allow us to test our specific assumptions concerning the relation between beauty, goodness, pragmatic quality (i.e., perceived usability), and hedonic quality. Specifically, we suggest that the link between beauty and pragmatic quality is fully mediated by goodness (i.e., evaluative consistency), whereas the link between beauty and hedonic quality is at least in part direct (i.e., probabilistic consistency). The four distinct data sets provide information about the stability of the obtained results by replication with different products and by subjects and materials analysis.

4.1. Method

Participants. The participants for Sets 1 to 3 were psychology undergraduate students at the University of York with a mean age around twenty years. The participants for Study 4 were anonymously recruited through the http://www.attrakdiff.de website. The age distribution was 18 below 20 years old, 313 between 21 and 40 years, 90 between 41 and 60, and 9 older than 60.

Measures. To measure pragmatic and hedonic quality a short, eight-item version of the 21 item AttrakDiff2 questionnaire (Hassenzahl, Burmester, & Koller, 2003) was constructed. A shorter test was required because of the large number of websites each participant had to rate. Accordingly, the original 21 items were first translated from German into English and then selected by a native English speaker according to their representativeness for the pragmatic and hedonic constructs. We used this "face validity" selection criterion rather than an empirical criterion, such as item factor loadings from earlier studies, because available data were all obtained with a German version of the questionnaire. No data were available for the English translation. Note, however, that results from Data Set 4 obtained in German are quite similar to Sets 1 to 3 obtained in English.

Because of the requirement to have a very brief questionnaire, we collapsed the originally proposed hedonic quality–identification and hedonic quality–stimulation to a single hedonic quality scale. A simple two-components model was sufficient for the questions addressed in the present study. In addition to the four pragmatic and four hedonic items, participants were asked to rate the goodness (i.e., "bad–good") and beauty (i.e., "ugly–beautiful") of each product. Data Set 4 was obtained with the full 21-item German version of the questionnaire. To make results comparable, only the data from the relevant subset of items were analyzed.

Websites Rated. For Studies 1 and 2, an initial sample of 60 E-commerce websites was selected (see the appendix), all available and remaining unchanged throughout the period of study. All had the primary aim of facilitating and supporting sales and online transactions to simulate real stores and travel agents. The types of sites selected were thought to be typical of the kind of sites people commonly make online purchases from. The websites were chosen for their diversity in picture and text content, color, density of information, and layout. Ten websites were selected randomly from the original 60 for Study 2 (see the appendix). Thirty websites were selected for Study 3 (see the appendix). These were of three types each selected randomly from a larger pool (see Figure 3). All sites were available and remained unaltered during the period of testing.

Set 4 was drawn from the http://www.attrakdiff.de website. This is a web-based, free-of-charge questionnaire tool. Users (i.e., evaluators) simply log on and define an evaluation project. A link is then generated and sent to potential participants. Evaluators will get an automatic result summary; the data are anonymized and stored for scientific purposes. Unfortunately, nothing can be said about the specific interactive products constituting this sample.

Procedure. For Sets 1 to 3 participants all used 17-in. 1024 × 768 color screen and a high bandwidth Ethernet connection. The browser window was "maximized" to fill the whole screen. For Set 1, each participant was given a separately randomized list of the 10 websites and was asked to work through them in the order given. Participants were required to view the home page of each website and encouraged to browse the site very briefly. Written instructions asked the user to spend "no more than a couple of minutes" doing this and then to rate it according to their "first impressions" using a paper questionnaire supplied. This took between 15 and 20 min. For Set 2, participants were required to rate 60 websites. Instead of being given a paper list of sites to visit, participants were sent an e-mail attachment containing hyperlinks to all 60 websites. Once again, the order of presentation was randomized for each participant, and they were told to visit the sites in the order listed. As the task took roughly two hours to complete, short breaks were allowed if needed.

For Set 3, a countdown timer was introduced to control more closely the amount of time that participants spent exploring the user interface. This was positioned at the top of the screen over the web browser toolbar and set to 30 sec. During this time, participants were instructed to try and get an overall feel for the "look" and layout of the site and to think about how they might use it. They were free to follow internal links in this time. The questionnaire items were printed on a sheet, but participants entered their ratings into a Microsoft Excel spreadsheet immediately after viewing the site. Participants worked through the 30 websites at their own pace but had to complete the task in no more than four separate instalments.

Set 4 was obtained via http://www.attrakdiff.de. No information is available about the exact individual procedure in which the questionnaire was embedded.

4.2. Results and Discussion

In general, for each question considered, six parallel analyses are presented with the expectation that they all provide essentially the same conclusion. The six data sets used are drawn from Studies 1 to 4 (see Figure 3). Data Set 1 (N_{par}= 60, aggregated by averaging ratings for the 10 websites in Study 1), Set 2 (N_{pro} = 60 websites, aggregated by averaging ratings for 10 participants in Study 2), Set 3a (N_{par} = 57 aggregated by averaging ratings for 30 websites in Study 3), Set 3b ($\mathrm{N}_{\mathrm{pro}}$ = 30 websites, aggregated by averaging ratings for 57 participants in Study 3), Set 4a ($\mathrm{N}_{\mathrm{par}}$ = 430, centered), and Set 4b ($\mathrm{N}_{\mathrm{pro}}$ = 21 websites aggregated by averaging ratings from a median of 18 ratings). Sets 1, 3a, and 4a thus provide subjects analyses, whereas Sets 2, 3b and 4b provide materials analyses (see Figure 3 for summary).

Descriptives

Figure 4 shows the internal consistency of the pragmatic and hedonic quality scales as well as the correlation between both constructs for each of the six data sets. In general, internal consistency was very good (lowest Cronbach's α = .79). More importantly, correlation between constructs was in general low (*M r* = .23), with lower intercorrelations in the more controlled laboratory sets (Set 1, 2, 3a, 3b) and higher correlations in the field data (Set 4a, 4b). These results unambiguously support the notion of pragmatic quality (i.e., perceived usability) and hedonic quality as distinct and measurable components of product perception (Hassenzahl, 2001) and lend further credit to the validity of the measures employed in the present study. Figure 5 gives means and standard errors of these measures.

Principal Components Analyses

Although the high internal consistencies combined with the low construct intercorrelations already point at the factorial validity of the pragmatic/hedonic dis-

FIGURE 4. Internal consistency (Cronbach's α) for pragmatic quality (i.e., perceived usability) and hedonic quality and intercorrelations (raw scores).

	Subjects[a]		Materials[a]	
Quality	Pragmatic Quality	Hedonic Quality	Pragmatic Quality	Hedonic Quality
	[Set 1][a]		[Set 2][a]	
Pragmatic quality	(.84)		(.86)	
Hedonic quality	.06	(.81)	.27*	(.95)
	[Set 3a][b]		[Set 3b][c]	
Pragmatic quality	(.80)		(.95)	
Hedonic quality	.00	(.79)	.05	(.92)
	[Set 4a][d]		[Set 4b][e]	
Pragmatic quality	(.82)		(.94)	
Hedonic quality	.52**	(.87)	.45*	(.86)

[a]*N* = 60. [b]*N* = 57. [c]*N* = 30. [d]*N* = 430. [e]*N* = 21.

FIGURE 5. Mean and standard error for pragmatic quality (i.e., perceived usability), hedonic quality, goodness, and beauty (raw scores, maximum = 7).

Construct	Subjects [Set 1][a]		Materials [Set 2][a]		Subjects [Set 3a][b]		Materials [Set 3b][c]		Subjects [Set 4a][d]		Materials [Set 4b][e]	
Pragmatic quality (4 items)	4.89	.07	4.65	.06	4.66	.06	4.66	.09	4.61	.06	4.57	.17
Hedonic quality (4 items)	4.12	.06	3.92	.10	4.06	.05	4.06	.10	4.35	.06	4.36	.12
Goodness (bad – good)	4.75	.08	4.20	.09	4.29	.06	4.29	.10	4.84	.07	4.86	.16
Beauty (ugly – beautiful)	3.96	.08	3.83	.11	3.93	.07	3.93	.13	4.34	.07	4.38	.11

[a]*N* = 60. [b]*N* = 57. [c]*N* = 30. [d]*N* = 430. [e]*N* = 21.

tinction, additional Principal Component Analyses (PCA) with Varimax rotation were carried out to get an idea of the fit of the two-component model and the adequacy of the single items. For Set 1 and 2 the "Eigenvalue > 1"-rule was used as extraction criterion. After having established the two-dimensional structure, the number of components to be extracted was set to two for all subsequent PCAs. Figure 6 shows the results of the PCAs for each of the six data sets.

The percentage of explained variance ranged from 65 (3a) to 85 (3b), with an overall better fit for the materials perspective. In general, the two-component structure replicated satisfactorily, with only nine out of 48 possible cross-loadings being larger than .30. The largest unwanted cross-loading was .415 ("confusing – structured" in Set 1). However, there was not a single item with loadings of an equal size on both components. All in all, the selection of items, that is, the abbreviated version of the AttrakDiff2 questionnaire, was appropriate for differentiating between pragmatic quality (i.e., perceived usability) and hedonic quality.

FIGURE 6. Component loadings, Eigenvalues, and percentage of variance explained.

Item	Verbal Anchors			Subjects [Set 1]		Materials [Set 2]	
				1	2	1	2
PQ 1	confusing	–	structured	.788	*.342*	*.415*	.848
PQ 2[a]	impractical	–	practical	.837		*.404*	.821
PQ 3	unpredictable	–	predictable	.825			.844
PQ 4	complicated	–	simple	.822			.801
HQ 1	dull	–	captivating		.761	.955	
HQ 2[a]	tacky	–	stylish		.847	.913	
HQ 3	cheap	–	premium		.766	.907	
HQ 4[a]	unimaginative	–	creative		.810	.934	
			Eigenvalue	2.74	2.67	3.80	2.80
			% explained variance	34	33 (67)	48	35 (83)
				Subjects [Set 3a]		Materials [Set 3b]	
				1	2	1	2
PQ 1	confusing	–	structured	.844		.962	
PQ 2[a]	impractical	–	practical	.853		.976	
PQ 3	unpredictable	–	predictable	.732		.891	–.319
PQ 4	complicated	–	simple	.753		.891	
HQ 1	dull	–	captivating		.791		.861
HQ 2[a]	tacky	–	stylish		.828		.917
HQ 3	cheap	–	premium		.672		.883
HQ 4[a]	unimaginative	–	creative		.828		.925
			Eigenvalue	2.62	2.57	3.63	3.37
			% explained variance	33	32 (65)	45	42 (87)
				Subjects [Set 4a]		Materials [Set 4b]	
				1	2	1	2
PQ 1	confusing	–	structured	*.385*	.752	.951	
PQ 3[a]	impractical	–	practical	*.356*	.744	.821	*.389*
PQ 2	unpredictable	–	predictable		.746	.900	
PQ 4	complicated	–	simple		.845	.905	
HQ 1	dull	–	captivating	.811			.876
HQ 2[a]	tacky	–	stylish	.824		*.320*	.760
HQ 3	cheap	–	premium	.787	*.306*	*.414*	.706
HQ 4[a]	unimaginative	–	creative	.871			.924
			Eigenvalue	3.01	2.59	3.50	2.92
			% explained variance	38	32 (70)	43	37 (80)

Note. Extraction: Principal Components, Extraction criterion for [Sets 1, 2]: Eigenvalue > 1, for [Sets 3a, 3b, 4a, 4b]: two factors; Varimax rotation, all loadings < .30 are suppressed. Unwanted cross-loadings are in italics.

[a]Item was reversed.

Component scores (i.e., factor scores) for pragmatic and hedonic quality were calculated by regression. Compared to scale means, these scores have the advantage of being perfectly uncorrelated. This facilitates interpretation of subsequent analyses. All subsequent analyses use these component scores rather than the mean scores reported in Figures 4 to 6, however, similar analyses using the mean score give essentially the same results.

Construct Correlations and Mediator Analyses

Figure 7 shows the bivariate correlations between (a) beauty and goodness, (b) beauty and pragmatic quality (i.e., perceived usability), and (c) beauty and hedonic quality for each data set. What strikes the eye is the consistently low correlation between beauty and pragmatic quality (*M r* = .19). Only two out of six coefficients obtained approached significance. The relationship between beauty and hedonic quality, however, was substantial. All of the bivariate correlations were highly significant, and had a mean value of .79.

We suggest that any correlation observed for the relationship between beauty and pragmatic quality is due to evaluative consistency that is mediated by goodness (value). For hedonic quality, however, the link is predicted to be direct (see Figure 1d). These two hypotheses can be tested empirically using mediation analysis (Baron & Kenny, 1986). The direct effect of beauty on perceived usability (i.e., relationship between beauty and pragmatic quality given the influence of goodness on pragmatic quality is controlled) is estimated by a multiple regression with beauty and goodness as predictors and pragmatic quality as criterion. The direct effect is simply the standardized regression coefficient (i.e., β-weight) of beauty. The indirect effect is the original bivariate correlation minus the direct effect. The significance of the indirect effect (i.e., the complete path from beauty to goodness and from goodness to perceived usability) can then be tested with the Sobel-test (Sobel, 1982). We used the SPSS-macros suggested by Preacher and Hayes (2004) for this analysis. A parallel analysis was done for the relationship between beauty and hedonic quality.

Figure 7 gives the standardized β-weights from the regression analyses estimating the direct effect of beauty on pragmatic quality. Only two of the six coefficients were

FIGURE 7. Bivariate correlations and mediation analyses (component scores are used for pragmatic and hedonic quality).

	Subjects			Materials			
Correlation	[Set 1][a]	[Set 3a][b]	[Set 4a][c]	[Set 2][a]	[Set 3b][d]	[Set 4b][e]	M[f]
Beauty – goodness							
Bivariate	.58**	.63**	.72**	.81**	.76**	.74**	.71
Beauty – Pragmatic quality							
Bivariate	.13	–.11	.29**	.18	.24	.39	.19
Direct effect, controlling for goodness	–.13	–.39*	–.09	–.20	–.48*	–.20	–.23
Indirect effect (Sobel's Z)	.26* (2.51)	.28* (2.48)	.38** (6.67)	.38* (2.10)	.72** (3.41)	.59* (2.53)	.47
Beauty – Hedonic quality							
Bivariate	.74**	.69**	.74**	.80**	.93**	.69**	.79
Direct effect, controlling for goodness	.65**	.32**	.51**	.19*	.89**	.50†	.56
Indirect effect (Sobel's Z)	.09 (1.38)	.37** (4.15)	.23** (4.47)	.61** (6.30)	.04 (0.48)	.19 (1.05)	.28

[a]*N* = 60. [b]*N* = 57. [c]*N* = 30. [d]*N* = 430. [e]*N* = 21. [f]Coefficients were Fisher transformed, averaged, and retransformed, *not* weighted by *N*.

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

significant, and if so, they were negative. The average direct effect was –.23 indicating, at best, no conceptual link between beauty and perceived usability. The indirect effect of beauty on perceived usability, however, was in all cases significant (*M* effect = .47). Overall, this confirms that the relation between beauty and perceived usability is indirect, mediated by the general evaluation (i.e., goodness) of the product. The correlations between beauty and usability reported in the literature, where participants have limited experience of using a website or product, may then be the consequence of a very potent but often uncontrolled third variable, goodness.

A quite different inference processes is indicated for the relationship between hedonic quality and beauty. Here there is evidence for a direct effect (probabilistic consistency). Figure 7 also includes the mediation analyses for the relation between beauty and hedonic quality. Although the indirect effect was significant for three out of the six analyses, the mean indirect effect (.28) was much smaller than the mean direct effect (.56). All the direct effects were significant. This supports the notion of a direct conceptual link between beauty and hedonic quality while acknowledging the existence of a general halo-effect of goodness on hedonic quality.

Subjects and materials analysis did not differ much. In general, correlations were slightly stronger in the materials analysis. The most noticeable difference is the size of the indirect effect for beauty and usability. In the subjects analysis the mean effect is .33 but is .58 in the materials analysis.

5. SUMMARY AND CONCLUSION

5.1. Methodology

This article began with a quantitative review of previous studies reporting correlations between beauty and usability. In general, the correlations reported were positive but very variable. Further examination for these studies revealed considerable inconsistency in method and analysis. To remedy this situation, we suggested that researchers need to pay more attention to the selection of the products that are rated and the way the data are aggregated for analysis. First, products should be selected in the same way as participants, the pool from which they are selected should be specified and then how the products were sampled from that pool. Second, theoretically critical patterns of correlation should be demonstrated both in materials and subjects analyses. A materials analysis aggregates by averaging the scores of participants to provide a rating for each product. A correlation here demonstrates the strength of the relationship across the product pool. A subjects analysis aggregates by averaging the scores for products to provide a rating for each participant. A correlation here demonstrates the strength of the relationship across the participant pool. This double requirement has been widely used in psychology since it was suggested in the area of psycholinguistics by Clark (1973), and many psychology journals make a point of requiring authors to establish generalizability across participants and items, where appropriate (see, e.g., the Editorial Policy for *Cognition*). Technically, the requirement arises from the fact that, conceptually, both participants and products are what statisticians call "random variables." Previous work has tended to treat products as a "fixed effect."

The data presented here demonstrate how this may be done. Study 3, for example had 57 participants rate all of 30 websites. The ratings were then aggregated separately for a subjects analysis (Set 3a) and then for a materials analysis (Set 3b). This is a lot of data and requires a degree of dedication from the participants. Fortunately, the reliability of the ratings was shown to be high, countering any suggestion that this highly repetitive task could have lead to random responding. Studies 1 and 2 used the same pool of websites but sampled fewer websites for the subjects analysis (Study 1) than the materials analyses (Study 2). By averaging more than 10 websites for 60 participants (Study 1) and then more than 10 participants for 60 websites (Study 2), data collection was made more tractable. The field data collected for Set 4 showed another way in which the large amounts of data needed to fulfil these requirements may be amassed. Overall, the results from the present studies were highly consistent, which lends support the robustness of the pragmatic quality and hedonic quality distinction and the suggested interference model of the relationship between those two constructs and beauty and goodness, respectively.

Another solution to the problem of collecting data that are statistically interpretable without requiring participants to rate large numbers of websites is to "deliberately confound" participants and materials. For example, Vilnai-Yavetz et al. (2005) asked people to rate their own office. Assuming that all the participants came from different offices, 148 sampling units can be thought of as both an N_{par} and an N_{pro}. This means that a significance test simultaneously tests the generality of the finding across the participant population sampled and the product population (offices) sampled. This statistical design would be particularly suitable for collecting data through a website, where obtaining a large number of participants is easier than getting each of them to do a lot of work for you. Another approach is hierarchical linear modeling or multilevel analysis (e.g., Hox, 2002), a statistical procedure that allows for the simultaneous test of relationships between variables on a person (i.e., subject perspective) and product level (i.e., material perspective), given a sufficient number of participants and products.

The generality of the conclusions drawn from correlations such as those presented here can always be questioned based on the nature of the products selected. For example, one might propose that participants could have visited them in the past and that this could have influenced the ratings. Computing a materials analysis (product perspective) provides a way round this problem. Publishing a list of the sites used allows the reader to make a judgement about the characteristics of the materials used. Computing materials effects demonstrates that the results hold over materials that vary with regard to these characteristics as well as other characteristics yet to be identified. This latter point is easiest explained by taking the more familiar example of a participants perspective. In a conventional subjects analysis, we don't worry that our participants vary in age and gender; we might want to know the age and gender distribution, but we also recognize that there are a myriad of other variables that could have affected the results that we don't know about. By sampling from a population in a repeatable way we can be confident that, within the limits imposed by sampling error, the distribution of these known and unknown participant qualities are representative of the distribution in the wider population implied by our selection procedure. A significance test then tells us how likely it is that the result could be due to sampling error. Exactly the same logic applies to a materials analysis. Previous knowledge of a site is only one of the many characteristics that sites may vary on and many of these are unknown. A correlation describes the extent that one variable covaries with another; hence if there is no variance, there can be no correlation. We need the sites to vary, and it would be pointless to try to control variation in arbitrary characteristics of the websites used.

One contribution of this article is to recommend strongly the more systematic selection of products and the use of participants and materials analyses. We would not argue, as is the case in certain areas of psychology, that only results presented as both a materials and a subjects analysis should be published. We would however ask authors to be clear about what they have done and to point out that the results need to be replicated using data of the other kind before strong conclusions are drawn.

Another contribution is to suggest, by example, the statistical analyses of the scales used that are required preliminary to evaluation of the key hypotheses. It seems to us more than just good practice to cite reliability coefficients (e.g., Cronbach's α) for the scales used, also the correlations between scales that are hypothesized to be independent. Without the demonstration of high reliability and discriminant validity of Pragmatic and Hedonic Quality in our data, the other key correlations our arguments rest upon would be hard to interpret. Principle component analyses are also presented to further reinforce the argument that Pragmatic and Hedonic Quality are distinct and measurable components of product perception. These preliminary analyses in turn require N_{par} and N_{pro} to be high enough for correlations to be stable, but this is not generally a problem in this area.

5.2. The Constructs

The empirical part of this article focused on four distinct constructs: goodness, beauty, pragmatic quality, and hedonic quality as conceptualized and operationalized by Hassenzahl (2003). Strictly speaking, this limits the results to exactly this set of constructs and the particular measurement approach taken (e.g., semantic differential format, composite scales for pragmatic and hedonic quality, single-item scales for beauty and goodness). However, when constructing the quantitative overview (Section 3), we assumed an overlap between these constructs and those used by other authors. Some of this overlap is discussed here.

Most commonly used in the studies we reviewed are Lavie and Tractinsky's (2004) influential classic and expressive aesthetics. Classic aesthetics is specified as "aesthetic," "pleasant," "clear," "clean," and "symmetric," whereas expressive aesthetics is specified as "creative," "fascinating," "original," "sophisticated," and "uses special effects." Of interest, "creative" is also an item in the present short version of the hedonic quality scale. In addition, "original" is a part of the longer hedonic quality–stimulation scale of the AttrakDiff2. We argue that expressive aesthetics and hedonic quality are strongly overlapping constructs. Thus, we would predict Tractinsky and Zmiri's (2006) finding of a low correlation of .03 between expressive aesthetics and usability on the basis of our own observation of a low correlation between hedonic and pragmatic quality, i.e., perceived usability (see, e.g., Hassenzahl, 2001).

In Figure 2, when there was no other direct measure of beauty, we assume an overlap between classic aesthetics and beauty. We prefer the direct, single-item measure because this captures the laypersons' concept of beauty. In Lavie and Tractinsky's (2004) factor analysis (reported in Table 1), "beautiful" loaded negatively on classic aesthetics (–.34). In Tractinsky and Zmiri's (2006) factor analysis (reported in Table 21.1) "beautiful" and "creative" both load highly on the same factor. Classic aesthetics, then, is different from beauty. A potential explanation would be to understand classic aesthetics as a form of "visual" usability (i.e., "clear," "clean," and "symmetric") complementing the usability of interaction. This would explain the higher correlations of classic aesthetics with usability compared to all other cases where only a single item measure of beauty was employed (see Figure 2).

The discussion just presented demonstrates the difficulty of resolving post hoc inconsistencies between constructs used in different studies. There is a clear need for new studies, which take integration and advancement of existing findings more seriously.

5.3. Causative Models

Advances in the software available for statistical modelling have made possible much more sophisticated approaches and simple path analyses, as exemplified in Figure 1, are now widely used to reason about possible relationships between constructs and then to test competing possibilities. This approach led us to test the possibility that the previously reported relationship between perceived usability and beauty could be indirect. Results from four independent data sets (laboratory and field) not only supported the idea of pragmatic quality (i.e., perceived usability) and hedonic quality as independent aspects of product perception but also lend strong support to the notion that pragmatic quality is only indirectly related to beauty. The small observed bivariate correlation appeared to be wholly explainable as being mediated by the general evaluation of the product (i.e., goodness). Any found correlation is rather the consequence of a "halo"-effect (Thorndike, 1920) than the consequence of any substantial conceptual or causal link. The "beautiful is usable" stereotype is, thus, rather a "beautiful is good and good is usable" stereotype. Through the very same mechanism, beauty could be related to almost any product attribute one chooses to study, for example, trustworthiness, functionality, or reliability (but see Tractinsky et al., 2000). In contrast, the relationship between beauty and hedonic quality was shown to be largely direct and reveals that beauty is self-referential, linked to the needs beyond the mere achievement of tasks. In this sense, beauty becomes a way to communicate self-serving messages to relevant others or even a source of pleasurable stimulation.

The basis of the model tested here is an inference perspective. We would suggest this as a fruitful extension of existing models (e.g., Hartmann, Sutcliffe, & De Angeli, 2007, 2008; Hassenzahl, 2003; Lavie & Tractinsky, 2004; Mahlke, 2002), which often assume people to have information about all product attributes and that this information serves as input to a general product evaluation. The inference perspective, however, acknowledges that information on product attributes might differ in terms of immediacy and type. Accordingly, an inference perspective appears especially appropriate for beauty and perceived usability. While beauty is immediately accessible through the product's visual presentation, usability reveals itself through interaction with the product only. Returning to Figure 1, it suggests that although the causative model depicted in Figure 1a may be the case in circumstances when a user has had a reasonable amount of experience of using the product, where the user does not, they may fall back on inference via beauty and goodness (i.e., Figure 1d). This raises an interesting area for further research to determine, for example, whether experience can be shown to weaken the mediated effect of beauty on usability. In the present study it can safely be assumed that all participants had some, at least very brief, hands-on experience with the products before rating. This may be an explanation for the rather low bivariate correlations between beauty and perceived usability apparent in the present study.

The present study is correlative. The causality implied by the models in Figure 1 is solely a matter of theoretical reasoning and cannot be tested by the present data. In fact, the assumed direction of the effects could be reversed without changing the actual effect sizes found. Accordingly, experimental studies that seek to test causative models through experimental manipulation must further test effects of beauty on perceived usability (and vice versa). Tractinsky et al. (2000) manipulated the usability of their ATM simulations by adding a 9-sec delay and buttons that did not operate and demonstrated nonsignificant effects on ratings of beauty. Monk and Lelos (2007) used four models of can opener as a proxy for personal electronic products such as MP3 players. These were painted to create a "pretty" set and an "ugly" set. Despite the fact that participants briefly used these can openers and either saw only the pretty or only the ugly set, there were significant effects of beauty on usability ratings. More studies must establish causal effects of varying product features on quality perception and even performance.

The present study is an example of how an explicit modelling can lead to a better understanding of the relationship between beauty and usability, specifically, and the study of product perception and evaluation in general. Future studies may apply a similar approach to the study of product attributes not considered in the present study, such as trust. We also put forward the methods and analyses used here as examples of solutions to the difficult methodological problem of generalizing across both participants and products.

NOTES

Acknowledgments. We thank the following University of York students who collected the data used in Studies 1, 2 and 3: Carol Finlayson, Vicky Kent and Lowri Davies. We also like to thank User Interface Design GmbH, which generously maintains http://www.attrakdiff.de and provided the data of Study 4.

Authors' Present Addresses. Marc Hassenzahl, Ergonomics and User Experience, Faculty of Design, Folkwang University, 45141 Essen, Germany. E-mail: marc.hassenzahl@folkwang-uni.de. Andrew Monk, Department of Psychology, University of York, York, YO24 1EA. E-mail: A.Monk@psych.york.ac.uk.

HCI Editorial Record. Received February 26, 2008. Revision received March 16, 2009. Accepted by Gary Olson. Final manuscript received March 26, 2010.

REFERENCES

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychology research. *Journal of Personality and Social Psychology, 51,* 1173–1182.

Batra, R., & Ahtola, O. T. (1990). Measuring the hedonic and utilitarian sources of consumer choice. *Marketing Letters, 2,* 159–170.

Clark, H. H. (1973). The language as a fixed-effect fallacy: a critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behaviour, 12,* 335–359.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

De Angeli, A., Sutcliffe, A., & Hartmann, J. (2006). Interaction, usability and aesthetics: what influences users' preferences? In *Proceedings of the 6th ACM Conference on Designing interactive Systems (DIS '06)* (pp. 271–280). New York, NY: ACM.

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

Ford, G. T., & Smith, R. A. (1987). Inferential beliefs in consumer evaluations: An assessment of alternative processing strategies. *Journal of Consumer Research, 14,* 363–371.

Hartmann, J., Sutcliffe, A., & De Angeli, A. (2007). Investigating attractiveness in web user interfaces. *Proceedings of the CHI 07 Conference on Human Factors in Computing Systems.* New York: ACM.

Hartmann, J., Sutcliffe, A., & De Angeli, A. (2008). Towards a theory of user judgment of aesthetics and user interface quality. *ACM Transactions on Computer–Human Interaction (TOCHI), 15*(4), 15.1–30.

Hassenzahl, M. (2001). 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, the Netherlands: Kluwer.

Hassenzahl, M. (2004). The interplay of beauty, goodness and usability in interactive products. *Human–Computer Interaction, 19,* 319–349.

Hassenzahl, M. (2008). Aesthetics in interactive products: Correlates and consequences of beauty. In H. N. J. Schifferstein & P. Hekkert (Eds.), *Product experience* (pp. 287–302). San Diego, CA: Elsevier.

Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. [AttrakDiff: 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, Germany: B.G. Teubner.

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 Systems.* New York, NY: ACM.

Hassenzahl, M., & Tractinsky, N. (2006). User experience—A research agenda [Editorial]. *Behavior & Information Technology, 25,* 91–97.

Hassenzahl, M., & Ullrich, D. (2007). To do or not to do: Differences in user experience and retrospective judgments depending on the presence or absence of instrumental goals. *Interacting with Computers, 19,* 429–437.

Hox, J. (2002). *Multilevel analysis. Techniques and applications.* Mahwah, NJ: Erlbaum.

Kardes, F. R., Posavac, S. S., & Cronley, M. L. (2004). Consumer inference: A review of processes, bases, and judgment contexts. *Journal of Consumer Research, 14*, 230–256.

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

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

Lindgaard, G., Fernandes, G., Dudek, C., & Brown, J. (2006). Attention web designers: You have 50 milliseconds to make a good first impression. *Behavior & Information Technology, 25,* 115–126.

Lindgaard, G., & Whitfield, T. W. A. (2004). Integrating aesthetics within an evolutionary and psychological framework. *Theoretical Issues in Ergonomics Science, 5,* 73–90.

Lingle, J. H., & Ostrom, T. M. (1979). Retrieval selectivity in memory-based impression judgments. *Journal of Personality and Social Psychology, 37,* 180–194.

Mahlke, S. (2002). Factors influencing the experience of web site usage. *Proceedings of the CHI 2002 Conference on Computer–Human Interaction.* New York, NY: ACM.

Mahlke, S. (2006). Aesthetic and symbolic qualities as antecedents of overall judgments of interactive products. In N. Bryan-Kinns, A. Blanford, P. Curzon, & L. Nigay (Eds.), *People and computers XX – Engage* (pp. 57–64). London, UK: Springer.

McCarthy, J., & Wright, P. (2004). *Technology as experience.* Cambridge, MA: MIT Press.

Monk, A. (2004). The product as a fixed-effect fallacy. *Human–Computer Interaction, 19,* 371–375.

Monk, A. F., & Lelos, K. (2007) Changing only the aesthetic features of a domestic product can affect its apparent usability. In A. Venkatesh, T. Gonzalvez, A. Monk, & B. Buckner (Eds.), *Home informatics and telematics: ICT for the next billion. Proceedings of HOIT 2007, Chennai, India* (pp. 221–234). New York, NY: Springer.

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

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers, 36,* 717–731.

Read, H. (1953). *Art and industry* (3rd ed.). London, UK: Faber and Faber.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. In S. Leinhart (Ed.), *Sociological methodology 1982* (pp. 290–312). San Francisco, CA: Jossey-Bass.

Sutcliffe, A., & De Angeli, A. (2005). Assessing interaction styles in web user interfaces. In *Proceedings of INTERACT 05* (pp. 405–417). Berlin, Germany: Springer.

Thorndike, E. L. (1920). A constant error on psychological rating. *Journal of Applied Psychology, 4,* 25–29.

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, NY: ACM.

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

Tractinsky, N., & Zmiri, D. (2006). Exploring attributes of skins as potential antecedents of emotion in HCI. In P. Fishwick (Ed.), *Aesthetic computing* (pp. 405–421). Cambridge, MA: MIT Press.

van Schaik, P., & Ling, J. (2003). The effect of link colour on information retrieval in educational intranet use. *Computers in Human Behavior, 19,* 533–564.

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

van Schaik, P., & Ling, J. (2009). The role of context in the perception of the aesthetics of web pages over time. *International Journal of Human Computer Studies, 67,* 79–89.

Vilnai-Yavetz, I., Rafaeli, A., & Schneider-Yaacov, C. (2005). Instrumentality, aesthetics, and symbolism of office design. *Environment and Behavior, 37,* 533–551.

APPENDIX

FIGURE A-1. Websites used in Study 1.

http://uk.dk.com
www.abebooks.co.uk
www.archersdirect.co.uk
www.audleytravel.com**
www.bargaincrazy.com
www.bartleby.com
www.beachcombertours.co.uk
www.bmibaby.com
www.boden.co.uk
www.bookatrip.com
www.bookitlate.com
www.booksbytesandbeyond.com
www.booksonline.co.uk**
www.cheaptickets.com
www.citybreaksguide.com
www.coopelectricalshop.co.uk
www.cosmosholidays.co.uk
www.cottontraders.co.uk
www.countrybookshop.co.uk
www.crestaholidays.co.uk
www.crystalholidays.co.uk
www.designerdiscount.co.uk**
www.easyjet.com
0www.empiredirect.co.uk
www.eveningdresses.co.uk
www.footprint-adventures.co.uk
www.funkyclothing.co.uk
towww.go-nowtravel.com
www.harrods.com
www.hippocampus.co.uk
www.holidayoasis.com
www.holidays4less.com
www.hudsonmusic.com
www.jessops.com**
www.joebrowns.co.uk
www.kuoni.co.uk**
www.leisuredirection.co.uk
www.manningtravel.co.uk
www.maplin.co.uk
www.nitro-shopping.com
www.offpeakluxury.com
www.olympicholidays.co.uk
www.pashmina-pashminas.co.uk
www.pixmania.co.uk**
www.play.com**
www.promod.com
www.responsibletravel.com
www.saga.co.uk/travel
www.scambag.co.uk
www.snowandrock.com
www.streetsonline.co.uk
www.teletextholidays.co.uk
www.topshop.co.uk**
www.travelbag.co.uk
www.travelcitydirect.com**
www.travelfetch.com**
www.travelwizard.com
www.unbeatable.co.uk
www.uniqueholiday.co.uk
www.venditor.com

Note. Double asterisks indicate that this website was one of the 10 randomly selected for Study 2.

FIGURE A-2. Websites Used in Study 3.

Iwantoneofthose.com	Asos.com	Chichester.ac.uk
Firebox.com	Girlsbits.com	Craven-college.ac.uk
Boysstuff.co.uk	Warehousefashion.com	Soton.ac.uk
Rocketdistribution.com	Accessorize.co.uk	Ed.ac.uk/studying/undergraduate
Boystoys.co.uk	Brandedbags.com	Exe-coll.ac.uk
Gadgetsuk.com	Wallis-fashion.com	Leedsthomasdanby.ac.uk
Thesharperedge.co.uk	Fashionshop.co.uk	Leeds-art.ac.uk
Celagadgets.com	Jolaby.co.uk	Lincoln.ac.uk
Gadgetsinc.co.uk	Jocasi.com	Ntu.ac.uk
Paramountzone.com	Net-a-porter.com	Swansea.ac.uk